A Kernel Test for Three-Variable Interactions with Random Processes

Paul K. Rubenstein[123] **Kacper P. Chwialkowski**[3] **Arthur Gretton**[4]
[1]Machine Learning Group, University of Cambridge
[2]Empirical Inference, MPI for Intelligent Systems, Tübingen, Germany
[3]Department of Computer Science, University College London
[4]Gatsby Computational Neuroscience Unit, University College London
pkr23@cam.ac.uk, kacper.chwialkowski@gmail.com, arthur.gretton@gmail.com

Abstract

We apply a wild bootstrap method to the Lancaster three-variable interaction measure in order to detect factorisation of the joint distribution on three variables forming a stationary random process, for which the existing permutation bootstrap method fails. As in the *i.i.d.* case, the Lancaster test is found to outperform existing tests in cases for which two independent variables individually have a weak influence on a third, but that when considered jointly the influence is strong. The main contributions of this paper are twofold: first, we prove that the Lancaster statistic satisfies the conditions required to estimate the quantiles of the null distribution using the wild bootstrap; second, the manner in which this is proved is novel, simpler than existing methods, and can further be applied to other statistics.

1 INTRODUCTION

Nonparametric testing of independence or interaction between random variables is a core staple of machine learning and statistics. The majority of nonparametric statistical tests of independence for continuous-valued random variables rely on the assumption that the observed data are drawn *i.i.d.* Feuerverger [1993], Gretton et al. [2007], Székely et al. [2007], Gretton and Gyorfi [2010], Heller et al. [2013]. The same assumption applies to tests of conditional dependence, and of multivariate interaction between variables Zhang et al. [2011], Kankainen and Ushakov [1998], Fukumizu et al. [2008], Sejdinovic et al. [2013], Patra et al. [2015]. For many applications in finance, medicine, and audio signal analysis, however, the *i.i.d.* assumption is unrealistic and overly restrictive. While many approaches exist for testing interactions between time series under strong parametric assumptions Kirchgässner et al. [2012], Ledford and Tawn [1996], the problem of testing for general, nonlinear interactions has seen far less analysis: tests of pairwise dependence have been proposed by Gaisser et al. [2010], Besserve et al. [2013], Chwialkowski et al. [2014], Chwialkowski and Gretton [2014], where the first publication also addresses mutual independence of more than two univariate time series. The two final works use as their statistic the Hilbert-Schmidt Indepenence Criterion, a general nonparametric measure of dependence [Gretton et al., 2005], which applies even for multivariate or non-Euclidean variables (such as strings and groups). The asymptotic behaviour and corresponding test threshold are derived using particular assumptions on the mixing properties of the processes from which the observations are drawn. These kernel approaches apply only to pairs of random processes, however.

The Lancaster interaction is a signed measure that can be used to construct a test statistic capable of detecting dependence between three random variables [Lancaster, 1969, Sejdinovic et al., 2013]. If the joint distribution on the three variables factorises in some way into a product of a marginal and a pairwise marginal, the Lancaster interaction is zero everywhere. Given observations, this can be used to construct a statistical test, the null hypothesis of which is that the joint distribution factorises thus. In the *i.i.d.* case, the null distribution of the test statistic can be estimated using a permutation bootstrap technique: this amounts to shuffling the indices of one or more of the variables and recalculating the test statistic on this bootstrapped data set. When our samples instead exhibit temporal dependence, shuffling the time indices destroys this dependence and thus doing so does not correspond to a valid resample of the test statistic.

Provided that our data-generating process satisfies some technical conditions on the forms of temporal dependence, recent work by Leucht and Neumann [2013], building on the work of Shao [2010], can come to our rescue. The wild bootstrap is a method that correctly resamples from the null distribution of a test statistic, subject to certain conditions on both the test statistic and the processes from which the observations have been drawn.

In this paper we show that the Lancaster interaction test statistic satisfies the conditions required to apply the wild bootstrap procedure; moreover, the manner in which we prove this is significantly simpler than existing proofs in the literature of the same property for other kernel test statistics [Chwialkowski et al., 2014, Chwialkowski and Gretton, 2014]. Previous proofs have relied on the classical theory of V-statistics to analyse the asymptotic distribution of the kernel statistic. In particular, the Hoeffding decomposition gives an expression for the kernel test statistic as a sum of other V-statistics. Understanding the asymptotic properties of the components of this decomposition is then conceptually tractable, but algebraically extremely painful. Moreover, as the complexity of the test statistic under analysis grows, the number of terms that must be considered in this approach grows factorially.[1] We conjecture that such analysis of interaction statistics of 4 or more variables would in practice be unfeasible without automatic theorem provers due to the sheer number of terms in the resulting computations.

In contrast, in the approach taken in this paper we explicitly consider our test statistic to be the norm of a Hilbert space operator. We exploit a Central Limit Theorem for Hilbert space valued random variables Dehling et al. [2015] to show that our test statistic converges in probability to the norm of a related population-centred Hilbert space operator, for which the asymptotic analysis is much simpler. Our approach is novel; previous analyses have not, to our knowledge, leveraged the Hilbert space geometry in the context of statistical hypothesis testing using kernel V-statistics in this way.

We propose that our method may in future be applied to the asymptotic analysis of other kernel statistics. In the appendix, we provide an application of this method to the Hilbert Schmidt Independence Criterion (HSIC) test statistic, giving a significantly shorter and simpler proof than that given in Chwialkowski and Gretton [2014]

The Central Limit Theorem that we use in this paper makes certain assumptions on the mixing properties of the random processes from which our data are drawn; as further progress is made, this may be substituted for more up-to-date theorems that make weaker mixing assumptions.

OUTLINE: In Section 2, we detail the Lancaster interaction test and provide our main results. These results justify use of the wild bootstrap to understand the null distribution of the test statistic. In Section 3, we provide more detail about the wild bootstrap, prove that its use correctly controls Type I error and give a consistency result. In Section 4, we evaluate the Lancaster test on synthetic data to identify cases in which it outperforms existing methods, as well as cases in which it is outperformed. In Section 6, we provide proofs of the main results of this paper, in particular the aforementioned novel proof. Further proofs may be found in the Supplementary material.

[1] See for example Lemma 8 in Supplementary material A.3 of Chwialkowski and Gretton [2014]. The proof of this lemma requires keeping track of 4! terms; an equivalent approach for the Lancaster test would have 6! terms. Depending on the precise structure of the statistic, this approach applied to a test involving 4 variables could require as many as $8! = 40320$ terms.

2 LANCASTER INTERACTION TEST

2.1 KERNEL NOTATION

Throughout this paper we will assume that the kernels k, l, m, defined on the domains $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ respectively, are characteristic [Sriperumbudur et al., 2011], bounded and Lipschitz continuous. We describe some notation relevant to the kernel k; similar notation holds for l and m. Recall that $\mu_X := \mathbb{E}_X k(X, \cdot) \in \mathcal{F}_k$ is the mean embedding [Smola et al., 2007] of the random variable X. Given observations X_i, an estimate of the mean embedding is $\tilde{\mu}_X = \frac{1}{n}\sum_{i=1}^n k(X_i, \cdot)$. Two modifications of k are used in this work:

$$\bar{k}(x, x') = \langle k(x, \cdot) - \mu_X, k(x', \cdot) - \mu_X \rangle, \tag{1}$$

$$\tilde{k}(x, x') = \langle k(x, \cdot) - \tilde{\mu}_X, k(x', \cdot) - \tilde{\mu}_X \rangle \tag{2}$$

These are called the *population centered kernel* and *empirically centered kernel* respectively.

2.2 LANCASTER INTERACTION

The Lancaster interaction on the triple of random variables (X, Y, Z) is defined as the signed measure $\Delta_L P = \mathbb{P}_{XYZ} - \mathbb{P}_{XY}\mathbb{P}_Z - \mathbb{P}_{XZ}\mathbb{P}_Y - \mathbb{P}_X\mathbb{P}_{YZ} + 2\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z$. This measure can be used to detect three-variable interactions. It is straightforward to show that if any variable is independent of the other two (equivalently, if the joint distribution $\mathbb{P}_{XYZ}$ factorises into a product of marginals in any way), then $\Delta_L P = 0$. That is, writing $\mathcal{H}_X = \{X \perp\!\!\!\perp (Y, Z)\}$ and similar for $\mathcal{H}_Y$ and $\mathcal{H}_Z$, we have that

$$\mathcal{H}_X \ \vee \ \mathcal{H}_Y \ \vee \ \mathcal{H}_Z \Rightarrow \Delta_L P = 0 \tag{3}$$

The reverse implication does not hold, and thus no conclusion about the veracity of the $\mathcal{H}_{\cdot}$ can be drawn when $\Delta_L P = 0$. Following Sejdinovic et al. [2013], we can consider the mean embedding of this measure:

$$\mu_L = \int k(x, \cdot)l(y, \cdot)m(z, \cdot)\Delta_L P \tag{4}$$

Given an *i.i.d.* sample $(X_i, Y_i, Z_i)_{i=1}^n$, the norm of the mean embedding μ_L can be empirically estimated using empirically centered kernel matrices. For example, for the kernel k with kernel matrix $K_{ij} = k(X_i, X_j)$, the empirically centered kernel matrix $\tilde{K}$ is given by

$$\tilde{K}_{ij} = \langle k(X_i, \cdot) - \tilde{\mu}_X, k(X_j, \cdot) - \tilde{\mu}_X \rangle,$$

By Sejdinovic et al. [2013], an estimator of the norm of the mean embedding of the Lancaster interaction for *i.i.d.* samples is

$$\|\hat{\mu}_L\|^2 = \frac{1}{n^2}\left(\tilde{K} \circ \tilde{L} \circ \tilde{M}\right)_{++} \tag{5}$$

where $\circ$ is the Hadamard (element-wise) product and $A_{++} = \sum_{ij} A_{ij}$, for a matrix A.

