A Wild Bootstrap for Degenerate Kernel Tests

Kacper Chwialkowski
Department of Computer Science
University College London
London, Gower Street, WC1E 6BT
kacper.chwialkowski@gmail.com

Dino Sejdinovic
Gatsby Computational Neuroscience Unit, UCL
17 Queen Square, London WC1N 3AR
dino.sejdinovic@gmail.com

Arthur Gretton
Gatsby Computational Neuroscience Unit, UCL
17 Queen Square, London WC1N 3AR
arthur.gretton@gmail.com

Abstract

A wild bootstrap method for nonparametric hypothesis tests based on kernel distribution embeddings is proposed. This bootstrap method is used to construct provably consistent tests that apply to random processes, for which the naive permutation-based bootstrap fails. It applies to a large group of kernel tests based on V-statistics, which are degenerate under the null hypothesis, and non-degenerate elsewhere. To illustrate this approach, we construct a two-sample test, an instantaneous independence test and a multiple lag independence test for time series. In experiments, the wild bootstrap gives strong performance on synthetic examples, on audio data, and in performance benchmarking for the Gibbs sampler. The code is available at https://github.com/kacperChwialkowski/wildBootstrap.

1 Introduction

Statistical tests based on distribution embeddings into reproducing kernel Hilbert spaces have been applied in many contexts, including two sample testing [17, 13, 27], tests of independence [15, 28, 4], tests of conditional independence [12, 28], and tests for higher order (Lancaster) interactions [21]. For these tests, consistency is guaranteed if and only if the observations are independent and identically distributed. Much real-world data fails to satisfy the i.i.d. assumption: audio signals, EEG recordings, text documents, financial time series, and samples obtained when running Markov Chain Monte Carlo, all show significant temporal dependence patterns.

The asymptotic behaviour of kernel test statistics becomes quite different when temporal dependencies exist within the samples. In recent work on independence testing using the Hilbert-Schmidt Independence Criterion (HSIC) [8], the asymptotic distribution of the statistic under the null hypothesis is obtained for a pair of independent time series, which satisfy an absolute regularity or a ϕ-mixing assumption. In this case, the null distribution is shown to be an infinite weighted sum of *dependent* χ^2-variables, as opposed to the sum of *independent* χ^2-variables obtained in the i.i.d. setting [15]. The difference in the asymptotic null distributions has important implications in practice: under the i.i.d. assumption, an empirical estimate of the null distribution can be obtained by repeatedly permuting the time indices of one of the signals. This breaks the temporal dependence within the permuted signal, which causes the test to return an elevated number of false positives, when used for testing time series. To address this problem, an alternative estimate of the null distribution is proposed in [8], where the null distribution is simulated by repeatedly *shifting* one signal relative to the other. This preserves the temporal structure within each signal, while breaking the cross-signal dependence.

A serious limitation of the shift procedure in [8] is that it is specific to the problem of independence testing: there is no obvious way to generalise it to other testing contexts. For instance, we might have two time series, with the goal of comparing their marginal distributions - this is a generalization of the two-sample setting to which the shift approach does not apply.

We note, however, that many kernel tests have a test statistic with a particular structure: the Maximum Mean Discrepancy (MMD), HSIC, and the Lancaster interaction statistic, each have empirical estimates which can be cast as normalized V-statistics, $\frac{1}{n^{m-1}} \sum_{1 \leq i_1, \ldots, i_m \leq n} h(Z_{i_1}, ..., Z_{i_m})$, where $Z_{i_1}, ..., Z_{i_m}$ are samples from a random process at the time points $\{i_1, \ldots, i_m\}$. We show that a method of external randomization known as the *wild bootstrap* may be applied [19, 24] to simulate from the null distribution. In brief, the arguments of the above sum are repeatedly multiplied by random, user-defined time series. For a test of level α, the $1 - \alpha$ quantile of the empirical distribution obtained using these perturbed statistics serves as the test threshold. This approach has the important advantage over [8] that it may be applied to *all* kernel-based tests for which V-statistics are employed, and not just for independence tests.

The main result of this paper is to show that the wild bootstrap procedure yields consistent tests for time series, i.e., tests based on the wild bootstrap have a Type I error rate (of wrongly rejecting the null hypothesis) approaching the design parameter α, and a Type II error (of wrongly accepting the null) approaching zero, as the number of samples increases. We use this result to construct a two-sample test using MMD, and an independence test using HSIC. The latter procedure is applied both to testing for instantaneous independence, and to testing for independence across multiple time lags, for which the earlier shift procedure of [8] cannot be applied.

We begin our presentation in Section 2, with a review of the τ-mixing assumption required of the time series, as well as of V-statistics (of which MMD and HSIC are instances). We also introduce the form taken by the wild bootstrap. In Section 3, we establish a general consistency result for the wild bootstrap procedure on V-statistics, which we apply to MMD and to HSIC in Section 4. Finally, in Section 5, we present a number of empirical comparisons: in the two sample case, we test for differences in audio signals with the same underlying pitch, and present a performance diagnostic for the output of a Gibbs sampler (the MCMC M.D.); in the independence case, we test for independence of two time series sharing a common variance (a characteristic of econometric models), and compare against the test of [4] in the case where dependence may occur at multiple, potentially unknown lags. Our tests outperform both the naive approach which neglects the dependence structure within the samples, and the approach of [4], when testing across multiple lags. Detailed proofs are found in the appendices.

2 Background

The main results of the paper are based around two concepts: τ-mixing [9], which describes the dependence within the time series, and V-statistics [23], which constitute our test statistics. In this section, we review these topics, and introduce the concept of wild bootstrapped V-statistics, which will be the key ingredient in our test construction.

τ-mixing. The notion of τ-mixing is used to characterise weak dependence. It is a less restrictive alternative to classical mixing coefficients, and is covered in depth in [9]. Let $\{Z_t, \mathcal{F}_t\}_{t \in \mathbb{N}}$ be a stationary sequence of integrable random variables, defined on a probability space Ω with a probability measure P and a natural filtration $\mathcal{F}_t$. The process is called τ-dependent if

$$\tau(r) = \sup_{l \in \mathbb{N}} \frac{1}{l} \sup_{r \leq i_1 \leq \ldots \leq i_l} \tau(\mathcal{F}_0, (Z_{i_1}, ..., Z_{i_l})) \overset{r \to \infty}{\longrightarrow} 0, \text{ where}$$

$$\tau(\mathcal{M}, X) = E\left(\sup_{g \in \Lambda} \left| \int g(t) P_{X|\mathcal{M}}(dt) - \int g(t) P_X(dt) \right| \right)$$

and Λ is the set of all one-Lipschitz continuous real-valued functions on the domain of X. $\tau(\mathcal{M}, X)$ can be interpreted as the minimal L_1 distance between X and X^* such that $X \overset{d}{=} X^*$ and X^* is independent of $\mathcal{M} \subset \mathcal{F}$. Furthermore, if $\mathcal{F}$ is rich enough, this X^* can be constructed, more information is provided in the Appendix B.

V-statistics. The test statistics considered in this paper are always V-statistics. Given the observations $Z = \{Z_t\}_{t=1}^n$, a V-statistic of a symmetric function h taking m arguments is given by

$$V(h, Z) = \frac{1}{n^m} \sum\nolimits_{i \in N^m} h(Z_{i_1}, ..., Z_{i_m}), \tag{1}$$

where N^m is a Cartesian power of a set $N = \{1, ..., n\}$. For simplicity, we will often drop the second argument and write simply $V(h)$.

We will refer to the function h as to the *core* of the V-statistic $V(h)$. While such functions are usually called kernels in the literature, in this paper we reserve the term kernel for positive-definite functions taking two arguments. A core h is said to be j-degenerate if for each $z_1, \ldots, z_j$ $Eh(z_1, \ldots, z_j, Z_{j+1}^*, \ldots, Z_m^*) = 0$, where $Z_{j+1}^*, \ldots, Z_m^*$ are independent copies of Z_1. If h is j-degenerate for all $j \leq m-1$, we will say that it is *canonical*. For a one-degenerate core h, we define an auxiliary function h_2, called the second component of the core, and given by $h_2(z_1, z_2) = Eh(z_1, z_2, Z_3^*, \ldots, Z_m^*)$. Finally we say that $nV(h)$ is a normalized V-statistic, and that a V-statistic with a one-degenerate core is a degenerate V-statistic. This degeneracy is common to many kernel statistics when the null hypothesis holds [13, 15, 21].

Our main results will rely on the fact that h_2 governs the asymptotic behaviour of normalized degenerate V-statistics. Unfortunately, the limiting distribution of such V-statistics is quite complicated - it is an infinite sum of *dependent* χ^2-distributed random variables, with a dependence determined by the temporal dependence structure within the process $\{Z_t\}$ and by the eigenfunctions of a certain integral operator associated with h_2 [5, 8]. Therefore, we propose a bootstrapped version of the V-statistics which will allow a consistent approximation of this difficult limiting distribution.

Bootstrapped V-statistic. We will study two versions of the bootstrapped V-statistics

$$B_{1,n}(h, Z) = \frac{1}{n^m} \sum\nolimits_{i \in N^m} W_{i_1,n} W_{i_2,n} h(Z_{i_1}, ..., Z_{i_m}), \tag{2}$$

$$B_{1,n}(h, Z) = \frac{1}{n^m} \sum\nolimits_{i \in N^m} \tilde{W}_{i_1,n} \tilde{W}_{i_2,n} h(Z_{i_1}, ..., Z_{i_m}), \tag{3}$$

where $\{W_{t,n}\}_{1 \leq t \leq n}$ is an auxiliary wild bootstrap process and $\tilde{W}_{t,n} = W_{t,n} - \frac{1}{n} \sum_{j=1}^n W_{j,n}$. This auxiliary process, proposed by [24, 19], satisfies the following assumption:

Bootstrap assumption: $\{W_{t,n}\}_{1 \leq t \leq n}$ is a row-wise strictly stationary triangular array independent of all Z_t such that $EW_{t,n} = 0$ and $\sup_n E|W_{t,n}^{2+\sigma}| < \infty$ for some $\sigma > 0$. The autocovariance of the process is given by $EW_{s,n}W_{t,n} = \rho(|s-t|/l_n)$ for some function ρ, such that $\lim_{u \to 0} \rho(u) = 1$ and $\sum_{r=1}^{n-1} \rho(|r|/l_n) = O(l_n)$. The sequence $\{l_n\}$ is taken such that $l_n = o(n)$ but $\lim_{n \to \infty} l_n = \infty$. The variables $W_{t,n}$ are τ-weakly dependent with coefficients $\tau(r) \leq C\zeta^{\frac{r}{l_n}}$ for $r = 1, ..., n$, $\zeta \in (0, 1)$ and $C \in \mathbb{R}$.

As noted in in [19, Remark 2], a simple realization of a process that satisfies this assumption is $W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}} \epsilon_t$ where $W_{0,n}$ and $\epsilon_1, \ldots, \epsilon_n$ are independent standard normal random variables. For simplicity, we will drop the index n and write W_t instead of $W_{t,n}$. A process that fulfils the *bootstrap assumption* will be called *bootstrap process*. Further discussion of the wild bootstrap is provided in the Appendix A. The versions of the bootstrapped V-statistics in (2) and (3) were previously studied in [19] for the case of canonical cores of degree $m = 2$. We extend their results to higher degree cores (common within the kernel testing framework), which are not necessarily one-degenerate. When stating a fact that applies to both B_1 and B_2, we will simply write B, and the argument h or index n will be dropped when there is no ambiguity.

3 Asymptotics of wild bootstrapped V-statistics

In this section, we present main Theorems that describe asymptotic behaviour of V-statistics, all the poofs are in the Appendix C. In the next section, these results will be used to construct kernel-based statistical tests applicable to dependent observations. Tests are constructed so that the V-statistic is degenerate under the null hypothesis and non-degenerate under the alternative. Theorem 1 guarantees that the bootstrapped V-statistic will converge to the same limiting null distribution as the simple V-statistic.

Throughout this paper we will make one mild assumption

$$\sup_{i \in N^m} Eh(Z_i)^2 < \infty,$$

where $Z_i = (Z_{i_1}, \cdots Z_{i_m})$. This assumption is almost always automatically satisfied, since most of the kernels used in practice are bounded.

Theorem 1. *Assume that the stationary process Z_t is τ-dependent with $\sum_{r=1}^{\infty} r^2 \sqrt{\tau(r)} < \infty$. If the core h is a Lipschitz continuous, one-degenerate and its h_2-component is a positive definite kernel, such that $Eh_2(Z_0, Z_0) < \infty$, then nB_n (2), (3), and nV_n (1) converge weakly to the same distribution V. Moreover $nB_n(h_2)$ and $nV_n(h_2)$ converge weakly to $\binom{m}{2}^{-1} V$.*

On the other hand, if the V-statistic is not degenerate, which is usually true under the alternative, it converges to some non-zero constant.

Theorem 2. *Assume that the stationary process Z_t is τ-dependent with $\tau(r) = o(r^{-4})$. If the core h is a Lipschitz continuous, and h_0 component is positive then V_n converges in mean squared to h_0.*

In this setting, Theorem 3 guarantees that the bootstrapped V-statistic will converge to zero in probability. This property is necessary in testing, as it implies that the test thresholds computed using the bootstrapped V-statistics will also converge to zero, and so will the corresponding Type II error.

Theorem 3. *Assume that the stationary process $\{Z_t\}$ is τ-dependent with a coefficient $\tau(r) = o(r^{-4})$. If the core h is a function of $m > 1$ arguments then $B_1(h)$ and $o(n)B_2(h)$ converge to zero in mean squared.*

Although both B_2 and B_1 converge to zero, the rate does not seem to be that same. As a consequence, tests that utilize B_2 usually give lower Type II error then the ones that use B_1. On the other hand, B_1 seems to better approximate V-statistic distribution under the null hypothesis. This agrees with our experiments in Section 5 as well as with those in [19, Section 5]). These results a sufficient for adopting kernel tests developed for i.i.d. data to tests that work on random processes. In particular Theorem 1 justifies usage of bootstraped V-statistics for estimating quantiles of the null distribution, while Theorems 23 guarantee consistency.

The general testing procedure is

- Calculate the test statistic $nV_n(h)$.
- Obtain wild bootstrap samples $\{B_n(h)\}_{i=1}^D$ and estimate the $1 - \alpha$ empirical quantile of these samples.
- If $nV_n(h)$ exceeds the quantile, reject.

4 Applications to Kernel Tests

In this section, we describe how the wild bootstrap for V-statistics can be used to construct kernel tests for independence and the two-sample problem, which are applicable to weakly dependent observations. We start by reviewing the main concepts underpinning the kernel testing framework.

For every symmetric, positive definite function, i.e., *kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is an associated reproducing kernel Hilbert space $\mathcal{H}_k$ [3, p. 19]. The kernel embedding of a probability measure P on $\mathcal{X}$ is an element $\mu_k(P) \in \mathcal{H}_k$, given by $\mu_k(P) = \int k(\cdot, x)\, dP(x)$ [3, 25]. If a measurable kernel k is bounded, the mean embedding $\mu_k(P)$ exists for all probability measures on $\mathcal{X}$, and for many interesting bounded kernels k, including the Gaussian, Laplacian and inverse multi-quadratics, the kernel embedding $P \mapsto \mu_k(P)$ is injective. Such kernels are said to be *characteristic* [26]. The RKHS-distance $\|\mu_k(P_x) - \mu_k(P_y)\|^2_{\mathcal{H}_k}$ between embeddings of two probability measures P_x and P_y is termed the Maximum Mean Discrepancy (MMD), and its empirical version serves as a popular statistic for non-parametric two-sample testing [13]. Similarly, given a sample of paired observations $\{(X_i, Y_i)\}_{i=1}^n \sim P_{xy}$, and kernels k and l respectively on X and Y domains, the RKHS-distance $\|\mu_\kappa(P_{xy}) - \mu_\kappa(P_x P_y)\|^2_{\mathcal{H}_\kappa}$ between embeddings of the joint distribution and of the product of the marginals, measures dependence between X and Y. Here, $\kappa((x, y), (x', y')) = k(x, x')l(y, y')$ is the kernel on the product space of X and Y domains. This quantity is called Hilbert-Schmidt

Independence Criterion (HSIC) [14, 15]. When characteristic RKHSs are used, the HSIC is zero iff $X \perp\!\!\!\perp Y$: this follows from [16]. The empirical statistic is written $\widehat{\mathrm{HSIC}}_\kappa = \frac{1}{n^2}\mathrm{Tr}(KHLH)$ for kernel matrices K and L and the centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$.

In this section, we describe how the wild bootstrap for V-statistics can be used to construct kernel tests for independence and the two-sample problem, in presence of weakly dependent observations.

4.1 Wild Bootstrap For MMD

Denote the observations by $\{X_i\}_{i=1}^{n_x} \sim P_x$, and $\{Y_j\}_{j=1}^{n_y} \sim P_y$. Our goal is to test the null hypothesis $\mathbf{H}_0 : P_x = P_y$ vs. the alternative $\mathbf{H}_1 : P_x \neq P_y$. In the case where samples have equal sizes, i.e., $n_x = n_y$, application of the wild bootstrap to MMD-based tests on dependent samples is straightforward: the empirical MMD can be written as a V-statistic with the core of degree two on pairs $z_i = (x_i, y_i)$ given by $h(z_1, z_2) = k(x_1, x_2) - k(x_1, y_2) - k(x_2, y_1) + k(y_1, y_2)$. It is clear that whenever k is Lipschitz continuous and bounded, so is h. Moreover, h is a valid positive definite kernel, since it can be represented as an RKHS inner product $\langle k(\cdot, x_1) - k(\cdot, y_1), k(\cdot, x_2) - k(\cdot, y_2)\rangle_{\mathcal{H}_k}$. Under the null hypothesis, h is also one-degenerate, i.e., $Eh\left((x_1, y_1), (X_2, Y_2)\right) = 0$. Therefore, we can use the bootstrapped statistics in (2) and (3) to approximate the null distribution and attain a desired test level.

When $n_x \neq n_y$, however, it is no longer possible to write the empirical MMD as a one-sample V-statistic. We will therefore require the following bootstrapped version of MMD

$$\widehat{\mathrm{MMD}}_{k,b} = \frac{1}{n_x^2}\sum_{i=1}^{n_x}\sum_{j=1}^{n_x}\tilde{W}_i^{(x)}\tilde{W}_j^{(x)}k(x_i, x_j) - \frac{1}{n_x^2}\sum_{i=1}^{n_y}\sum_{j=1}^{n_y}\tilde{W}_i^{(y)}\tilde{W}_j^{(y)}k(y_i, y_j) - \frac{2}{n_x n_y}\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\tilde{W}_i^{(x)}\tilde{W}_j^{(y)}k(x_i, y_j), \tag{4}$$

where $\tilde{W}_t^{(x)} = W_t^{(x)} - \frac{1}{n_x}\sum_{i=1}^{n_x} W_i^{(x)}$, $\tilde{W}_t^{(y)} = W_t^{(y)} - \frac{1}{n_y}\sum_{j=1}^{n_y} W_j^{(y)}$; $\{W_t^{(x)}\}$ and $\{W_t^{(y)}\}$ are two auxiliary wild bootstrap processes that are independent of $\{X_t\}$ and $\{Y_t\}$ and also independent of each other, both satisfying the bootstrap assumption in Section 2. The following Proposition shows that the bootstrapped statistic has the same asymptotic null distribution as the empirical MMD. The proof follows that of [19, Theorem 3.1], and is given in the Appendix.

Proposition 1. *Let k be bounded and Lipschitz continuous, and let $\{X_t\}$ and $\{Y_t\}$ both be τ-dependent with coefficients $\tau(r) = O(r^{-6-\epsilon})$, but independent of each other. Further, let $n_x = \rho_x n$ and $n_y = \rho_y n$ where $n = n_x + n_y$. Then, under the null hypothesis $P_x = P_y$, $\varphi\left(\rho_x\rho_y n\widehat{MMD}_k, \rho_x\rho_y n\widehat{MMD}_{k,b}\right) \to 0$ in probability as $n \to \infty$, where φ is the Prokhorov metric and $\widehat{MMD}_k$ is the MMD between empirical measures.*

4.2 Wild Bootstrap For HSIC

Using HSIC in the context of random processes is not new in the machine learning literature. For a 1-approximating functional of an absolutely regular process [6], convergence in probability of the empirical HSIC to its population value was shown in [29]. No asymptotic distributions were obtained, however, nor was a statistical test constructed. The asymptotics of a normalized V-statistic were obtained in [8] for absolutely regular and ϕ-mixing processes [11]. Due to the intractability of the null distribution for the test statistic, the authors propose a procedure to approximate its null distribution using circular shifts of the observations leading to tests of instantaneous independence, i.e., of $X_t \perp\!\!\!\perp Y_t$, $\forall t$. This was shown to be consistent under the null (i.e., leading to the correct Type I error), however consistency of the shift procedure under the alternative is a challenging open question (see [8, Section A.2] for further discussion). In contrast, as shown below in Propositions 2 and **??** (which are direct consequences of the Theorems 1 and 3), the wild bootstrap guarantees test consistency under both hypotheses: null and alternative, which is a major advantage. In addition, the wild bootstrap can be used in constructing a test for the harder problem of determining independence across multiple lags simultaneously, similar to the one in [4].