2.3 TESTING PROCEDURE

In this paper, we construct a statistical test for three-variable interaction, using $n\|\hat{\mu}_L\|^2$ as the test statistic to distinguish between the following hypotheses:

$\mathcal{H}_0 : \mathcal{H}_X \;\vee\; \mathcal{H}_Y \;\vee\; \mathcal{H}_Z$
$\mathcal{H}_1 : \mathbb{P}_{XYZ}$ does not factorise in any way

The null hypothesis $\mathcal{H}_0$ is a composite of the three 'sub-hypotheses' $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$. We test $\mathcal{H}_0$ by testing each of the sub-hypotheses separately and we reject if and only if we reject each of $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$. Hereafter we describe the procedure for testing $\mathcal{H}_Z$; similar results hold for $\mathcal{H}_X$ and $\mathcal{H}_Y$.

Sejdinovic et al. [2013] show that, under $\mathcal{H}_Z$, $n\|\hat{\mu}_L\|^2$ converges to an infinite sum of weighted χ-squared random variables. By leveraging the *i.i.d.* assumption of the samples, any given quantile of this distribution can be estimated using simple permutation bootstrap, and so a test procedure is proposed.

In the time series setting this approach does not work. Temporal dependence within the samples makes study of the asymptotic distribution of $n\|\hat{\mu}_L\|^2$ difficult; in Section 4.2 we verify experimentally that the permutation bootstrap used in the *i.i.d* case fails. To construct a test in this setting we will use asymptotic and bootstrap results for mixing processes.

Mixing formalises the notion of the temporal structure within a process, and can be thought of as the rate at which the process forgets about its past. For example, for Gaussian processes this rate can be captured by the autocorrelation function; for general processes, generalisations of autocorrelation are used. The exact assumptions we make about the mixing properties of processes in this paper are discussed in Supplementary material A.7. For brevity in statements of results throughout this paper, however, we define sufficient *suitable mixing assumptions* in Section 3.

2.4 MAIN RESULTS

It is straightforward to show that the norm of the mean embedding (5) can also be written as

$$\|\hat{\mu}_L\|^2 = \frac{1}{n^2}\left(\widetilde{\tilde{K} \circ \tilde{L} \circ \tilde{M}}\right)_{++}$$

Our first contribution is to show that the (difficult) study of the asymptotic null distribution of $\|\hat{\mu}_L\|^2$ can be reduced to studying population centered kernels

$$\|\hat{\mu}_{L,2}^{(Z)}\|^2 = \frac{1}{n^2}\left(\overline{\overline{K} \circ \overline{L}} \circ \overline{M}\right)_{++}$$

where e.g.

$$\overline{K}_{ij} = \langle k(X_i,\cdot) - \mu_X, k(X_j,\cdot) - \mu_X\rangle,$$

Specifically, we prove the following:

Theorem 1. *Suppose that $(X_i, Y_i, Z_i)_{i=1}^n$ are drawn from a random process satisfying suitable mixing assumptions. Under $\mathcal{H}_Z$, $\lim_{n\to\infty}(n\|\hat{\mu}_{L,2}^{(Z)}\|^2 - n\|\hat{\mu}_L\|^2) = 0$ in probability.*

Our proof of Theorem 1 relies crucially on the following Lemma which we prove in Supplementary material A.1

Lemma 1. *Suppose that $(X_i)_{i=1}^n$ is drawn from a random process satisfying suitable mixing assumptions and that k is a bounded kernel on $\mathcal{X}$. Then $\|\hat{\mu}_X - \mu_X\|_k = O_P(n^{-\frac{1}{2}})$*

Proof. (Theorem 1) We provide a short sketch of the proof here; for a full proof, see Section 6.

The key idea is to note that we can rewrite $n\|\hat{\mu}_L\|^2$ in terms of the population centred kernel matrices $\overline{K}$, $\overline{L}$ and $\overline{M}$. Each of the resulting terms can in turn be converted to an inner product between quantities of the form $\hat{\mu} - \mu$, where $\hat{\mu}$ is an empirical estimator of μ, and each μ is a mean embedding or covariance operator.

By applying Lemma 1 to the $\hat{\mu} - \mu$, we show that most of these terms converge in probability to 0, with the residual terms equaling $n\|\hat{\mu}_{L,2}^{(Z)}\|^2$. □

As discussed in Section 1, the essential idea of this proof is novel and the resulting proof is significantly more concise than previous approaches [Chwialkowski and Gretton, 2014, Chwialkowski et al., 2014].

Theorem 1 is useful because the statistic $\|\hat{\mu}_{L,2}^{(Z)}\|^2$ is much easier to study under the non-*i.i.d.* assumption than $\|\hat{\mu}_L\|^2$. Indeed, it can expressed as a V-statistic (see Section 3.2)

$$V_n = \frac{1}{n^2}\sum_{1\le i,j\le n} \overline{\overline{k} \otimes \overline{l}} \otimes \overline{m}(S_i, S_j)$$

where $S_i = (X_i, Y_i, Z_i)$. The crucial observation is that

$$h := \overline{\overline{k} \otimes \overline{l}} \otimes \overline{m}$$

is well behaved in the following sense.

Theorem 2. *Suppose that k, l and m are bounded, symmetric, Lipschitz continuous kernels. Then h is also bounded symmetric and Lipschitz continuous, and is moreover degenerate under $\mathcal{H}_Z$ i.e $\mathbb{E}_S h(S, s) = 0$ for any fixed s.*

Algorithm 1 Test $\mathcal{H}_Z$ with Wild Bootstrap

Input: $\tilde{K}$, $\tilde{L}$, $\tilde{M}$, each size $n \times n$, N= number of bootstraps, $\alpha =$ p-value threshold

$n\|\hat{\mu}_L\|^2 = \frac{1}{n}\left(\left(\widetilde{\tilde{K} \circ \tilde{L}}\right) \circ \tilde{M}\right)_{++}$

samples = zeros(1,N)

for $i = 1$ **to** N **do**

  Draw random vector W according to Equation 6

  samples[i] = $\frac{1}{n} W^\intercal \left(\left(\widetilde{\tilde{K} \circ \tilde{L}}\right) \circ \tilde{M}\right) W$ $(*)$

end for

if sum($n\|\hat{\mu}_L\|^2 >$ samples)$> \frac{\alpha}{N}$ **then**

  Reject $\mathcal{H}_Z$

else

  Do not reject $\mathcal{H}_Z$

end if

Proof. See Section 6 □

The asymptotic analysis of such a V-statistic for non-*i.i.d.* data is still complex, but we can appeal to prior work: Leucht and Neumann [2013] showed a way to estimate any given quantile of such a V-statistic under the null hypothesis using a method called the wild bootstrap (introduced in Section 3). This, combined with analysis of the V-statistic under the alternative hypothesis provided in Theorem 2 of Chwialkowski et al. [2014][2], results in a statistical test given in Algorithm 1. The wild bootstrap is used in line $(*)$ to generate samples of the null distribution.

In Section 3 we discuss the wild bootstrap and provide results regarding consistency and Type I error control.

2.5 MULTIPLE TESTING CORRECTION

In the Lancaster test, we reject the composite null hypothesis $\mathcal{H}_0$ if and only if we reject all three of the components. In Sejdinovic et al. [2013], it is suggested that the Holm-Bonferroni correction be used to account for multiple testing [Holm, 1979]. We show here that more relaxed conditions on the p-values can be used while still bounding the Type I error, thus increasing test power.

Denote by $\mathcal{A}_*$ the event that $\mathcal{H}_*$ is rejected. Then

$$\begin{aligned}\mathbb{P}(\mathcal{A}_0) &= \mathbb{P}(\mathcal{A}_X \wedge \mathcal{A}_Y \wedge \mathcal{A}_Z) \\ &\leq \min\{\mathbb{P}(\mathcal{A}_X), \mathbb{P}(\mathcal{A}_Y), \mathbb{P}(\mathcal{A}_Z)\}\end{aligned}$$

If $\mathcal{H}_0$ is true, then so must one of the components. Without loss of generality assume that $\mathcal{H}_X$ is true. If we use significance levels of α in each test individually then $\mathbb{P}(\mathcal{A}_X) \leq \alpha$ and thus $\mathbb{P}(\mathcal{A}_0) \leq \alpha$.

[2]Note that similar results are presented in Leucht and Neumann [2013] as specific cases.

Therefore rejecting $\mathcal{H}_0$ in the event that each test has p-value less than α individually guarantees a Type I error overall of at most α. In contrast, the Holm-Bonferonni method requires that the sorted p-values be lower than $[\frac{\alpha}{3}, \frac{\alpha}{2}, \alpha]$ in order to reject the null hypothesis overall. It is therefore more conservative than necessary and thus has worse test power compared to the 'simple correction' proposed here. This is experimentally verified in Section 4.