Following symmetrisation, it is shown in [15, 8] that the empirical HSIC can be written as a degree four V-statistic with core given by

$$h(z_1, z_2, z_3, z_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)})[l(y_{\pi(1)}, y_{\pi(2)}) + l(y_{\pi(3)}, y_{\pi(4)}) - 2l(y_{\pi(2)}, y_{\pi(3)})],$$

where we denote by S_n the group of permutations over n elements. One-degeneracy of the core under the null hypothesis was stated in [15, Theorem 2], [15, Section A.2, following eq. (11)] shows that h_2 is a kernel; $h_0 \geq 0$ follows from the fact that HSIC is a distance. Using Theorems 1,3,2 we can construct an independence test using h. Drawback of this test, when implemented in the most straightforward way, is its quadruple computational complexity. To achieve quadratic time complexity, that matches time complexity of HSIC test for i.i.d. data, we modify our bootstrapped statistic.

Quadratic time HSIC. In this section we assume that kernels k, l are positive and bounded. We define empirical mean embedding $\tilde{\mu}_X(x) = \frac{1}{n}\sum_i^n k(x, X_i)$ and centred kernels

$$\begin{aligned}
\bar{k}(x, x') =& k(x', x) - Ek(x, X) - Ek(X', x') + Ek(X, X') \\
=& \langle k(x, \cdot) - \mu_X, k(x', \cdot) - \mu_X \rangle. \\
\tilde{k}(x, x') =& k(x, x') - \frac{1}{n}\sum_i^n k(x, X_i) - \frac{1}{n}\sum_i^n k(x', X_i) + \frac{1}{n^2}\sum_{i,j}^n k(X_j, X_i) \\
=& \langle k(x, \cdot) - \tilde{\mu}_X, k(x', \cdot) - \tilde{\mu}_X \rangle.
\end{aligned}$$

where X, X' are i.i.d. copies of X_1. Same definitions hold for the kernel l. Let Q_i denote W_i or $\tilde{W}_i$ (where it is necessary, we check claims for both W_i and $\tilde{W}_i$ separately). We further define

$$S_n = \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i(\phi(X_i) - \tilde{\mu}_X) \otimes (\phi(Y_i) - \tilde{\mu}_Y), \tag{5}$$

$$T_n = \frac{1}{\sqrt{n}} \sum_{i \in N} Q_i(\phi(X_i) - \mu_X) \otimes (\phi(Y_i) - \mu_Y). \tag{6}$$

First, we relate T_n to $B(h_2)$.

Statement 1. *[15, section A.2, following eq. (11)] The second component of h is $h_2(z_1, z_2) = \frac{1}{6}\bar{k}(x_1, x_2)\bar{l}(y_1, y_2)$.*

Lemma 1. *Squared norm of T_n is equal to $6B(h_2)$.*

Proof.

$$\begin{aligned}
\|T_n\|^2 =& \frac{1}{n} \sum_{i,j \in N} Q_i Q_j \Big\langle (\phi(X_i) - \mu_X) \otimes (\phi(Y_i) - \mu_Y), (\phi(X_j) - \mu_X) \otimes (\phi(Y_j) - \mu_Y) \Big\rangle \\
=& \frac{1}{n} \sum_{i,j \in N} Q_i Q_j \bar{k}(X_i, X_j)\bar{l}(Y_i, Y_j) \\
=& 6B(h_2).
\end{aligned}$$

□

Next we relate S_n to T_n – we show that the difference between them is asymptotically negligible. We start with a technical lemma.

Lemma 2. *If $(\bar{k} \times \bar{k}, Z_i)$ is of type Δ of order $O(r^{-4})$ (see Definition 2), then*

$$\lim_{n \to \infty} E \left\| \sqrt{n}(\tilde{\mu}_X - \mu_X) \right\|^4 = O(1).$$

Proof.

$$E\left\|\sqrt{n}(\tilde{\mu}_X - \mu_X)\right\|^4 = E\left\|\frac{1}{\sqrt{n}}\sum_{i\in N}\phi(X_i) - \mu_X\right\|^4$$
$$= E\left(\frac{1}{n}\sum_{i\in N}\langle\phi(X_j) - \mu_X, \phi(X_i) - \mu_X\rangle\right)^2$$
$$= \frac{1}{n^2}E\sum_{i\in N^4}\bar{k}\times\bar{k}(Z_i).$$

Since $(\bar{k}\times\bar{k}, X_i)$ is of type Δ, by Lemma 4, the expected value is of order $O(1)$. □

Lemma 3. *If $(\bar{k}\times\bar{k}, Z_i)$, $(\bar{l}\times\bar{l}, Z_i)$ are of type Δ of order $O(r^{-4})$, then, under the null, $\|S_n\|^2 - \|T_n\|^2$ converges to zero in mean square. Under the alternative $\frac{1}{n}(\|S_n\|^2 - \|T_n\|^2)$ converges to zero in mean square.*

Proof. We first show that $E\|S_n - T_n\|^2 \to 0$ both under the null and the alternative. Then, using the fact that $\|T_n\|^2 < \infty$ under the null and $\frac{1}{n}\|T_n\|^2 < \infty$ under alternative we will conclude the proof. The difference $S_n - T_n$ is

$$\frac{1}{\sqrt{n}}\sum_{i\in N}Q_i\Big[(\phi(X_i) - \tilde{\mu}_X)\otimes(\phi(Y_i) - \tilde{\mu}_Y) - (\phi(X_i) - \mu_X)\otimes(\phi(Y_i) - \mu_Y)\Big]$$
$$= \frac{1}{\sqrt{n}}\sum_{i\in N}Q_i\Big[\phi(X_i)\otimes\mu_Y - \phi(X_i)\otimes\tilde{\mu}_Y\Big]$$
$$+ \frac{1}{\sqrt{n}}\sum_{i\in N}Q_i\Big[\phi(Y_i)\otimes\mu_X - \phi(Y_i)\otimes\tilde{\mu}_X\Big]$$
$$+ \frac{1}{\sqrt{n}}\sum_{i\in N}Q_i(\tilde{\mu}_X\otimes\tilde{\mu}_Y - \mu_Y\otimes\mu_X).$$

We examine differences separately – it is sufficient to show that each difference converges to zero in mean square.

The expected norm of the first difference is

$$E\left\|\frac{1}{\sqrt{n}}\sum_{i\in N}Q_i\Big[\phi(X_i)\otimes\mu_Y - \phi(X_i)\otimes\tilde{\mu}_Y\Big]\right\|^2$$
$$= E\left\|\sqrt{n}(\mu_Y - \tilde{\mu}_Y)\otimes\frac{1}{\sqrt{n}}\sum_{i\in N}Q_i\phi(X_i)\right\|^2$$
$$\leq \sqrt{E\left\|\sqrt{n}(\tilde{\mu}_Y - \mu_Y)\right\|^4 E\left\|\frac{1}{\sqrt{n}}\sum_{i\in N}Q_i\phi(X_i)\right\|^4}.$$

We used $\|v\otimes u\| = \|v\|\|u\|$ and Cauchy-Schwarz inequality. By Lemma 2 the first term is $O(1)$. The second term is equal to

$$E\|\frac{1}{n}\sum_{i\in N}Q_i\phi(X_i)\|^4 = E\left(\frac{1}{n^2}\sum_{i,j}k(X_i, X_j)Q_iQ_j\right)^2.$$

The expected value converges to zero in mean square by Lemma 4 (the assumption $\sup_{i,j}k(X_i, X_j) < \infty$ is satisfied). Using similar reasoning, the second term

$$E\left\|\frac{1}{\sqrt{n}}\sum_{i\in N}Q_i\Big[\phi(Y_i)\otimes\tilde{\mu}_X - \phi(Y_i)\otimes\mu_X\Big]\right\|^2$$

also converges to zero. The final term is

$$E\left\|\frac{1}{\sqrt{n}}\sum_{i\in N}Q_i(\tilde{\mu}_X\otimes\tilde{\mu}_Y-\mu_Y\otimes\mu_X)\right\|^2$$
$$=E\left|\frac{1}{n}\sum_{i\in N}Q_i\right|E\left\|\sqrt{n}(\tilde{\mu}_X\otimes\tilde{\mu}_Y-\mu_Y\otimes\mu_X)\right\|^2$$

$\frac{1}{n}\sum_{i\in N}Q_i$ converges in mean square to zero (Lemmas 6, 7). We rewrite the second term

$$E\left\|\sqrt{n}(\tilde{\mu}_X\otimes\tilde{\mu}_Y-\tilde{\mu}_Y\otimes\mu_X+\tilde{\mu}_Y\otimes\mu_X-\mu_Y\otimes\mu_X)\right\|^2$$

It is sufficient to bound

$$E\left\|\sqrt{n}\tilde{\mu}_Y\otimes(\tilde{\mu}_X-\mu_X)\right\|^2\leq E\sqrt{\left\|\tilde{\mu}_Y\right\|^4E\left\|\sqrt{n}(\tilde{\mu}_X-\mu_X)\right\|^4}$$
$$E\left\|\sqrt{n}\mu_X\otimes(\tilde{\mu}_Y-\mu_Y)\right\|^2=\left\|\mu_X\right\|^2E\left\|\sqrt{n}(\tilde{\mu}_Y-\mu_Y)\right\|^2$$

$E\|\tilde{\mu}_Y\|^4=E\frac{1}{n^4}\sum_{i\in N^4}(l\times l)(Y_i)=O(1)$, since l is bounded. By Lemma 2 $E\left\|\sqrt{n}(\tilde{\mu}_X-\mu_X)\right\|^4$ and $E\left\|\sqrt{n}(\tilde{\mu}_Y-\mu_Y)\right\|^2$ are finite. Thus, to whole expression converges to zero. We proved that T_n-S_n converges in mean square to zero. We have

$$E|\|T_n\|^2-\|S_n\|^2|\leq E\big|\|T_n\|-\|S_n\|\big|\big|\|T_n\|+\|S_n\|\big|$$
$$\leq\sqrt{E\big|\|T_n\|-\|S_n\|\big|^2E\big|\|T_n\|+\|S_n\|\big|^2}$$