3 THE WILD BOOTSTRAP

In this section we discuss the wild bootstrap and provide consistency and Type I error results for the proposed Lancaster test.

3.1 TEMPORAL DEPENDENCE

There are various formalisations of memory or 'mixing' of a random process [Doukhan, 1994, Bradley et al., 2005, Dedecker et al., 2007]; of relevance to this paper is the following :

Definition 1. *A process $(X_t)_t$ is β-*mixing *(also known as* absolutely regular*) if $\beta(m) \longrightarrow 0$ as $m \longrightarrow \infty$, where*

$$\beta(m) = \frac{1}{2} \sup_n \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$$

where the second supremum is taken over all finite partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of the sample space such that $A_i \in \mathcal{F}_1^n$ and $B_j \in \mathcal{F}_{n+m}^\infty$ and $\mathcal{F}_b^c = \sigma(X_b, X_{b+1}, \ldots, X_c)$

SUITABLE MIXING ASSUMPTIONS

We assume that the random process $S_i = (X_i, Y_i, Z_i)$ is β mixing with mixing coefficients satisfying $\beta(m) = o(m^{-6})$. Throughout this paper we refer to this assumption as *suitable mixing assumptions*. For a more in depth discussion of the mixing assumptions made in this paper, see Supplementary materials A.7.

3.2 V-STATISTICS

A V-statistic of a 2-argument, symmetric function h given observations $\mathcal{S}_n = \{S_1, \ldots, S_n\}$ is [Serfling, 2009]:

$$V_n = \frac{1}{n^2} \sum_{1 \leq i,j \leq n} h(S_i, S_j)$$

We call nV_n a *normalised* V-statistic. We call h the *core* of V and we say that h is *degenerate* if, for any s_1, $\mathbb{E}_{S_2 \sim \mathbb{P}}[h(s_1, S_2)] = 0$, in which case we say that V is a *degenerate V-statistic*. Many kernel test statistics can be viewed as normalised V-statistics which, under the null hypothesis, are degenerate. As mentioned in the previous section, $\|\hat{\mu}_{L,2}^{(Z)}\|^2$ is a V-statistic. Theorems 1 and 2 together imply that, under $\mathcal{H}_Z$, it can be treated as a degenerate V-statistic.

3.3 WILD BOOTSTRAP

If the test statistic has the form of a normalised V-statistic, then provided certain extra conditions are met, the wild bootstrap of Leucht and Neumann [2013] is a method to directly resample the test statistic under the null hypothesis. These conditions can be categorised as concerning: (1) appropriate mixing of the process from which our observations are drawn; (2) the core of the V-statistic.

The condition on the core that is of crucial importance to this paper is that it must be degenerate. Theorem 2 justifies our use of the wild bootstrap in the Lancaster interaction test.

Given the statistic nV_n, Leucht and Neumann [2013] tells us that a random vector W of length n can be drawn such that the bootstrapped statistic[3]

$$nV_b = \frac{1}{n}\sum_{i,j} W_i h(S_i, S_j) W_j$$

is distributed according to the null distribution of nV_n.

By generating many such W and calculating nV_b for each, we can estimate the quantiles of nV.

3.4 GENERATING W

The process generating W must satisfy conditions (B2) given on page 6 of Leucht and Neumann [2013] for nV_b to correctly resample from the null distribution of nV_n. For brevity, we provide here only an example of such a process; the interested reader should consult Leucht and Neumann [2013] or Appedix A of Chwialkowski et al. [2014] for a more detailed discussion of the bootstrapping process. The following bootstrapping process was used in the experiments in Section 4:

$$W_t = e^{-1/l_n} W_{t-1} + \sqrt{1 - e^{-2/l_n}}\epsilon_t \tag{6}$$

where $W_1, \epsilon_1, \dots, \epsilon_t$ are independent $\mathcal{N}(0,1)$ random variables. l_n should be taken from a sequence $\{l_n\}$ such that $\lim_{n \longrightarrow \infty} l_n = \infty$; in practice we used $l_n = 20$ for all of the experiments since the values of n were roughly comparable in each case.

3.5 CONTROL OF TYPE I ERROR

The following theorem shows that by estimating the quantiles of the wild bootstrapped statistic nV_b we correctly control the Type I error when testing $\mathcal{H}_Z$.

Theorem 3. *Suppose that $(X_i, Y_i, Z_i)_{i=1}^n$ are drawn from a random process satisfying suitable mixing conditions, and that W is drawn from a process satisfying (B2) in Leucht and Neumann [2013]. Then asymptotically, the quantiles of*

$$nV_b = \frac{1}{n} W^{\intercal} \left(\overline{(\bar{K} \circ \bar{L})} \circ \bar{M} \right) W$$

converge to those of $n\|\hat{\mu}_L\|^2$.

Proof. See Supplementary material A.3 □

3.6 (SEMI-)CONSISTENCY OF TESTING PROCEDURE

Note that in order to achieve consistency for this test, we would need that $\mathcal{H}_0 \iff \Delta_L P = 0$. Unfortunately this does not hold - in Sejdinovic et al. [2013] examples are given of distributions for which $\mathcal{H}_0$ is false, and yet $\Delta_L P = 0$.

However, the following result does hold:

Theorem 4. *Suppose that $\Delta_L P \neq 0$. Then as $n \longrightarrow \infty$, the probability of correctly rejecting $\mathcal{H}_0$ converges to 1.*

Proof. See Supplementary material A.4 □

At the time of writing, a characterisation of distributions for which $\mathcal{H}_0$ is false yet $\Delta_L P = 0$ is unknown. Therefore, if we reject $\mathcal{H}_0$ then we conclude that the distribution does not factorise; if we fail to reject $\mathcal{H}_0$ then we cannot conclude that the distribution factorises.

4 EXPERIMENTS

The Lancaster test described above amounts to a method to test each of the sub-hypotheses $\mathcal{H}_X, \mathcal{H}_Y, \mathcal{H}_Z$. Rather than using the Lancaster test statistic with wild bootstrap to test each of these, we could instead use HSIC. For example, by considering the pair of variables (X, Y) and Z with kernels $k \otimes l$ and m respectively, HSIC can be used to test $\mathcal{H}_Z$. Similar grouping of the variables can be used to test $\mathcal{H}_X$ and $\mathcal{H}_Y$. Applying the same multiple testing correction as in the Lancaster test, we derive an alternative test of dependence between three variables. We refer to this HSIC based procedure as *3-way HSIC*.

In the case of *i.i.d.* observations, it was shown in Sejdinovic et al. [2013] that Lancaster statistical test is more sensitive to dependence between three random variables than the above HSIC-based test when pairwise interaction is weak but joint interaction is strong. In this section, we demonstrate that the same is true in the time series case on synthetic data.

[3] Note that for fixed $\mathcal{S}_n$, nV_b is a random variable through the randomness introduced by W

4.1 WEAK PAIRWISE INTERACTION, STRONG JOINT INTERACTION

This experiment demonstrates that the Lancaster test has greater power than 3-way HSIC when the pairwise interaction is weak, but joint interaction is strong.

Synthetic data were generated from autoregressive processes X, Y and Z according to:

$$X_t = \frac{1}{2}X_{t-1} + \epsilon_t$$
$$Y_t = \frac{1}{2}Y_{t-1} + \eta_t$$
$$Z_t = \frac{1}{2}Z_{t-1} + d|\theta_t|\text{sign}(X_tY_t) + \zeta_t$$

where $X_0, Y_0, Z_0, \epsilon_t, \eta_t, \theta_t$ and ζ_t are *i.i.d.* $\mathcal{N}(0,1)$ random variables and $d \in \mathbb{R}$, called the *dependence* coefficient, determines the extent to which the process $(Z_t)_t$ is dependent on $(X_t, Y_t)_t$.

Data were generated with varying values of d. For each value of d, 300 datasets were generated, each consisting of 1200 consecutive observations of the variables. Gaussian kernels with bandwidth parameter 1 were used on each variable, and 250 bootstrapping procedures were used for each test on each dataset.

Observe that the random variables are pairwise independent but jointly dependent. Both the Lancaster and 3-way HSIC tests should be able to detect the dependence and therefore reject the null hypothesis in the limit of infinite data. In the finite data regime, the value of d affects drastically how hard it is to detect the dependence. The results of this experiment are presented in Figure 1, which shows that the Lancaster test achieves very high test power with weak dependence coefficients compared to 3-way HSIC. Note also that when using the simple multiple testing correction a higher test power is achieved than with the Holm-Bonferroni correction.

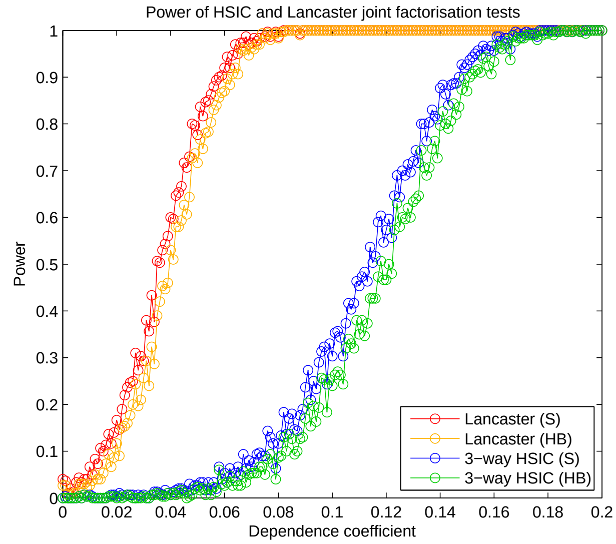

Figure 1: Results of experiment in Section 4.1. (S) refers to the simple multiple correction; (HB) refers to Holm-Bonferroni. The Lancaster test is more sensitive to dependence than 3-way HSIC, and test power for both tests is higher when using the simple correction rather than the Holm-Bonferroni multiple testing correction.

4.2 FALSE POSITIVE RATES

This experiment demonstrates that in the time series case, existing permutation bootstrap methods fail to control the Type I error, while the wild bootstrap correctly identifies test statistic thresholds and appropriately controls Type I error.

Synthetic data were generated from autoregressive processes X, Y and Z according to:

$$X_t = aX_{t-1} + \epsilon_t$$
$$Y_t = aY_{t-1} + \eta_t$$
$$Z_t = aZ_{t-1} + \zeta_t$$

where $X_0, Y_0, Z_0, \epsilon_t, \eta_t$ and ζ_t are *i.i.d.* $\mathcal{N}(0,1)$ random variables and a, called the *dependence coefficient*, determines how temporally dependent the processes are. The null hypothesis in this example is true as each process is independent of the others.

The Lancaster test was performed using both the Wild Bootstrap and the simple permutation bootstrap (used in the *i.i.d.* case) in order to sample from the null distributions of the test statistic. We used a fixed desired false positive rate $\alpha = 0.05$ with sample of size 1000, with 200 experiments run for each value of a. Figure 2 shows the false positive rates for these two methods for varying a. It shows that as the processes become more dependent, the false positive rate for the permutation method becomes very large, and is not bounded by the fixed α, whereas the false positive rate for the Wild Bootstrap method is bounded by α.

4.3 STRONG PAIRWISE INTERACTION

This experiment demonstrates a limitation of the Lancaster test. When pairwise interaction is strong, 3-way HSIC has greater test power than Lancaster.

Synthetic data were generated from autoregressive pro-

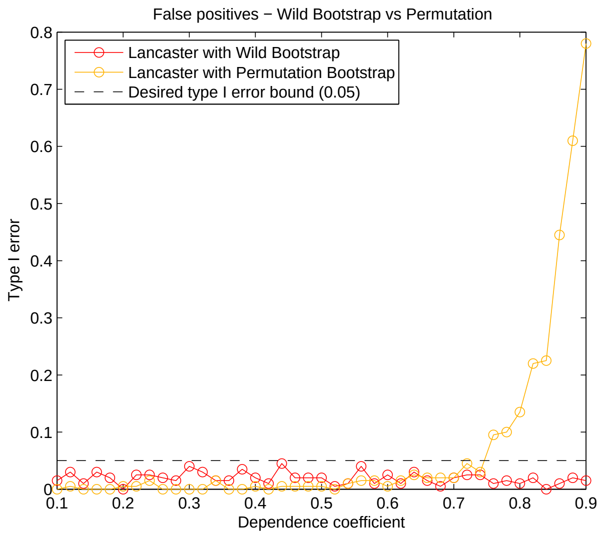

Figure 2: Results of experiment in section 4.2. Whereas the wild bootstrap succeeds in controlling the Type I error across all values of the dependence coefficient, the permutation bootstrap fails to control the Type I error as it does not sample from the correct null distribution as temporal dependence between samples increases.

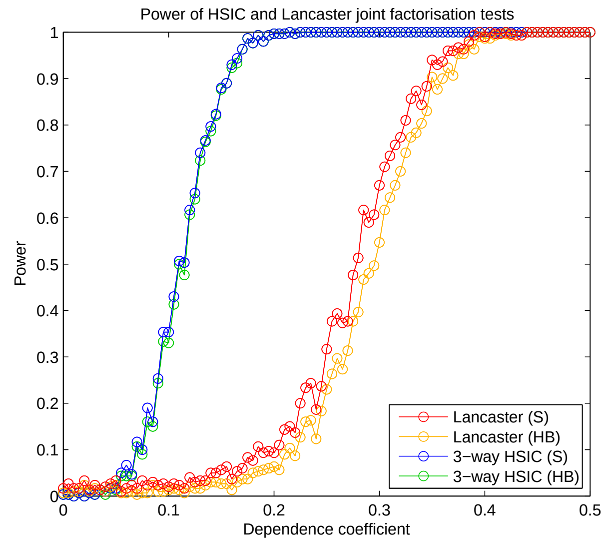

Figure 3: Results of experiment in Section 4.3. (S) refers to the simple multiple correction; (HB) refers to Holm-Bonferroni. The Lancaster test is less sensitive to dependence than 3-way HSIC, and test power in both cases is higher when using the simple correction rather than the Holm-Bonferroni multiple testing correction.

cesses X, Y and Z according to:

$$
\begin{aligned}
X_t &= \frac{1}{2}X_{t-1} + \epsilon_t \\
Y_t &= \frac{1}{2}Y_{t-1} + \eta_t \\
Z_t &= \frac{1}{2}Z_{t-1} + d(X_t + Y_t) + \zeta_t
\end{aligned}
$$

where $X_0, Y_0, Z_0, \epsilon_t, \eta_t$ and ζ_t are *i.i.d.* $\mathcal{N}(0,1)$ random variables and $d \in \mathbb{R}$, called the *dependence* coefficient, determines the extent to which the process $(Z_t)_t$ is dependent on X_t and Y_t.

Data were generated with varying values for the dependence coefficient. For each value of d, 300 datasets were generated, each consisting of 1200 consecutive observations of the variables. Gaussian kernels with bandwidth parameter 1 were used on each variable, and 250 bootstrapping procedures were used for each test on each dataset.

In this case Z_t is pairwise-dependent on both of X_t and Y_t, in addition to all three variables being jointly dependent. Both the Lancaster and 3-way HSIC tests should be capable of detecting the dependence and therefore reject the null hypothesis in the limit of infinite data. The results of this experiment are presented in Figure 3, which demonstrates that in this case the 3-way HSIC test is more sensitive to the dependence than the Lancaster test.

4.4 FOREX DATA

Exchange rates between three currencies (GBP, USD, EUR) at 5 minute intervals over 7 consecutive trading days were obtained. The data were processed by taking the returns (difference between consecutive terms within each time series, $x_t^r = x_t - x_{t-1}$) which were then normalised (divided by standard deviation). We performed the Lancaster test, 3-way HSIC and pairwise HSIC on using the first 800 entries of each processed series. All tests rejected the null hypothesis. The Lancaster and 3-way HSIC tests both returned p-values of 0 for each of $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$ with 10000 bootstrapping procedures.

We then shifted one of the time series and repeated the tests (i.e. we used entries 1 to 800 of two of the processed series and entries 801 to 1600 of the third). In this case, pairwise HSIC still detected dependence between the two unshifted time series, and both Lancaster and 3-way HSIC did not reject the null hypothesis that the joint distribution factorises. The Lancaster test returned p-values of 0.2708, 0.2725 and 0.1975 for $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$ respectively. 3-way HSIC restrurn p-values of 0.3133, 0.0000 and 0.0000 respectively.

In both cases, the Lancaster test behaves as expected. Due to arbitrage, any two exchange rates should determine the third and the Lancaster test correctly identifies a joint dependence in the returns. However, when we shift one of the time series, we break the dependence between it and the other series. Lancaster correctly identifies here that the underlying distribution does factorise.

5 DISCUSSION AND FUTURE RESEARCH

We demonstrated that the Lancaster test is more sensitive than 3-way HSIC when pairwise interaction is weak, but that the opposite is true when pairwise interaction is strong. It is curious that the two tests have different strengths in this manner, particularly when considering the very similar forms of the statistics in each case. Indeed, to test $\mathcal{H}_Z$ using the Lancaster statistic, we bootstrap the following:

$$n\|\Delta_L \hat{P}\|^2 = \frac{1}{n}\left(\left(\widetilde{\tilde{K} \circ \tilde{L}}\right) \circ \tilde{M}\right)_{++}$$

while for the 3-way HSIC test we bootstrap:

$$nHSIC_b = \frac{1}{n}\left(\widetilde{(K \circ L)} \circ \tilde{M}\right)_{++}$$

These two quantities differ only in the centring of K and L, amounting to constant shifts in the respective feature spaces of the kernels k and l. This difference has the consequence of quite drastically changing the types of dependency to which each statistic is sensitive. A formal characterisation of the cases in which the Lancaster statistic is more sensitive than 3-way HSIC would be desirable.

6 PROOFS

An outline of the proof of Theorem 1 was given in Section 2; here we provide the full proof, as well as a proof of Theorem 2.