To show that the above expression converges to zero it is sufficient to show that $E\|T_n\|^2<\infty$ and $E\|S_n\|^2<\infty$. Under the null hypothesis, by Lemma 4, expected value of $E\|T_n\|^2=nB_n(h_2)$ is finite. Since $E\|T_n-S_n\|^2\to 0$ we also have $E\|T_n-S_n\|\to 0$. $E\|S_n\|$ is therefore finite since

$$E\|S_n\|=E\|S_n-T_n+T_n\|\leq E\|S_n-T_n\|+E\|T_n\|<\infty$$

Therefore we have

$$E\|S_n\|^2\leq E\|S_n-T_n+T_n\|^2$$
$$\leq E\|S_n-T_n\|^2+E\|T_n-S_n\|E\|T_n\|+E\|T_n\|^2<\infty$$

Under the alternative we have

$$n^{-1}E|\|T_n\|^2-\|S_n\|^2|\leq n^{-1}E\big|\|T_n\|-\|S_n\|\big|\big|\|T_n\|+\|S_n\|\big|$$
$$\leq\sqrt{E\big|\|T_n\|-\|S_n\|\big|^2n^{-1}E\big|\|T_n\|+\|S_n\|\big|^2}$$

it is sufficient to show that $n^{-1}E\|T_n\|^2<\infty$ and $n^{-1}E\|S_n\|^2<\infty$. By Theorem 3, $n^{-1}E\|T_n\|^2<\infty$ is finite and, using the reasoning similar to the one above, we have that $n^{-1}E\|S_n\|^2<\infty$. □

This shows that we can use squared norm of S_n as a bootstrapped test statistic. For HSIC we redefine B_n

$$B_n^*:=\|S_n\|^2=\frac{1}{n}\sum_{i,j\in N}Q_i,Q_j\tilde{k}(X_i,X_j)\tilde{l}(X_i,X_j).\tag{7}$$

B_1^* corresponds to $Q_i = W_i$, B_2^* corresponds to $Q_i = \tilde{W}_i$. This bootstrapped statistic interestingly coincides with $V_n(h)$. [15] showed that

$$V_n(h) = \frac{1}{n} \sum_{i,j \in N} \tilde{k}(X_i, X_j) \tilde{l}(X_i, X_j). \tag{8}$$

Finally, notice that both statistics 7 and 8 can be calculated in quadratic time.

Proposition 2. *Let $Z_t = (X_t, Y_t)$ be a stationary process that is τ-dependent such that $\sum_{r=1}^{\infty} r^2 \sqrt{\tau(r)} < \infty$. Under the null hypothesis B_n^* (7) and $nV_n(h)$ (8)converge weakly to the same distribution. Under the alternative hypothesis B_n^* converges to zero in probability, while $V_n(h)$ converges to a positive constant.*

Proof. We calculate

$$nV_n(h) - B_n^* = nV_n(h) - 6nB_n(h_2) + 6nB_n(h_2) - B_n^*.$$

By Lemma 1, $6nB_n(h_2) = \|T_n\|^2$. By definition (7), $B_n^* = \|S_n\|^2$. By Lemma 3, $6nB_n(h_2) - B_n^*$ converges to zero in mean square. We check assumptions; since process Z_t is τ-mixing (of order $o(r^{-4})$) and both $\bar{k}$, $\bar{l}$ are canonical, Lemma 11 guarantees that $(\bar{k}, Z_i)$, $(\bar{l}, Z_i)$ are of type Δ of order $O(r^{-4})$.

Under the null hypothesis, by Theorem 1, $nV_n(h) - 6nB_n(h_2)$ converges to zero. We check assumptions; by Lemma 1, h_2 is a symmetric, one-degenerate, bounded kernel, assumptions concerning τ-mixing are satisfied.

Under the alternative, by Theorem 3 and Lemma 3 respectively, $6B_n(h_2)$ and $\frac{1}{n}B_n^* - 6B_n(h_2)$ converge to zero in mean square. By Theorem 3, $V_n(h)$ converges to a positive constant. □

We consider two types of tests: instantaneous independence and independence at multiple time lags.

Test of instantaneous independence Here, the null hypothesis $\mathbf{H_0}$ is that X_t and Y_t are independent at all times t, and the alternative hypothesis $\mathbf{H_1}$ is that they are dependent. We use Proposition2 directly to bootstrap an appropriate quantile and compare it with a test statistic.

Lag-HSIC Proposition 2 allows us to construct a test of time series independence that is similar to one designed by [4]. Here, we will be testing against a broader null hypothesis: X_t and $Y_{t'}$ are independent for $|t - t'| < M$ for an arbitrary large but fixed M.

Since the time series $Z_t = (X_t, Y_t)$ is stationary, it suffices to check whether there exists a dependency between X_t and Y_{t+m} for $-M \leq m \leq M$. Since each lag corresponds to an individual hypothesis, we will require a Bonferroni correction to attain a desired test level α. We therefore define $q = 1 - \frac{\alpha}{2M+1}$. The shifted time series will be denoted $Z_t^m = (X_t, Y_{t+m})$. Let $S_{m,n} = nV_n(h, Z^m)$ denote the value of the normalized HSIC statistic calculated on the shifted process Z_t^m. Let $F_{b,n}$ denote the empirical cumulative distribution function obtained by the bootstrap procedure using B_n^* (7). The test will then reject the null hypothesis if the event $\mathcal{A}_n = \left\{ \max_{-M \leq m \leq M} S_{m,n} > F_{b,n}^{-1}(q) \right\}$ occurs. By a simple application of the union bound, it is clear that the asymptotic probability of the Type I error will be $\lim_{n \to \infty} P_{\mathbf{H_0}}(\mathcal{A}_n) \leq \alpha$. On the other hand, if the alternative holds, there exists some m with $|m| \leq M$ for which $V_n(h, Z^m) = n^{-1} S_{m,n}$ converges to a non-zero constant. In this case

$$P_{\mathbf{H_1}}(\mathcal{A}_n) \geq P_{\mathbf{H_1}}(S_{m,n} > F_{b,n}^{-1}(q)) = P_{\mathbf{H_1}}(n^{-1} S_{m,n} > n^{-1} F_{b,n}^{-1}(q)) \to 1 \tag{9}$$

as long as $n^{-1} F_{b,n}^{-1}(q) \to 0$, which follows from the convergence of B_n^* (7) to zero in probability shown in Proposition 2. Therefore, the Type II error of the multiple lag test is guaranteed to converge to zero as the sample size increases. Our experiments in the next Section demonstrate that while this procedure is defined over a finite range of lags, it results in tests more powerful than the procedure for an infinite number of lags proposed in [4]. We note that a procedure that works for an infinite number of lags, although possible to construct, does not add much practical value under the present assumptions. Indeed, since the τ-mixing assumption applies to the joint sequence $Z_t = (X_t, Y_t)$, dependence between X_t and Y_{t+m} is bound to disappear at a rate of $o(m^{-6})$, i.e., the variables both within and across the two series are assumed to become gradually independent at large lags.

Table 1: Rejection rates for two-sample experiments. **MCMC**: sample size=500; a Gaussian kernel with bandwidth $\sigma = 1.7$ is used; every second Gibbs sample is kept (i.e., after a pass through both dimensions). **Audio**: sample sizes are $(n_x, n_y) = \{(300, 200), (600, 400), (900, 600)\}$; a Gaussian kernel with bandwidth $\sigma = 14$ is used. **Both**: wild bootstrap uses blocksize of $l_n = 20$; averaged over at least 200 trials. The Type II error for all tests was zero

	experiment \ method	permutation	$\widehat{\text{MMD}}_{k,b}$	V_{b1}	V_{b2}
MCMC	i.i.d. vs i.i.d. ($\mathbf{H}_0$)	.040	.025	.012	.070
	i.i.d. vs Gibbs ($\mathbf{H}_0$)	.528	.100	.052	.105
	Gibbs vs Gibbs ($\mathbf{H}_0$)	.680	.110	.060	.100
Audio	$\mathbf{H}_0$	{.970,.965,.995}	{.145,.120,.114}		
	$\mathbf{H}_1$	{1,1,1}	{.600,.898,.995}		

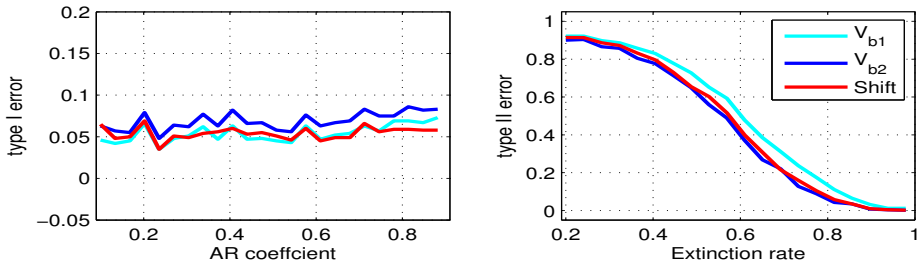

Figure 1: Comparison of Shift-HSIC and tests based on V_{b1} and V_{b2}. The left panel shows the performance under the null hypothesis, where a larger AR coefficient implies a stronger temporal dependence. The right panel show the performance under the alternative hypothesis, where a larger extinction rate implies a greater dependence between processes.

5 Experiments

The MCMC M.D. We employ MMD in order to diagnose how far an MCMC chain is from its stationary distribution [22, Section 5], by comparing the MCMC sample to a benchmark sample. A hypothesis test of whether the sampler has converged based on the standard permutation-based bootstrap leads to too many rejections of the null hypothesis, due to dependence within the chain. Thus, one would require heavily thinned chains, which is wasteful of samples and computationally burdensome. Our experiments indicate that the wild bootstrap approach allows consistent tests directly on the chains, as it attains a desired number of false positives.

To assess performance of the wild bootstrap in determining MCMC convergence, we consider the situation where samples $\{X_i\}$ and $\{Y_i\}$ are bivariate, and both have the identical marginal distribution given by an elongated normal $P = \mathcal{N}\left([\ 0 \quad 0\ ], \begin{bmatrix} 15.5 & 14.5 \\ 14.5 & 15.5 \end{bmatrix}\right)$. However, they could have arisen either as independent samples, or as outputs of the Gibbs sampler with stationary distri-

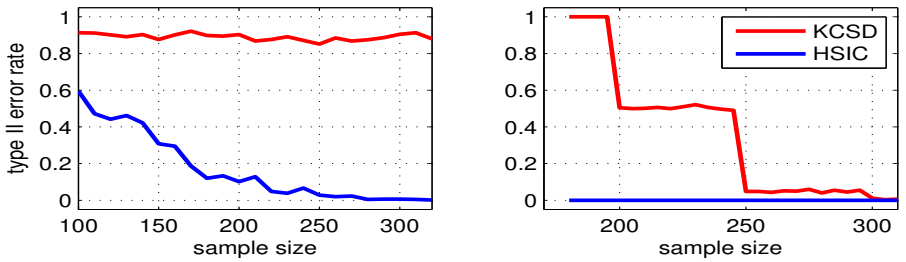

Figure 2: In both panel Type II error is plotted. The left panel presents the error of the lag-HSIC and KCSD algorithms for a process following dynamics given by the equation (10). The errors for a process with dynamics given by equations (11) and (12) are shown in the right panel. The X axis is indexed by the time series length, i.e., sample size. The Type I error was around 5%.

bution P. Table 1 shows the *rejection rates* under the significance level $\alpha = 0.05$. It is clear that in the case where at least one of the samples is a Gibbs chain, the permutation-based test has a Type I error much larger than α. The wild bootstrap using V_{b1} (without artificial degeneration) yields the correct Type I error control in these cases. Consistent with findings in [19, Section 5], V_{b1} mimics the null distribution better than V_{b2}. The bootstrapped statistic $\widehat{\mathrm{MMD}}_{k,b}$ in (4) which also relies on the artificially degenerated bootstrap processes, behaves similarly to V_{b2}. In the alternative scenario where $\{Y_i\}$ was taken from a distribution with the same covariance structure but with the mean set to $\mu = [\ 2.5 \quad 0\ ]$, the Type II error for all tests was zero.