Proof. (Theorem 1)

By observing that

$$\begin{aligned} &\phi_X(X_i) - \frac{1}{n}\sum_k \phi_X(X_k) \\ =\ &(\phi_X(X_i) - \mu_X) - \frac{1}{n}\sum_k (\phi_X(X_k) - \mu_X) \\ =\ &\bar{\phi}_X(X_i) - \frac{1}{n}\sum_k \bar{\phi}_X(X_k) \end{aligned}$$

we can therefore expand $\tilde{K}$ in terms of $\bar{K}$ as

$$\begin{aligned} &\tilde{K}_{ij} \\ &= \langle \phi_X(X_i) - \frac{1}{n}\sum_k \phi_X(X_k), \phi_X(X_j) - \frac{1}{n}\sum_k \phi_X(X_k)\rangle \\ &= \langle \bar{\phi}_X(X_i) - \frac{1}{n}\sum_k \bar{\phi}_X(X_k), \bar{\phi}_X(X_j) - \frac{1}{n}\sum_k \bar{\phi}_X(X_k)\rangle \\ &= \bar{K}_{ij} - \frac{1}{n}\sum_k \bar{K}_{ik} - \frac{1}{n}\sum_k \bar{K}_{jk} + \frac{1}{n^2}\sum_{kl} \bar{K}_{kl} \end{aligned}$$

and expanding $\tilde{L}$ and $\tilde{M}$ in a similar way, we can rewrite the Lancaster test statistic as

$$\begin{aligned} n\|\hat{\mu}_L\|^2 = &\frac{1}{n}(\bar{K} \circ \bar{L} \circ \bar{M})_{++} &&- \frac{2}{n^2}((\bar{K} \circ \bar{L})\bar{M})_{++} \\ &- \frac{2}{n^2}((\bar{K} \circ \bar{M})\bar{L})_{++} &&- \frac{2}{n^2}((\bar{M} \circ \bar{L})\bar{K})_{++} \\ &+ \frac{1}{n^3}(\bar{K} \circ \bar{L})_{++}\bar{M}_{++} &&+ \frac{1}{n^3}(\bar{K} \circ \bar{M})_{++}\bar{L}_{++} \\ &+ \frac{1}{n^3}(\bar{L} \circ \bar{M})_{++}\bar{K}_{++} &&+ \frac{2}{n^3}(\bar{M}\bar{K}\bar{L})_{++} \\ &+ \frac{2}{n^3}(\bar{K}\bar{L}\bar{M})_{++} &&+ \frac{2}{n^3}(\bar{K}\bar{M}\bar{L})_{++} \\ &+ \frac{4}{n^3}tr(\bar{K}_+ \circ \bar{L}_+ \circ \bar{M}_+) &&- \frac{4}{n^4}(\bar{K}\bar{L})_{++}\bar{M}_{++} \\ &- \frac{4}{n^4}(\bar{K}\bar{M})_{++}\bar{L}_{++} &&- \frac{4}{n^4}(\bar{L}\bar{M})_{++}\bar{K}_{++} \\ &+ \frac{4}{n^5}\bar{K}_{++}\bar{L}_{++}\bar{M}_{++} \end{aligned}$$

We denote by $C_{XYZ} = \mathbb{E}_{XYZ}[\bar{\phi}_X(X) \otimes \bar{\phi}_Y(Y) \otimes \bar{\phi}_Z(Z)]$ the population centred covariance operator with empirical estimate $\bar{C}_{XYZ} = \frac{1}{n}\sum_i \bar{\phi}_X(X_i) \otimes \bar{\phi}_Y(Y_i) \otimes \bar{\phi}_Z(Z_i)$. We define similarly the quantities $C_{XY}, C_{YZX}, \ldots$ with corresponding empirical counterparts $\bar{C}_{XY}, \bar{C}_{YZX}, \ldots$ where for example $C_{YZ} = \mathbb{E}_{YZ}[\bar{\phi}_Y(Y) \otimes \bar{\phi}_Z(Z)]$

Each of the terms in the above expression for $n\|\hat{\mu}_L\|^2$ can be expressed as inner products between empirical estimates of population centred covariance operators and tensor products of mean embeddings. Rewriting them as such yields:

$$\begin{aligned} n\|\hat{\mu}_L\|^2 = &\ n\langle \bar{C}_{XYZ}, \bar{C}_{XYZ}\rangle \\ &- 2n\langle \bar{C}_{XYZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle \\ &- 2n\langle \bar{C}_{XZY}, \bar{C}_{XZ} \otimes \bar{\mu}_Y\rangle \\ &- 2n\langle \bar{C}_{YZX}, \bar{C}_{YZ} \otimes \bar{\mu}_X\rangle \\ &+ n\langle \bar{C}_{XY} \otimes \bar{\mu}_Z, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle \\ &+ n\langle \bar{C}_{XZ} \otimes \bar{\mu}_Y, \bar{C}_{XZ} \otimes \bar{\mu}_Y\rangle \\ &+ n\langle \bar{C}_{YZ} \otimes \bar{\mu}_X, \bar{C}_{YZ} \otimes \bar{\mu}_X\rangle \\ &+ 2n\langle \bar{\mu}_Z \otimes \bar{C}_{XY}, \bar{C}_{ZX} \otimes \bar{\mu}_Y\rangle \\ &\ \vdots \end{aligned}$$

$$
\begin{aligned}
&+ 2n\langle \bar{\mu}_X \otimes \bar{C}_{YZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z \rangle \\
&+ 2n\langle \bar{\mu}_X \otimes \bar{C}_{ZY}, \bar{C}_{XZ} \otimes \bar{\mu}_Y \rangle \\
&+ 4n\langle \bar{C}_{XYZ}, \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z \rangle \\
&- 4n\langle \bar{C}_{XY} \otimes \bar{\mu}_Z, \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z \rangle \\
&- 4n\langle \bar{C}_{XZ} \otimes \bar{\mu}_Y, \bar{\mu}_X \otimes \bar{\mu}_Z \otimes \bar{\mu}_Y \rangle \\
&- 4n\langle \bar{C}_{YZ} \otimes \bar{\mu}_X, \bar{\mu}_Y \otimes \bar{\mu}_Z \otimes \bar{\mu}_X \rangle \\
&+ 4n\langle \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z, \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z \rangle
\end{aligned}
$$

By assumption, $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$ and thus the expectation operator also factorises similarly. As a consequence, $C_{XYZ} = 0$. Indeed, given any $A \in \mathcal{F}_X \otimes \mathcal{F}_\mathcal{Y} \otimes \mathcal{F}_\mathcal{Z}$, we can consider A to be a bounded linear operator $\mathcal{F}_\mathcal{Z} \longrightarrow \mathcal{F}_X \otimes \mathcal{F}_\mathcal{Y}$. It follows that[4]

$$
\begin{aligned}
&\mathbb{E}_{XYZ}\langle A, \bar{C}_{XYZ}\rangle \\
&= \frac{1}{n}\sum_i \mathbb{E}_{XY}\mathbb{E}_Z\langle A, \bar{\phi}_X(X_i) \otimes \bar{\phi}_Y(Y_i) \otimes \bar{\phi}_Z(Z_i)\rangle \\
&= \frac{1}{n}\sum_i \mathbb{E}_{XY}\mathbb{E}_Z\langle \bar{\phi}_X(X_i) \otimes \bar{\phi}_Y(Y_i), A\bar{\phi}_Z(Z_i)\rangle_{\mathcal{F}_X\otimes\mathcal{F}_Y} \\
&= \frac{1}{n}\sum_i \mathbb{E}_{XY}\langle \bar{\phi}_X(X_i) \otimes \bar{\phi}_Y(Y_i), A\mathbb{E}_Z\bar{\phi}_Z(Z_i)\rangle_{\mathcal{F}_X\otimes\mathcal{F}_Y} \\
&= 0
\end{aligned}
$$

We conclude that $C_{XYZ} = \mathbb{E}_{XYZ}\bar{C}_{XYZ} = 0$.

Similarly, C_{XZY}, C_{YZX}, C_{XZ}, C_{YZ} are all 0 in their respective Hilbert spaces. Lemma 2 tells us that each subprocess of (X_i, Y_i, Z_i) satisfies the same β-mixing conditions as (X_i, Y_i, Z_i), thus by applying Lemma 1 it follows that $\|\bar{C}_{XZY}\|$, $\|\bar{C}_{YZX}\|$, $\|\bar{C}_{XZ}\|$, $\|\bar{C}_{YZ}\|$, $\|\bar{\mu}_X\|$, $\|\bar{\mu}_Y\|$, $\|\bar{\mu}_Z\| = O_P\left(n^{-\frac{1}{2}}\right)$. Therefore

$$
\begin{aligned}
n\|\hat{\mu}_L\|^2 &\xrightarrow{O_P(n^{-\frac{1}{2}})} n\langle \bar{C}_{XYZ}, \bar{C}_{XYZ}\rangle \\
&- 2n\langle \bar{C}_{XYZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle - 2n\langle \bar{C}_{XZY}, \bar{C}_{XZ} \otimes \bar{\mu}_Y\rangle \\
&= \frac{1}{n}((\bar{K} \circ \bar{L}) \circ \bar{M})_{++} \\
&- \frac{2}{n^2}((\bar{K} \circ \bar{L})\bar{M})_{++} + \frac{1}{n^3}(\bar{K} \circ \bar{L})_{++}\bar{M}_{++}
\end{aligned}
$$

since all the other terms decay at least as quickly as $O_P(\frac{1}{\sqrt{n}})$. This is shown here for $n\langle \bar{\mu}_X \otimes \bar{C}_{YZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle$; the proofs for the other terms are similar.

$$
\begin{aligned}
&n\langle \bar{\mu}_X \otimes \bar{C}_{YZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle \\
&\leq n\|\bar{\mu}_X \otimes \bar{C}_{YZ}\|\|\bar{C}_{XY} \otimes \bar{\mu}_Z\| \\
&= n\sqrt{\langle \bar{\mu}_X \otimes \bar{C}_{YZ}, \bar{\mu}_X \otimes \bar{C}_{YZ}\rangle}\sqrt{\langle \bar{C}_{XY} \otimes \bar{\mu}_Z, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle} \\
&= n\sqrt{\langle \bar{\mu}_X, \bar{\mu}_X\rangle\langle \bar{C}_{YZ}, \bar{C}_{YZ}\rangle}\sqrt{\langle \bar{C}_{XY}, \bar{C}_{XY}\rangle\langle \bar{\mu}_Z, \bar{\mu}_Z\rangle} \\
&= n\|\bar{\mu}_X\|\|\bar{C}_{YZ}\|\|\bar{C}_{XY}\|\|\bar{\mu}_Z\| \\
&= nO_P\left(\tfrac{1}{\sqrt{n}}\right)O_P\left(\tfrac{1}{\sqrt{n}}\right)O_P(1)O_P\left(\tfrac{1}{\sqrt{n}}\right) = O_P\left(\tfrac{1}{\sqrt{n}}\right)
\end{aligned}
$$

It can be shown that $\bar{K} \circ \bar{L}$ in the above expression can be replaced with $\overline{K \circ L}$ while preserving equality. That is, we can equivalently write

$$
\begin{aligned}
n\|\Delta_L\hat{P}\|^2 \longrightarrow &\frac{1}{n}((\overline{K \circ L}) \circ \bar{M})_{++} \\
&- \frac{2}{n^2}((\overline{K \circ L})\bar{M})_{++} + \frac{1}{n^3}(\overline{K \circ L})_{++}\bar{M}_{++}
\end{aligned}
$$

This is equivalent to treating $\bar{k} \otimes \bar{l}$ as a kernel on the single variable $T := (X, Y)$ and performing another recentering trick as we did at the beginning of this proof. By rewriting the above expression in terms of the operator $\bar{C}_{TZ}$ and mean embeddings μ_T and μ_Z, it can be shown by a similar argument to before that the latter two terms tend to 0 at least as $O_P(n^{-\frac{1}{2}})$, and thus, substituting for the definition of $\|\hat{\mu}_{L,2}^{(Z)}\|^2$,

$$
n\|\hat{\mu}_L\|^2 \xrightarrow{O_P(\frac{1}{\sqrt{n}})} n\|\hat{\mu}_{L,2}^{(Z)}\|^2
$$

as required. □

Proof. (Theorem 2)

Note that $\mathbb{E}_{XYZ} = \mathbb{E}_{XY}\mathbb{E}_Z$ under $\mathcal{H}_Z$. Therefore, fixing any $s_j = (x_j, y_j, z_j)$ we have that

$$
\begin{aligned}
\mathbb{E}_{S_i}h(S_i, s_j) &= \mathbb{E}_{X_iY_i}\mathbb{E}_{Z_i}\overline{\bar{k} \otimes \bar{l}} \otimes \bar{m}(S_i, s_j) \\
&= \langle \mathbb{E}_{X_iY_i}\bar{\phi}(X_i) \otimes \bar{\phi}(Y_i) - C_{XY}, \bar{\phi}(x_j) \otimes \bar{\phi}(y_j) - C_{XY}\rangle \\
&\qquad \times \langle \mathbb{E}_{Z_i}\bar{\phi}(Z_i), \bar{\phi}(z_j)\rangle \\
&= \langle 0, \bar{\phi}(x_j) \otimes \bar{\phi}(y_j) - C_{XY}\rangle \\
&\qquad \times \langle 0, \bar{\phi}(z_j)\rangle = 0
\end{aligned}
$$

Therefore h is degenerate. Symmetry follows from the symmetry of the Hilbert space inner product.

For boundedness and Lipschitz continuity, it suffices to show the two following rules for constructing new kernels from old preserve both properties (see Supplementary materials A.5 for proof):

- $k \mapsto \bar{k}$
- $(k, l) \mapsto k \otimes l$

It then follows that $h = \overline{\bar{k} \otimes \bar{l}} \otimes \bar{m}$ is bounded and Lipschitz continuous since it can be constructed from k, l and m using the two above rules. □

ORCID 0000-0003-3169-7624

[4]We can bring the $\mathbb{E}_Z$ inside the inner product in the penultimate line due to the Bochner integrability of $\bar{\phi}_Z(Z)$, which follows from the conditions required for μ_Z to exist [Steinwart and Christmann, 2008].

References

M. Besserve, N. Logothetis, and B. Schölkopf. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *NIPS*, pages 2535–2543, 2013.

R. C. Bradley et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(2):107–144, 2005.

K. Chwialkowski and A. Gretton. A kernel independence test for random processes. *arXiv preprint arXiv:1402.4501*, 2014.

K. P. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pages 3608–3616, 2014.

J. Dedecker and C. Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236, 2005.

J. Dedecker, P. Doukhan, G. Lang, L. R. J. Rafael, S. Louhichi, and C. Prieur. Weak dependence. In *Weak Dependence: With Examples and Applications*, pages 9–20. Springer, 2007.

H. Dehling, O. S. Sharipov, and M. Wendler. Bootstrap for dependent hilbert space-valued random variables with application to von mises statistics. *Journal of Multivariate Analysis*, 133:200–215, 2015.

P. Doukhan. *Mixing*. Springer, 1994.

A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61(3):419–433, 1993.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, pages 489–496, Cambridge, MA, 2008. MIT Press.

S. Gaisser, M. Ruppert, and F. Schmid. A multivariate version of hoeffding's phi-square. *Journal of Multivariate Analysis*, 101(10):2571–2586, 2010.

A. Gretton and L. Gyorfi. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423, 2010.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592, 2007.

R. Heller, Y. Heller, and M. Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

A. Kankainen and N. G. Ushakov. A consistent modification of a test for independence based on the empirical characteristic function. *Journal of Mathematical Sciencies*, 89:1582–1589, 1998.

G. Kirchgässner, J. Wolters, and U. Hassler. *Introduction to modern time series analysis*. Springer Science & Business Media, 2012.

H. O. Lancaster. *Chi-Square Distribution*. Wiley Online Library, 1969.

A. W. Ledford and J. A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1): 169–187, 1996.

A. Leucht and M. H. Neumann. Dependent wild bootstrap for degenerate u-and v-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.

R. Patra, B. Sen, and G. Szekely. On a nonparametric notion of residual and its applications. *Statist. Probab. Lett.*, 106:208–213, 2015.

D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems*, pages 1124–1132, 2013.

R. J. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

X. Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235, 2010.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *The Journal of Machine Learning Research*, 12:2389–2410, 2011.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

G. Székely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

K. Zhang, J. Peters, D. Janzing, and B. Schoelkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813, 2011.

A SUPPLEMENTARY MATERIAL

This supplementary section contains proofs omitted from the main paper and includes a proof that the HSIC statistic asymptotically satisfies the hypothesis of the Wild Bootstrap.

A.1 HILBERT SPACE RANDOM VARIABLE CLT

In this paper we exploit a Central Limit Theorem for Hilbert space valued random variables that are functions of random processes [Dehling et al., 2015]. One of the conditions required to apply this theorem concerns appropriate β-mixing of the underlying processes. This theorem is used as a black-box, and it is hoped by the authors that as further theorems concerning CLT-properties of Hilbert space random variables are developed, the conditions required of the processes may be weakened.

Proof. (Lemma 1) We exploit Theorem 1.1 from Dehling et al. [2015]. Using the language of this paper, $\bar{\phi}(X_i)$ is a 1-approximating functional of $(X_i)_i$, following straightforwardly from the definition of 1-approximating functionals given.

Since our kernels are bounded, $\exists C: \quad \|\bar{\phi}(X_i)\| < C$ and so

$$\mathbb{E}\|\bar{\phi}(X_1)\|^{2+\delta} < C^{2+\delta} < \infty \;\; \forall \delta > 0$$

Thus condition (1) is satisfied.

We can take $f_m = \bar{\phi}(X_0) \;\; \forall m$ and so achieve $a_m = 0 \;\; \forall m$, thus condition (2) is satisfied.

By assumption on the time series, condition (3) is satisfied.

Thus, by Theorem 1.1 in Dehling et al. [2015]

$$\sqrt{n}(\tilde{\mu}_X - \mu_X) \overset{n \longrightarrow \infty}{\rightsquigarrow} N$$

where N is a Hilbert space valued Gaussian random variable and convergence is in distribution. Thus

$$\|\tilde{\mu}_X - \mu_X\| = O_P(\frac{1}{\sqrt{n}})$$

□

A.2 SUB-PROCESSES OF β-MIXING PROCESSES ARE β-MIXING

Lemma 2. *Suppose that the process $(X_t, Y_t, Z_t)_t$ is β-mixing. Then any 'sub-process' is also β-mixing (for example $(X_t, Y_t)_t$ or $(X_t)_t$)*

Proof. (Lemma 2)

Let us consider $(X_t, Y_t)_t$. Let us call $\beta_{XYZ}(m)$ the coefficients for the process $(X_t, Y_t, Z_t)_t$, and $\beta_{XY}(m)$ the coefficients for the process $(X_t, Y_t)_t$.