Pitch-evoking sounds Our second experiment is a two sample test on sounds studied in the field of pitch perception [18]. We synthesise the sounds with the fundamental frequency parameter of treble C, subsampled at 10.46kHz. Each i-th period of length Ω contains $d = 20$ audio samples at times $0 = t_1 < \ldots < t_d < \Omega$ – we treat this whole vector as a single observation X_i or Y_i, i.e., we are comparing distributions on $\mathbb{R}^{20}$. Sounds are generated based on the AR process $a_i = \lambda a_{i-1} + \sqrt{1-\lambda^2}\epsilon_i$, where $a_0, \epsilon_i \sim \mathcal{N}(0, I_d)$, with $X_{i,r} = \sum_j \sum_{s=1}^{d} a_{j,s} \exp\left(-\frac{(t_r - t_s - (j-i)\Omega)^2}{2\sigma^2}\right)$. Thus, a given pattern – a smoothed version of a_0 – slowly varies, and hence the sound deviates from periodicity, but still evokes a pitch. We take X with $\sigma = 0.1\Omega$ and $\lambda = 0.8$, and Y is either an independent copy of X (null scenario), or has $\sigma = 0.05\Omega$ (alternative scenario) (Variation in the smoothness parameter changes the width of the spectral envelope, i.e., the brightness of the sound). n_x is taken to be different from n_y. Results in Table 1 demonstrate that the approach using the wild bootstrapped statistic in (4) allows control of the Type I error and reduction of the Type II error with increasing sample size, while the permutation test virtually always rejects the null hypothesis. As in [19] and the MCMC example, the artificial degeneration of the wild bootstrap process causes the Type I error to remain above the design parameter of 0.05, although it can be observed to drop with increasing sample size.

Instantaneous independence To examine instantaneous independence test performance, we compare it with the Shift-HSIC procedure [8] on the 'Extinct Gaussian' autoregressive process proposed in the [8, Section 4.1]. Using exactly the same setting we compute type I error as a function of the temporal dependence and type II error as a function of extinction rate. Figure 1 shows that all three tests (Shift-HSIC and tests based on V_{b1} and V_{b2}) perform similarly.

Lag-HSIC The KCSD [4] is, to our knowledge, the only test procedure to reject the null hypothesis if there exist t,t' such that Z_t and $Z_{t'}$ are dependent. In the experiments, we compare lag-HSIC with KCSD on two kinds of processes: one inspired by econometrics and one from [4].

In lag-HSIC, the number of lags under examination was equal to $\max\{10, \log n\}$, where n is the sample size. We used Gaussian kernels with widths estimated by the median heuristic. The cumulative distribution of the V-statistics was approximated by samples from nV_{b2}. To model the tail of this distribution, we have fitted the generalized Pareto distribution to the bootstrapped samples ([20] shows that for a large class of underlying distribution functions such an approximation is valid).

The first process is a pair of two time series which share a common variance,

$$X_t = \epsilon_{1,t}\sigma_t^2, \quad Y_t = \epsilon_{2,t}\sigma_t^2, \sigma_t^2 = 1 + 0.45(X_{t-1}^2 + Y_{t-1}^2), \quad \epsilon_{i,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \quad i \in \{1,2\}. \tag{10}$$

The above set of equations is an instance of the VEC dynamics [2] used in econometrics to model market volatility. The left panel of the Figure 2 presents the Type II error rate: for KCSD it remains at 90% while for lag-HSIC it gradually drops to zero. The Type I error, which we calculated by sampling two independent copies $(X_t^{(1)}, Y_t^{(1)})$ and $(X_t^{(2)}, Y_t^{(2)})$ of the process and performing the tests on the pair $(X_t^{(1)}, Y_t^{(2)})$, was around 5% for both of the tests.

Our next experiment is a process sampled according to the dynamics proposed by [4],

$$X_t = \cos(\phi_{t,1}), \qquad \phi_{t,1} = \phi_{t-1,1} + 0.1\epsilon_{1,t} + 2\pi f_1 T_s, \quad \epsilon_{1,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{11}$$

$$Y_t = [2 + C\sin(\phi_{t,1})]\cos(\phi_{t,2}), \quad \phi_{t,2} = \phi_{t-1,2} + 0.1\epsilon_{2,t} + 2\pi f_2 T_s, \quad \epsilon_{2,t} \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{12}$$

with parameters $C = .4$, $f_1 = 4Hz$,$f_2 = 20Hz$, and frequency $\frac{1}{T_s} = 100Hz$. We compared performance of the KCSD algorithm, with parameters set to vales recommended in [4], and the lag-HSIC algorithm. The Type II error of lag-HSIC, presented in the right panel of the Figure 2,

is substantially lower than that of KCSD. The Type I error ($C = 0$) is equal or lower than 5% for both procedures. Most oddly, KCSD error seems to converge to zero in steps. This may be due to the method relying on a spectral decomposition of the signals across a fixed set of bands. As the number of samples increases, the quality of the spectrogram will improve, and dependence will become apparent in bands where it was undetectable at shorter signal lengths.

References

[1] M.A. Arcones. The law of large numbers for U-statistics under absolute regularity. *Electron. Comm. Probab*, 3:13–19, 1998.

[2] L. Bauwens, S. Laurent, and J.V.K. Rombouts. Multivariate GARCH models: a survey. *J. Appl. Econ.*, 21(1):79–109, January 2006.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

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

[5] I.S. Borisov and N.V. Volodko. Orthogonal series and limit theorems for canonical U- and V-statistics of stationary connected observations. *Siberian Adv. Math.*, 18(4):242–257, 2008.

[6] S. Borovkova, R. Burton, and H. Dehling. Limit theorems for functionals of mixing processes with applications to U-statistics and dimension estimation. *Trans. Amer. Math. Soc.*, 353(11):4261–4318, 2001.

[7] R. Bradley et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(107-44):37, 2005.

[8] K. Chwialkowski and A. Gretton. A kernel independence test for random processes. In *ICML*, 2014.

[9] J. Dedecker, P. Doukhan, G. Lang, S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190. Springer, 2007.

[10] Jérôme Dedecker and Clémentine Prieur. New dependence coefficients. examples and applications to statistics. *Probability Theory and Related Fields*, 132(2):203–236, 2005.

[11] P. Doukhan. *Mixing*. Springer, 1994.

[12] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

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

[14] 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.

[15] A. Gretton, K. Fukumizu, C Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *NIPS*, volume 20, pages 585–592, 2007.

[16] Arthur Gretton. A simpler condition for consistency of a kernel independence test. arXiv:1501.06103, 2015.

[17] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *NIPS*. 2008.

[18] P. Hehrmann. *Pitch Perception as Probabilistic Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2011.

[19] A. Leucht and M.H. Neumann. Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.

[20] J. Pickands III. Statistical inference using extreme order statistics. *Ann. Statist.*, pages 119–131, 1975.

[21] D. Sejdinovic, A. Gretton, and W. Bergsma. A kernel test for three-variable interactions. In *NIPS*, pages 1124–1132, 2013.

[22] D. Sejdinovic, H. Strathmann, M. Lomeli Garcia, C. Andrieu, and A. Gretton. Kernel Adaptive Metropolis-Hastings. In *ICML*, 2014.

[23] R. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

[24] X. Shao. The dependent wild bootstrap. *J. Amer. Statist. Assoc.*, 105(489):218–235, 2010.

[25] A. J Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, volume LNAI4754, pages 13–31, Berlin/Heidelberg, 2007. Springer-Verlag.

[26] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

[27] M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011.

[28] K. Zhang, J. Peters, D. Janzing, B., and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, pages 804–813, 2011.

[29] X. Zhang, L. Song, A. Gretton, and A. Smola. Kernel measures of independence for non-iid data. In *NIPS*, volume 22, 2008.

A An Introduction to the Wild Bootstrap

Bootstrap methods aim to evaluate the accuracy of the sample estimates - they are particularly useful when dealing with complicated distributions, or when the assumptions of a parametric procedure are in doubt. Bootstrap methods randomize the dataset used for the sample estimate calculation, so that a new dataset with a similar statistical properties is obtained, e.g. one popular method is resampling. In the wild bootstrap method the observations in the dataset are multiplied by appropriate random numbers. To present the idea behind the wild bootstrap we will discuss a toy example similar to that in [24], and then relate it to the wild bootstrap method used in this article.

Consider a stationary autoregressive-moving-average random process $\{X_i\}_{i\in\mathbf{Z}}$ with zero mean. The normalized sample mean of the process X_t has normal distribution

$$\frac{\sum_{i=1}^{N} X_i}{\sqrt{n}} \xrightarrow{d} N(0,\sigma_\infty^2), \tag{13}$$

where $\sigma_\infty^2 = \sum_{j=-\infty}^{j=\infty} cov(X_0, X_j)$. The variance σ_∞^2 is not easy to estimate (the naive approach of approximating different covariances separately and summing them has several drawbacks, e.g. how many empirical covariances should be calculated?). Using the wild bootstrap method we will construct processes that mimic behaviour of the X_t process and then use them to approximate the distribution of the normalized sample mean, $\frac{\sum_{i=1}^{N} X_i}{\sqrt{n}}$. The bootstrap process used to to randomize the sample meets the following criteria:

- $\{W_{t,n}\}_{1\le t\le n}$ is a row-wise strictly stationary triangular array independent of all X_t, such that $EW_{t,n} = 0$ and $\sup_n E|W_{t,n}^{2+\sigma}| < \infty$ for some $\sigma > 0$.
- The autocovariance of the process is given by $EW_{s,n}W_{t,n} = \rho(|s-t|/l_n)$ for some function ρ, such that $\lim_{u\to 0}\rho(u) = 1$.
- The sequence $\{l_n\}$ is taken such that $\lim_{n\to\infty} l_n = \infty$.

A process that fulfils those criteria, given also in the main article, is

$$W_{t,n} = e^{-1/l_n} W_{t-1,n} + \sqrt{1 - e^{-2/l_n}}\epsilon_t \tag{14}$$

We need to show that the distribution of the normalized sample mean of the process $Y_t^n = W_t^n X_t$, where $|t| \le n$, mimics the distribution $N(0,\sigma_\infty^2)$. It suffices to calculate the expected value and correlations:

$$EY_t^n = EW_t^n X_t = 0, \tag{15}$$

$$cov(Y_0^n, Y_t^n) = cov(X_0, X_t) cov(Y_0^n, Y_t^n) = cov(X_0, X_t)\rho(|t|/l_n) \to cov(X_0, X_t) \tag{16}$$

The asymptotic auto-covariance structure of the process Y_t is the same as the auto-covariance structure of the process X_t. Therefore

$$\frac{\sum_{i=1}^{N} Y_i}{\sqrt{n}} \xrightarrow{d} N(0,\sigma_\infty). \tag{17}$$

This mechanism is used in [19]. Recall that, under some assumptions, a normalized V-statistic can be written as

$$\sum_{k=0}^{\infty} \lambda_k \left(\frac{\sum_{i=1}^{n}\phi_k(X_i)}{\sqrt{n}}\right)^2 \overset{P}{=} \frac{1}{n}\sum_{1\le i,j\le n} h(X_i, X_j)$$

where λ_k are eigenvalues and ϕ_k are eigenfunction of the kernel h, respectively. Since $E\phi_k(X_i) = 0$ (degeneracy condition) one may replace

$$\frac{\sum_{i=1}^{n}\phi_k(X_i)}{\sqrt{n}}$$

with a bootstrapped version

$$\frac{\sum_{i=1}^{n} W_t^n \phi_k(X_i)}{\sqrt{n}},$$

and conclude, as in the toy example, that the limiting distribution of the single component of the sum $\sum_k \lambda_k ...$ remains the same. One of the main contributions of [19] is in showing that the distribution of the whole sum $\sum_k \lambda_k \left(\frac{\sum_{i=1}^n W_t^n \phi_k(X_i)}{\sqrt{n}} \right)^2$ with the components bootstrapped converges in a particular sense (in probability in Prokhorov metric) to the distribution of the normalized V-statistic, $\frac{1}{n} \sum_{1 \le i,j \le n} h(X_i, X_j)$.

B Relation between β,ϕ and τ mixing

Strong mixing coefficients. A process is called absolutely regular (β-mixing) if $\beta(m) \to 0$, where

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

The second supremum in the $\beta(m)$ definition is taken over all pairs of finite partitions $\{A_1, \cdots, A_I\}$ and $\{B_1, \cdots, B_J\}$ of the sample space such that $A_i \in \mathcal{A}_1^n$ and $B_j \in \mathcal{A}_{n+m}^{\infty}$, and $\mathcal{A}_b^c$ is a sigma field spanned by a subsequence, $\mathcal{A}_b^c = \sigma(Z_b, Z_{b+1}, ..., Z_c)$. A process is called uniform mixing (ϕ-mixing) if $\phi(m) \to 0$, where

$$\phi(m) = \sup_n \sup_{A \in \mathcal{A}_1^n} \sup_{B \in \mathcal{A}_{n+m}^{\infty}} |P(B|A) - P(B)|.$$

The process is called strongly mixing (α-mixing) if $\alpha(m) \to 0$, where

$$\alpha(m) = \sup_n \sup_{A \in \mathcal{A}_1^n} \sup_{B \in \mathcal{A}_{n+m}^{\infty}} |P(B \cap A) - P(B)P(A)|.$$

By [7] we have $\alpha(m) \le \beta(m) \le \phi(m)$.

Weak mixing The expected value and variance ofcoefficients. The process is called $\tilde{\alpha}$-mixing if $\tilde{\alpha}(m) \to 0$, where

$$\tilde{\alpha}(m) = \sup_{l \in \mathbb{N}} \frac{1}{l} \sup_{m \le i_1 \le ... \le i_l} \tilde{\alpha}(\mathcal{F}_0, (Z_{i_1}, ..., Z_{i_l})) \overset{r \to \infty}{\longrightarrow} 0, \text{ where}$$
$$\tilde{\alpha}(\mathcal{M}, X) = \sup_{g \in \Lambda} \parallel E(g(X)|\mathcal{M}) - Eg(X) \parallel_1$$

and Λ is the set of all one-Lipschitz continuous real-valued functions on the domain of X. The other weak mixing coefficient, already introduced, is τ-mixing. [9, Remark 2.4] show that $\tilde{\alpha}(m) \le 2\alpha(m)$. [10, Proposition 2] relates τ-mixing and $\tilde{\alpha}$ mixing, as follows: if Q_x is the generalized inverse of the tail function

$$Q_x(u) = \inf_{t \in R} \{P(|X| > t) \le u\},$$

then

$$\tau(\mathcal{M}, X) \le 2 \int_0^{\tilde{\alpha}(\mathcal{M}, X)} Q_x(u) du.$$

While this definition can be hard to interpret, it can be simplified in the case $E|X|^p = M$ for some $p > 1$, since via Markov's inequality $P(|X| > t) \le \frac{M}{t^p}$, and thus $\frac{M}{t^p} \le u$ implies $P(|X| > t) \le u$. Therefore $Q'(u) = \frac{M}{\sqrt[p]{u}} \ge Q_x(u)$. As a result, under the assumption that the real valued random variable is p-integrable for some $p > 1$, we have the following inequality

$$\frac{\sqrt[p]{\tilde{\alpha}(\mathcal{M}, X)}}{M} \ge C\tau(\mathcal{M}, X)$$

C Proofs

In this section we prove the main theorems. As for the notation, n denotes number of observations, $N = \{1, \cdots, n\}$, if h is function then $h \times h$ denotes a product of h with itself, $\lim_{n \to \infty} X_n \overset{L_2}{=} X$ denotes convergence in mean square

C.1 Proof of the Theorem 1

Hoeffding decomposition reduces any V-statistic to a sum of canonical V-statistics with canonical cores h_c, which are easier to study in context of non-iid data. As an illustration, consider a canonical core h of m arguments and fix some indexes $i_1 \leq \cdots \leq i_{m-1} \ll i_m$, for a sake of example we may assume that indexes represent time. If observations $Z_{i_1}, \cdots, Z_{i_{m-1}}$ are independent of the observation Z_{i_m}, then the expected value of $h(Z_{i_1}, \cdots, Z_{i_m})$, by degeneracy, is equal to zero. If it is reasonable to assume that Z_{i_m} is almost independent of $Z_{i_1}, \cdots, Z_{i_{m-1}}$, maybe because it is so distant in time, then it is also reasonable to expect that for a canonical core h (which is not too complicated)

$$Eh(Z_{i_1}, \cdot, Z_{i_m}) \approx 0.$$

which follows from the following approximate calculation

$$\int h(z_{i_1}, \cdot, z_{i_m}) dP_{Z_{i_1}, \cdots, Z_{i_m}} \approx \int h(z_{i_1}, \cdots, z_{i_m}) dP_{Z_{i_1}, \cdots, Z_{i_{m-1}}} dP_{Z_{i_m}} = 0$$

We formalize this intuition.

Definition 1. *Associate with any set of indexes $i_1, \cdots, i_m$ its nearest neighbor within the set. Suppose i_r is is an index with the most distant nearest neighbor. We will call i_r the most isolated index, and we will refer to its distance to the nearest neighbor as an isolation distance.*

Consider a following example, for the set $\{1, 5, 7\}$, 1 is the most isolated index and the isolation distance is 4.

Definition 2. *Given a sequence of random variables Z_t and a function h, if for all sets of indexes $i_1, \cdots, i_m$, with the isolation distance equal to r*

$$|Eh(Z_{i_1}, \cdots, Z_{i_m})| \leq \Delta(h, r)$$

for some some function Δ, then we say that the pair (h, Z_t) is of type Δ.

The next theorem shows a growth rate of a canonical V-statistic when a pair h, Z_t is of type Δ.

Theorem 4. *Let (Z_t, h), where h is a function of $m > 1$ arguments, be a of type Δ, with $\Delta(h, r) = o(r^{-k})$ for some k, then*

$$\sum_{i \in N^m} |Eh(Z_i)| = O\left(n^{\lfloor \frac{m}{2} \rfloor}\right) + o\left(n^{2\lfloor \frac{m}{2} \rfloor + 2 - k}\right).$$

Proof. The proof uses a technique similar to [1, Lemma 3]. We will focus on ordered m-tuples $1 \leq i_1 \leq \ldots \leq i_m \leq n$, and by considering all possible permutations of their indices, we obtain an upper bound

$$\sum_{i \in N^m} |Eh\left(Z_{i_1}, \ldots, Z_{i_m}\right)| < \sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} \sum_{\pi \in S_m} \left|Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right|,$$

where (strict) inequality stems from the fact that the m-tuples with some coinciding entries appear multiple times on the right.