Observe that for $A \in \sigma((X_b, Y_b), \ldots, (X_c, Y_c))$, it is the case that $A \times \mathcal{Z} \in \sigma((X_b, Y_b, Z_b), \ldots, (X_c, Y_c, Z_c))$ and $\mathbb{P}_{XY}(A) = \mathbb{P}_{XYZ}(A \times \mathcal{Z})$.

Thus

$$
\begin{aligned}
\beta_{XY}(m) &= \frac{1}{2}\sup_n \sup_{\{A_i^{XY}\},\{B_j^{XY}\}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}_{XY}(A_i^{XY} \cap B_j^{XY}) - \mathbb{P}_{XYZ}(A_i^{XY})\mathbb{P}_{XYZ}(B_j^{XY})| \\
&= \frac{1}{2}\sup_n \sup_{\{A_i^{XY}\},\{B_j^{XY}\}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}_{XYZ}((A_i^{XY} \times \mathcal{Z}) \cap (B_j^{XY} \times \mathcal{Z})) \\
&\qquad\qquad\qquad\qquad\qquad - \mathbb{P}_{XYZ}(A_i^{XY} \times \mathcal{Z})\mathbb{P}_{XYZ}(B_j^{XY} \times \mathcal{Z})| \\
&\le \frac{1}{2}\sup_n \sup_{\{A_i^{XYZ}\},\{B_j^{XYZ}\}} \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}_{XYZ}(A_i^{XYZ} \cap B_j^{XYZ}) - \mathbb{P}_{XYZ}(A_i^{XYZ})\mathbb{P}_{XYZ}(B_j^{XYZ})| \\
&= \beta_{XYZ}(m)
\end{aligned}
$$

Thus we have shown that $\beta_{XYZ}(m) \longrightarrow 0 \implies \beta_{XY}(m) \longrightarrow 0$. That is, if $(X_t, Y_t, Z_t)_t$ is β-mixing then so is $(X_t, Y_t)_t$

A similar argument holds for any other sub-process. □

A.3 CONTROL OF TYPE I ERROR

Theorem 3 shows that the quantiles of the bootstrapped statistic nV_b (which we can estimate by drawing a large number of samples) converge to those of the test statistic $\|\hat{\mu}_L\|^2$ under the null hypothesis. Therefore, we can estimate rejection thresholds to appropriately control Type I error.

Proof. (Theorem 3)

We use Theorem 3.1 from Leucht and Neumann [2013]. By assumption, condition (B2) is satisfied by the random matrix W. (A2) is satisfied due to Theorem 2. (B1) is satisfied due to the suitable mixing assumptions.

Therefore, Theorem 3.1 implies that nV_b converges in probability to the null distribution of $n\|\hat{\mu}_{L,2}^{(Z)}\|^2$. Since $n\|\mu_L\|^2$ also converges in probability to $n\|\hat{\mu}_{L,2}^{(Z)}\|^2$, it follows that nV_b converges to $n\|\mu_L\|^2$ in probability, and thus also in distribution. Convergence in distribution implies that the quantiles converge. □

A.4 SEMI-CONSISTENCY

Theorem 4 provides a consistency result: if $\Delta_L P \neq 0$, then we correctly reject $\mathcal{H}_0$ with probability 1 in the limit $n \longrightarrow \infty$.

Proof. By Theorem 2 from Chwialkowski et al. [2014], nV_b converges to some random variable with finite variance, while $n\|\hat{\mu}_L\|^2 \longrightarrow \infty$. Thus if Q_α is the α-quantile of nV_b, then $P(n\|\hat{\mu}_L\|^2 > Q_\alpha) \longrightarrow 1$ for any α. □

A.5 PROOF THAT BOUNDEDNESS AND LIPSCHITZ CONTINUITY IS PRESERVED

Recall that a kernel k defined on $\mathcal{X}$ is Lipschitz continuous iff $\exists C_k : \forall w \;\; |k(x,w) - k(x',w)| \le C_k d_{\mathcal{X}}(x,x')$ where $d_{\mathcal{X}}$ is the metric on $\mathcal{X}$ with respect to which k is Lipschitz continuous.

Claim 1. *k bounded and Lipschitz continuous $\implies$ $\bar{k}$ is bounded and Lipschitz continuous*

Proof. k bounded implies there exists B_k such that $|k(x,w)| \le B_k \; \forall x, w \in \mathcal{X}$. It follows that

$$
\begin{aligned}
|\bar{k}(x,w)| &= |k(x,w) - \mathbb{E}_X[k(X,w)] - \mathbb{E}_W[k(x,W)] + \mathbb{E}_{XW}[k(X,W)]| \\
&\le |k(x,w)| + \mathbb{E}_X|k(X,w)| + \mathbb{E}_W|k(x,W)| + \mathbb{E}_{XW}|k(X,W)| \\
&\le 4B_k
\end{aligned}
$$

And thus $\bar{k}$ is bounded. For Lipschitz continuity, observe that for any $w \in \mathcal{X}$

$$
\begin{aligned}
|\bar{k}(x,w) - \bar{k}(x',w)| &= |k(x,w) - \mathbb{E}_X[k(X,w)] - \mathbb{E}_W[k(x,W)] + \mathbb{E}_{XW}[k(X,W)] \\
&\qquad - k(x',w) + \mathbb{E}_X[k(X,w)] + \mathbb{E}_W[k(x',W)] - \mathbb{E}_{XW}[k(X,W)]| \\
&= |k(x,w) - k(x',w) + \mathbb{E}_W[k(x',W)] - \mathbb{E}_W[k(x,W)]| \\
&\leq |k(x,w) - k(x',w)| + |\mathbb{E}_W[k(x',W)] - \mathbb{E}_W[k(x,W)]| \\
&\leq |k(x,w) - k(x',w)| + \mathbb{E}_W|k(x',W) - k(x,W)| \\
&\leq 2C_k d_{\mathcal{X}}(x,x')
\end{aligned}
$$

and thus $\bar{k}$ is Lipschitz continuous.

□

Claim 2. *k and l bounded and Lipschitz continuous with respect to the metrics $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ respectively $\implies$ $k \otimes l$ is bounded and Lipschitz continuous with respect to any metric on $\mathcal{X} \times \mathcal{Y}$ equivalent to $d((x,y),(x',y')) = d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y')$*

Note that all norms on finite dimensional vector spaces are equivalent, and so if $\mathcal{X}$ and $\mathcal{Y}$ are finite dimensional vector spaces then $k \otimes l$ is Lipschitz continuous with respect to *any* norm on $\mathcal{X} \times \mathcal{Y}$

Proof. Let k and l be bounded by B_k and B_l respectively. Then

$$
\begin{aligned}
|k \otimes l\,((x,y),(w,z))\,| &= |k(x,w)l(y,z)| \\
&= |k(x,w)||l(y,z)| \\
&\leq B_k B_l
\end{aligned}
$$

Let k and l have Lipschitz constants C_k and C_l respectively. Then, for any $(w,z) \in \mathcal{X} \times \mathcal{Y}$

$$
\begin{aligned}
&|k \otimes l\,((x,y),(w,z)) - k \otimes l\,((x',y'),(w,z))\,| \\
&\quad = |k(x,w)l(y,z) - k(x',w)l(y',z)| \\
&\quad = |k(x,w)l(y,z) - k(x',w)l(y,z) + k(x',w)l(y,z) - k(x',w)l(y',z)| \\
&\quad \leq |l(y,z)||k(x,w) - k(x',w)| + |k(x',w)||l(y,z) - l(y',z)| \\
&\quad \leq B_l C_k d_{\mathcal{X}}(x,x') + B_k C_l d_{\mathcal{Y}}(y,y') \\
&\quad \leq \max(B_l C_k, B_k C_l)\ \ d((x,y),(x',y'))
\end{aligned}
$$

□

A.6 PROOF THAT HSIC CAN BE WILD BOOTSTRAPPED

Given samples $\{(X_i,Y_i)\}_{i=1}^n$, and taking all notation involving kernels and base spaces as before, the HSIC statistic is defined to be the squared RKHS distance between the empirical embeddings of the distributions $\mathbb{P}_{XY}$ and $\mathbb{P}_X\mathbb{P}_Y$:

$$
\begin{aligned}
HSIC_b &= \|\frac{1}{n}\sum_i \phi_X(X_i) \otimes \phi_Y(Y_i) - \left(\frac{1}{n}\sum_i \phi_X(X_i)\right) \otimes \left(\frac{1}{n}\sum_i \phi_Y(Y_i)\right)\|^2 \\
&= \frac{1}{n^2}(K \circ L)_{++} - \frac{2}{n^3}(KL)_{++} + \frac{1}{n^4}K_{++}L_{++} \\
&= \frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++}
\end{aligned}
$$

where the last equality can be shown easily by expanding $\tilde{K}$ (and $\tilde{L}$ similarly) as

$$\tilde{K}_{ij} = \langle \phi_X(X_i) - \frac{1}{n}\sum_k \phi_X(X_k), \phi_X(X_j) - \frac{1}{n}\sum_k \phi_X(X_k)\rangle$$
$$= K_{ij} - \frac{1}{n}\sum_k K_{ik} - \frac{1}{n}\sum_k K_{jk} + \frac{1}{n^2}\sum_{kl} K_{kl}$$

Theorem 5. *Suppose that $(X_i, Y_i)_{i=1}^n$ are drawn from a process that is β-mixing with coefficients $\beta(m)$ satisfying $\sum_{m=1}^{\infty} \beta(m)^{\frac{\delta}{2+\delta}} < \infty$ for some $\delta > 0$. Under $\mathcal{H}_0 = \{\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y\}$, $\lim_{n\to\infty}(nHSIC_b - \frac{1}{n}(\bar{K} \circ \bar{L})_{++}) = 0$ in probability.*

Similar to the case with the Lancaster statistic, $\frac{1}{n}(\bar{K} \circ \bar{L})_{++}$ is much easier to study than $nHSIC_b$ under the non-*i.i.d.* assumption. It can be written as a normalised V-statistic as:

$$nV_n = \frac{1}{n}\sum_{1\leq i,j\leq n} \bar{k} \otimes \bar{l}(S_i, S_j)$$

where $S_i = (X_i, Y_i)$. Again, the crucial observation is that

$$h = \bar{k} \otimes \bar{l}$$

is well behaved in the following sense

Theorem 6. *Suppose that k and l are bounded symmetric Lipschitz contentious kernels. Then h is also bounded symmetric and Lipschitz continuous, which is moreover degenerate under $\mathcal{H}_0$.*

Together, Theorems 5 and 6 justify use of the Wild Bootstrap in estimating the quantiles of the null distribution of the test statistic $nHSIC_b$.

Proof. (Theorem 5) We can equivalently write $HSIC_b$ as the norm of the empirically centred covariance operator, which is invariant to population centering the feature maps:

$$HSIC_b = \left\| \frac{1}{n}\sum_i \left(\phi_X(X_i) - \frac{1}{n}\sum_j \phi_X(X_j)\right) \otimes \frac{1}{n}\sum_i \left(\phi_Y(Y_i) - \frac{1}{n}\sum_j \phi_Y(Y_j)\right) \right\|^2$$
$$= \left\| \frac{1}{n}\sum_i \left(\bar{\phi}_X(X_i) - \frac{1}{n}\sum_j \bar{\phi}_X(X_j)\right) \otimes \frac{1}{n}\sum_i \left(\bar{\phi}_Y(Y_i) - \frac{1}{n}\sum_j \bar{\phi}_Y(Y_j)\right) \right\|^2$$

Expanding this, we can rewrite the above in terms of inner products involving the population centred covariance operator and the population centred mean embeddings:

$$nHSIC_b = n\|\bar{C}_{XY}\|^2 - 2n\langle \bar{C}_{XY}, \bar{\mu}_X \otimes \bar{\mu}_Y\rangle + n\|\bar{\mu}_X \otimes \bar{\mu}_Y\|^2$$

The first term in this expression can be written as $n\|\bar{C}_{XY}\|^2 = \frac{1}{n}\sum_{ij}\bar{k}(X_i, X_j)\bar{l}(Y_i, Y_j) = \frac{1}{n}\sum_{ij} h(S_i, S_j)$. We show that the remaining two terms decay to zero in probability.

By assumption, $\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ and thus the expectation operator factorises similarly. Therefore, for any $A \in HS(\mathcal{F}_Y, \mathcal{F}_X)$,

$$\begin{aligned}\mathbb{E}_{XY}\langle A, \bar{C}_{XY}\rangle &= \frac{1}{n}\sum_i \mathbb{E}_X\mathbb{E}_Y\langle A, (\phi_X(X_i)-\mu_X)\otimes(\phi_Y(Y_i)-\mu_Y)\rangle_{HS} \\ &= \frac{1}{n}\sum_i \mathbb{E}_X\mathbb{E}_Y\langle \phi_X(X_i)-\mu_X, A\,(\phi_Y(Y_i)-\mu_Y)\rangle_{\mathcal{F}_X} \\ &= \frac{1}{n}\sum_i \mathbb{E}_Y\langle \mathbb{E}_X\,(\phi_X(X_i)-\mu_X)\,, A\,(\phi_Y(Y_i)-\mu_Y)\rangle_{\mathcal{F}_X} \\ &= 0\end{aligned}$$

where the commutativity of $\mathbb{E}_X$ with the inner product in the penultimate line follows from the Bochner integrability of the quantity $\phi_X(X)-\mu_X$, which in turn follows from the conditions under which μ_X exists [Steinwart and Christmann, 2008]. It follows that $\mathbb{E}_{XY}\bar{C}_{XY} = 0$.

Thus by Lemma 1 as before, it follows that $\|\bar{C}_{XY}\|, \|\bar{\mu}_X\|, \|\bar{\mu}_Y\| = O_P(n^{-\frac{1}{2}})$.

It thus follows that the two latter quantities in the above expression for $nHSIC_b$ decay to 0 in probability.

$$\begin{aligned}n\langle\bar{C}_{XY}, \bar{\mu}_X\otimes\bar{\mu}_Y\rangle &\leq n\|\bar{C}_{XY}\|\|\bar{\mu}_X\|\|\bar{\mu}_Y\| \\ &= O_P(n^{-\frac{1}{2}})\end{aligned}$$

$$\begin{aligned}\|\bar{\mu}_X\otimes\bar{\mu}_Y\|^2 &= n\|\bar{\mu}_X\|^2\|\bar{\mu}_Y\|^2 \\ &= nO_P(n^{-2}) \\ &= O_P(n^{-1})\end{aligned}$$

It follows that $nHSIC_b \xrightarrow{O(n^{-\frac{1}{2}})}= n\|\bar{C}_{XY}\|^2 = \frac{1}{n}(\bar{K}\circ\bar{L})_{++}$, as required. □

Proof. (Theorem 6)

To show degeneracy, fix any s_i and observe that

$$\begin{aligned}\mathbb{E}_S h(s_i, S) &= \mathbb{E}_X\mathbb{E}_Y\langle\bar{\phi}(x_i), \bar{\phi}(X)\rangle\langle\bar{\phi}(y_i), \bar{\phi}(Y)\rangle \\ &= \langle\bar{\phi}(x_i), \mathbb{E}_X\bar{\phi}(X)\rangle\langle\bar{\phi}(y_i), \mathbb{E}_Y\bar{\phi}(Y)\rangle \\ &= \langle\bar{\phi}(x_i), 0\rangle\langle\bar{\phi}(y_i), 0\rangle = 0\end{aligned}$$

Symmetry is inherited from symmetry of k and l. Boundedness and Lipschitz continuity are implied by application of the claims in Section A.5.

□

A.7 DISCUSSION OF MIXING ASSUMPTIONS

Throughout this paper, there are (related) assumptions that need to be made on the random processes we consider in order to satisfy the conditions of (1) the wild bootstrap; and (2) the Hilbert space CLT. For simplicity, we wrapped up the assumptions into the single "suitable mixing assumptions". We discuss here the precise assumptions that are needed, how they relate to the suitable mixing assumptions and the applicability of the suitable mixing assumptions.

(1) For their proof of the consistency of the wild bootstrap, Leucht and Neumann [2013] invoke the notion of τ-mixing. We require that the τ-mixing coefficients $\tau(n)$ satisfy the hypothesis of their theorem, namely that $\sum_{n=1}^{\infty} n^2\tau(n) < \infty$.

Properties of τ-mixing, for example its relationship to other types of more commonly understood mixing or models that satisfy τ-mixing, are discussed in Dedecker and Prieur [2005]. In particular, under the assumption that X_i has finite pth moment for any $p > 1$, τ-mixing implies beta-mixing. Examples of systems that are τ-mixing are: causal functions of stationary sequences, iterated random functions, Markov chains and expanding maps.

(2) In order to use the Hilbert space CLT, we require that our processes are β-mixing with coefficients $\beta(n)$ satisfying $\sum_{n=1}^{\infty} \beta(n)^{\delta/(2+\delta)} < \infty$ for some $\delta > 0$ and $\sum_{n=1}^{\infty} n\beta(n) < \infty$

In particular, both (1) and (2) are satisfied by a process that is β-mixing with coefficients $\beta(n) = o(n^{-6})$ as stated in the "Suitable Mixing" section.

Many commonly studied processes satisfy the "suitable mixing" condition. In particular Corollary 3.6 of Bradley et al. [2005] states that Harris recurrent and aperiodic markov chains satisfy absolute regularity and Theorem 3.7 of Bradley et al. [2005] states that geometric ergodicity implies geometric decay of beta coefficients. Interestingly Theorem 3.3 of Bradley et al. [2005] describes situations in which a non-stationary chain mixes exponentially.

Note, however, that our novel proof idea relies on the Hilbert space CLT and so requires only assumption (2) above to be used. Therefore our proof idea could be applied to the asymptotic study of other V-statistics in the case that the processes are beta-mixing with $\beta(n) = o(n^{-3+\epsilon})$ for any $\epsilon > 0$.