Since (h, Z_t) is a of type Δ

$$\forall i \in N^m \sum_{\pi \in S_m} \left|Eh\left(Z_{i_{\pi(1)}}, \ldots, Z_{i_{\pi(m)}}\right)\right| = O(\Delta(h, w(i))),$$

where $w(i)$ is an isolating distance of the index set $i = i_1, \cdots i_m$. We need to estimate order of the sum

$$\sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} O(\Delta(h, w(i))).$$

Let us upper bound the number of ordered m-tuples i with $w(i) = w$. Denote $s = \left\lfloor \frac{m}{2} \right\rfloor + 1$. i_1 can take n different values, but since $i_2 \leq i_1 + w$, i_2 can take at most $w + 1$ different values. For $2 \leq l \leq s - 1$, since $\min\{i_{2l} - i_{2l-1}, i_{2l-1} - i_{2l-2}\} \leq w$, we can either let i_{2l-1} take up to n different values and let i_{2l} take up to $w+1$ different values (if $i_{2l} - i_{2l-1} \leq i_{2l-1} - i_{2l-2}$) or let i_{2l-1}

take up to $w+1$ different values and let i_{2l} take up to n different values (if $i_{2l}-i_{2l-1} > i_{2l-1}-i_{2l-2}$), upper bounding the total number of choices for $[i_{2l-1}, i_{2l}]$ by $2n(w+1)$. Finally, the last term i_m can always have at most $w+1$ different values. This brings the total number of m-tuples with $w(i) = w$ to at most $2^{s-2}n^{s-1}(w+1)^s$. Thus, the number of m-tuples with $w(i) = 0$ is $O(n^{s-1})$ and since $Eh\left(Z_{i_1}, \ldots, Z_{i_m}\right) < \infty$, we have

$$
\begin{aligned}
&\sum_{1 \leq i_1 \leq \ldots \leq i_m \leq n} O(\Delta(h, w(i))) \\
&\leq O(n^{s-1}) + \sum_{w=1}^{n-1} \sum_{\substack{1 \leq i_1 \leq \ldots \leq i_m \leq n: \\ w(i)=w}} O(\Delta(h, w(i))) \\
&\leq O(n^{s-1}) + n^{s-1} \sum_{w=1}^{n-1} (w+1)^s O(\Delta(h, w)) \\
&\leq O(n^{s-1}) + n^{s-1} \sum_{w=1}^{n-1} o(w^{s-k}) \\
&\leq O(n^{s-1}) + n^{s-1} \max(o(n^{s-k+1}), O(1)) \\
&\leq O(n^{s-1}) + o(n^{2s-k}) + O(n^{s-1})) \\
&= O(n^{s-1}) + o(n^{2s-k}),
\end{aligned}
$$

which proves the claim. We have used $\Delta(h, w) = o(w^{-k})$. □

The previous theorem states sufficient conditions for a V-statistic or a bootstrapped V-statistic to converge to zero.

Lemma 4. *Let h be a function of $m > 1$ arguments and let $(\{Z_t\}_{t \in N}, h \times h)$ be a of type Δ, with $\Delta(h \times h, r) = o(r^{-4})$. If $\{G_i\}_{i \in N}$ is a random process, independent of Z_t, such that $\sup_i EG_i^4 < \infty$, with notation $T_n = \frac{1}{n^{m-1}} \sum_{i \in N^m} G_{i_1} G_{i_2} h(Z_i)$,*

$$
\begin{cases} \lim_{n \to \infty} o(1) T_n \overset{L_2}{=} 0 & m = 2, \\ \lim_{n \to \infty} T_n \overset{L_2}{=} 0 & m > 2 \end{cases}
$$

since,

$$
\begin{cases} ET_n^2 = O(1) & m = 2, \\ ET_n^2 = o(1) & m > 2. \end{cases}
$$

Proof. First we verify that for $i, j \in N^m$

$$
a_{i,j} = EG_{i_1} G_{i_2} G_{j_1} G_{j_2}
$$

is uniformly bounded. We get the bound by applying Cauchy-Schwarz iteratively and using assumption $\sup_i EG_i^4 < \infty$.

We check that the second non-central moment converges to zero,

$$
\begin{aligned}
&E\left(T_n\right)^2 \\
&= \frac{1}{n^{2m-2}} \sum_{i,j \in N^m} EG_{i_1} G_{i_2} G_{j_1} G_{j_2} Eh(Z_i) h(Z_j) \\
&\leq \frac{1}{n^{2m-2}} \sum_{i,j \in N^m} |a_{i,j} Eh(Z_i) h(Z_j)| \\
&\leq \left(\sup_n \sup_{i,j \in N^m} |a_{i,j}| \right) \frac{1}{n^{2m-2}} \sum_{i,j \in N^m} |Eh(Z_i) h(Z_j)|.
\end{aligned}
$$

Supremum over n is needed since $EG_{i_1}G_{i_2}G_{j_1}G_{j_2}$ might change with n. Lemma 4, by the assumption that $(h(\cdots)\times h(\cdots), Z_t)$ is of type Δ, the growth of the inner sum $\sum_{i,j\in N^m}|Eh(Z_i)h(Z_j)|$ is at most of order

$$O(n^m)+o(n^{2m+2-k}).$$

Since $\Delta(h\times h, r)=o(r^{-4})$, the growth rate is

$$E\left(T_n\right)^2=\frac{O\left(n^m\right)+o(n^{2m-2})}{n^{2m-2}}=\begin{cases}O(1) & m=2\\ o(1) & m>2\end{cases}$$

For $m=2$ we have assumed existence of an extra term $o(1)$, which concludes the proof. □

We next prove that the asymptotic distribution of a V-statistic depends on number of terms in the Hoeffding decomposition that are equal to zero.

Lemma 5. *Let h be a core with m arguments. If $h_0=h_1=0$, and for all $c>2$ component $(h_c\times h_c, Z_t)$ is of type Δ, with $\Delta(h_c\times h_c, r)=o(r^{-4})$ then*

$$\lim_{n\to\infty}\left(nV_n(h)-\binom{m}{2}nV_n(h_2)\right)\overset{L_2}{=}0$$

Proof. Using Hoeffding decomposition we write the core h as a sum of the components h_c ,

$$\begin{aligned}nV_n(h)=&nV_n(h_m)+\binom{m}{1}nV_n(h_{m-1})+...\\&+\binom{m}{m-2}nV_n(h_2)+\binom{m}{m-1}nV_n(h_1)+h_0.\end{aligned}$$

$h_0=0$ and $h_1=0$. By Lemma 4, for $c\geq 3$, $nV_n(h_c)$ converges to zero in mean squared. To see that it suffices to put $Q=1$ and verify that $(h_c\times h_c, Z_t)$ is of Δ type, which is explicitly assumed. □

Before we study the asymptotic distribution of a bootstrapped statistic B_n we need to sate three simple lemmas that will be frequently used.

Lemma 6. *If W_i is a bootstrap process then*

$$\lim_{n\to\infty}\frac{l_n}{n}\sum_{i=1}^{n}W_i\overset{L_2}{=}0.$$

Proof. By the definition of W_i, $E(\sum_{i=1}^n W_i)^2\leq n2\sum_{r=1}^n Cov(W_0,W_r)=nO(l_n)$, where $\sum_{r=1}^n Cov(W_0,W_r)=O(l_n)$ follows from bootstrap assumption. Also, by the Bootstrap assumptions we have $\lim_{n\to\infty}\frac{l_n^3}{n^2}=0$. Therefore $\frac{1}{n}\sum_{i=1}^n W_i$ converges to zero in mean squared. □

Lemma 7. *If $\{W_i\}$ is a bootstrap process then*

$$\sum_{i=1}^{n}\tilde{W}_i=\sum_{i=1}^{n}\left(W_i-\frac{1}{n}\sum_{j=1}^{n}W_j\right)=0.$$

Lemma 8. *Let f be a function and let $j=\{j_1,\ldots,j_q\}$ be a subset of $\{1,\ldots,m\}$. Then*

$$\sum_{i\in N^m}f(Z_{i_{j_1}},...,Z_{i_{j_q}})=n^{m-q}\sum_{i\in N^q}f(Z_{i_1},...,Z_{i_q})$$

Proof. Each element $f(Z_{i_{j_1}},...,Z_{i_{j_q}})$ is repeated exactly n^{m-q} times. □

We now prove an analogue of the Lemma 5 for bootstrapped statistics B.

Lemma 9. *Let h be a core of a m arguments and let Q_i denote W_i or $\tilde{W}_i$. If*

$$\frac{1}{n^2}\sum_{i\in N^2} Q_{i_1}Q_{i_2}h_0 = 0,$$
$$\frac{1}{n^m}\sum_{i\in N^m}\sum_{1\le j\le m} Q_{i_1}Q_{i_2}h_1(Z_{i_j}) = 0.$$

and (h_c, Z_t) for $c > 2$ are of type Δ, with $\Delta(h_c \times h_c, r) = o(r^{-4})$ then

$$\lim_{n\to\infty}\left(nB(h) - \binom{m}{2} nB(h_2)\right) \overset{L_2}{=} 0$$

Proof. Where it is necessary, we check claims for both W_i and $\tilde{W}_i$ separately. We will frequently use the fact that $\frac{l_n}{n}\sum_{i=1}^n Q_i, \frac{1}{n}\sum_{i=1}^n Q_i$ converge to zero in mean square.

Using Hoeffding decomposition we write core h as a sum of components h_c (the ones with h_0, h_1 are equal to zero and therefore omitted)

$$nB_1(h) = \frac{1}{n^{m-1}}\sum_{i\in N^m}\Big[Q_{i_1}Q_{i_2}h_m(Z_{i_1},...,Z_{i_m}) + \sum_{1\le j_1<...<j_{m-1}\le m} Q_{i_1}Q_{i_2}h_{m-1}(Z_{i_{j_1}},...,Z_{i_{j_{m-1}}}) + ... + \sum_{1\le j_1<j_2\le m} Q_{i_1}Q_{i_2}h_2(Z_{i_{j_1}},Z_{i_{j_2}})\Big].$$

Consider the sum associated with h_c

$$\frac{1}{n^{m-1}}\sum_{i\in N^m}\sum_{1\le j_1<...<j_c\le m} Q_{i_1}Q_{i_2}h_c(Z_{i_{j_1}},...,Z_{i_{j_c}}). \tag{18}$$

We will show that for almost all fixed $j_1 < \cdots < j_c$ the sum 18 converges to zero.

Suppose $j_1 > 2$. The sum 18 can be written

$$\frac{1}{n^{m-1}}\sum_{i\in N^m} Q_{i_1}Q_{i_2}h_c(Z_{i_{j_1}},...,Z_{i_{j_c}}) \overset{L.8}{=\!=} \frac{1}{n^{c+1}}\sum_{i\in N^{c+2}} Q_{i_1}Q_{i_2}h_c(Z_{i_3},...,Z_{i_{c+2}})$$
$$= \left(\frac{1}{n^{c-1}}\sum_{i\in N^c} h_c(Z_{i_1},...,Z_{i_c})\right)\left(\frac{1}{n}\sum_{i=1}^n Q_i\right)^2 = \frac{n}{l_n}V_n(h_c)\left(\frac{l_n}{n}\sum_{i=1}^n Q_i\right)^2.$$

By Lemma 4, for $c \ge 3$, $\frac{n}{l_n}V_n(h_c)$ converges to zero in mean squared. Indeed, it is sufficient to put $G_i = 1$ and $T_n = nV_n(h_c)$ and notice that $\frac{n}{l_n}V_n(h_c) = \frac{1}{l_n} = o(1)T_n$, since $l_n \to \infty$. Consequently, since $(\frac{1}{n}\sum_{i=1}^n Q_i)^2$ converges to zero in mean square 6, the product, converges to zero in mean square i.e.

$$V_n(h_c)\left(\frac{1}{n}\sum_{i=1}^n Q_i\right)^2 \overset{L_2}{\to} 0$$

Suppose $j_1 = 2$. The sum 18 can be written

$$\frac{1}{n^{m-1}}\sum_{i\in N^m} Q_{i_1}Q_{i_2}h_c(Z_{i_2},...,Z_{i_{j_c}}) \overset{L.8}{=\!=} \frac{1}{n^c}\sum_{i\in N^{c+1}} Q_{i_1}Q_{i_2}h_c(Z_{i_2},\cdots,Z_{i_{j_c}}) =$$
$$\left(\frac{1}{l_n n^{c-1}}\sum_{i\in N^c} Q_{i_1}h_c(Z_{i_1},\cdots,Z_{i_c})\right)\left(\frac{l_n}{n}\sum_{i=1}^n Q_i\right). \tag{19}$$

The latter expression $\frac{l_n}{n}\sum_{i=1}^n Q_i$ converges to zero in mean square. The former expression can be further decomposed

$$\frac{1}{l_n}n^{-c+1}\sum_{i\in N^c} Q_{i_1}h_c(Z_{i_1},\cdots,Z_{i_c}) = \frac{1}{4}\frac{1}{l_n}(T_+ - T_-) \text{ where,}$$
$$\frac{1}{l_n}T_- = \frac{1}{l_n}n^{-c+1}\sum_{i\in N^2}(Q_{i_1}-1)h_c(Z_{i_1},\cdots,Z_{i_c})(Q_{i_2}-1),$$
$$\frac{1}{l_n}T_+ = \frac{1}{l_n}n^{-c+1}\sum_{i\in N^2}(Q_{i_1}+1)h_c(Z_{i_1},\cdots,Z_{i_c})(Q_{i_2}+1),$$

We use Lemma 4 for $\frac{1}{l_n}T_+$ and $\frac{1}{l_n}T_-$, to show that they converge to zero. We need to check that

$$\sup_i E(Q_i + / - 1)^4 < \infty$$

If $Q_i = W_i$ this follows from the Bootstrap assumptions $\sup_n \sup_{i\leq n} EW_{i,n}^4 < \infty$. If $Q_i = \tilde{W}_i$ we check that

$$E(\frac{1}{n}\sum_{i=1}^n W_i)^4 \leq \sup_n \sup_{i\leq n} EW_{i,n}^4,$$

and so $\leq \sup_i E(\tilde{W}_i) < \infty$. Now we conclude that both $\frac{1}{l_n}T_+$ and $\frac{1}{l_n}T_-$ converge to zero. Therefore their sum (even though they are not independent) converges to zero.

Suppose $j_1 = 1$ and $j_2 > 2$. This case is identical to the previous case, up to swapping i_1, i_2 in the equation 19.

Finally, suppose $j_1 = 1$ and $j_2 = 2$ and $c > 2$. The sum 18 can be written

$$\frac{1}{n^{m-1}}\sum_{i\in N^m} Q_{i_1}Q_{i_2}h_c(Z_{i_1},Z_{i_2},...,Z_{i_{j_c}}) \overset{L.8}{=\!=} \frac{1}{n^c}\sum_{i\in N^{c+1}} Q_{i_1}Q_{i_2}h_c(Z_{i_1},Z_{i_2},...,Z_{i_{j_c}})$$

We again use Lemma 4 to see that this sum converges to zero in mean squared (we checked the assumptions above). We have proved that

$$\lim_{n\to\infty}\left(nB(h) - \binom{m}{2}nB(h_2)\right) \overset{L_2}{=} 0$$

□

So far we avoided expressing results in terms of τ-mixing and degeneracy of a core, now we relate Δ formalism to those concepts. We start with a technical lemma.

Lemma 10. *If h is a Lipschitz continuous core then its components are also Lipschitz continuous.*

Proof. The auxiliary function used in the Hoeffding decomposition

$$g_c(z_1,...z_c) = Eh(z_1,...,z_c,Z_{c+1}^*,...,Z_m^*).$$

is Lipschitz, since h is Lipschitz continuous.

$$
\begin{aligned}
&|g_c(z_1,...z_c) - g_c(z'_1,...z'_c)| \\
&\leq \left| \int [h(z_1,...,z_c,z_{c+1},...,z_m) - h(z'_1,...,z'_c,z_{c+1},...,z_m)] dP(z_{c+1}) \cdots dP(z_m) \right| \\
&\leq \left| \int Lip(h) \left(\sum_{i=1}^{c} |z_i - z'_i| + \sum_{i=c+1}^{m} |z_i - z_i| \right) dP(z_{c+1}) \cdots dP(z_m) \right| \\
&\leq \left| \int Lip(h) \left(\sum_{i=1}^{c} |z_i - z'_i| \right) dP(z_{c+1}) \cdots dP(z_m) \right| \\
&= |Lip(h) \sum_{i=1}^{c} |z_i - z'_i| \int dP(z_{c+1}) \cdots dP(z_m)| \\
&= |Lip(h) \sum_{i=1}^{c} |z_i - z'_i||.
\end{aligned}
$$

h_0 is obviously Lipschitz continuous. If h_k for $k < c$ are Lipschitz continuous then, since g_c is Lipschitz continuous, h_c is also Lipschitz continuous as a sum of Lipschitz continuous functions. □

Lemma 11. *Let $\{Z_t\}$ be a τ-dependent stationary process and h be a Lipschitz core of m arguments, If for all $c > 0$ $(h_c \times h_c, Z_t)$ and (h, Z_t) are of type Δ with the rate $O(\tau(d))$ then*

$$\Delta(h,d) = \Delta(h_c \times h_c, d) = O(\tau(d))$$

Proof. Let $f = h_c \times h_c$ or $f = h$. f is canonical and Lipschitz continuous (if $f = h_c \times h_c$ it follows from Lemma 10). Suppose i_r is the isolating index. Further suppose there are k indexes $a_1, \cdots, a_k$ smaller than i_r and $m - k - 1$ indexes greater than i_r, namely $a_{k+2}, \cdots, a_m$. In this notation $a_{k+1} = i_r$.

Let us partition the vector $(Z_{i_1}, \ldots, Z_{i_m})$ into three parts:

$$A = (Z_{a_1}, \ldots, Z_{a_k}),\ B = Z_{a_{k+1}},\ C = \left(Z_{a_{k+2}}, \ldots, Z_{a_m}\right).$$

where a_{k+1} is the isolating index. If $k = 0$, A is empty and if $k = m - 1$, C is empty but this does not change our arguments below. Using Lemma [9, Lemma 5.3], we will construct B^* and C^{**} that are independent of A and independent of each other and

$$E \left\| (A,B,C) - (A,B^*,C^{**}) \right\|_1 = O(\tau(w)), \tag{20}$$

where w is an isolating distance [1]. Let $D = (B, C)$ The [9, Lemma 5.3] guarantees that there exist D^* independent of A, such that

$$
\begin{aligned}
\|Ed(D,D^*)|\sigma(A)\|_1 &= E|Ed(D,D^*)|\sigma(A)| \\
&= E(Ed(D,D^*)|\sigma(A)) = Ed(D,D^*) = O(\tau(w)),
\end{aligned}
$$

where d is the L_1 distance on Euclidean space (non-negativity justifies dropping absolute value). By definition of τ-mixing, $\tau(w) \geq \tau(\sigma(A), D)$. Since $D^* = (B^*, C^*)$ has the same distribution as D (in particular it has the same τ dependence structure) we use the lemma again to construct C^{**}, independent of A and B^*, such that

$$Ed(C, C^{**}) = O(\tau(w)).$$

By the triangle inequality we obtain equation 20.

$$
\begin{aligned}
&Ed\big((A,B,C) - (A,D^*) + (A,D^*) - (A,B^*,C^{**})\big) \leq \\
&Ed\big((A,B,C) - (A,D^*)\big) + Ed\big((A,D^*) - (A,B^*,C^{**})\big) = \\
&Ed(D,D^*) + Ed(C,C^{**}) = O(\tau(w)).
\end{aligned}
$$

[1] [9, Lemma 5.3] assumes that there exists a random variable δ independent of the vector (A, B, C). This assumption is important only if CDF of the vector is not continuous, we can assume that our space is endowed with such δ.

Since B^* is a singleton, independent of both A and C^{**}, by degeneracy of f

$$Ef(A, B^*, C^{**}) = 0. \tag{21}$$

Note that $f(A, B^*, C^{**})$ is just a shorthand, random variables A, B^*, C^{**} are inserted in the right order. Thus, we have that

$$\begin{aligned} |Ef\left(Z_{i_1}, \ldots, Z_{i_m}\right)| &\leq E\left|f\left(A, B, C\right) - f\left(A, B^*, C^{**}\right)\right| + \left|Ef(A, B^*, C^{**})\right| \\ &\leq \text{Lip}(f) E \left\|(A, B, C) - (A, B^*, C^{**})\right\|_1 + 0 \\ &= O(\tau(w)). \end{aligned}$$

□

Finally we can prove Theorem 1.

Proof. In the proof we are going to use [19][Theorems 2.1, 3.1], which characterise asymptotic properties of $nV_n(h_2)$ and $nB(h_2)$. Both theorems use similar set of assumptions which we verify upfront.
Assumption A2.

- (i) h_2 is one-degenerate and symmetric - this follows from the Hoeffding decomposition;
- (ii) h_2 is a kernel - is one of the assumptions of this theorem;
- (iii) $Eh_2(Z_1, Z_1) < \infty$ – follows from $\sup_{i \in N^6} |Eh(Z_i)| < \infty$;
- (iv) h_2 is Lipschitz continuous - follows from the Lemma 10.

Assumption B1, A1. Assumption $B1$, $\sum_{r=1}^{n} r^2 \sqrt{\tau(r)} < \infty$, is the same as ours, assumption $A1$, $\sum_{r=1}^{n} \sqrt{\tau(r)} < \infty$ is implied.
Assumption B2. This assumption about the bootstrap process W_t is the same as our Bootstrap assumptions.

Denote by V the weak limit of $nV_n(h_2)$, which exits by the [19][Theorem 2.1], and let $\mathcal{F} = \sigma(Z_1, \cdots, Z_n)$. By [19, Theorem 3.1], since the distribution of V is continuous, we have

$$\sup_{x \in R} |P(nB_n(h_2) < x|\mathcal{F}) - P(V < x)| \to 0$$

in probability. We show that $nB_n(h_2)$ converges to V weakly, by showing pointwise convergence of CDF

$$\begin{aligned} \lim_{n\to\infty} P(nB_n(h_2) < x) &= \lim_{n\to\infty} EP(nB_n(h_2) < x|\mathcal{F}) \\ &= E \lim_{n\to\infty} P(nB_n(h_2) < x|\mathcal{F}) = EP(V < x) = P(V < x) \end{aligned}$$

To change the order of limit and expectation we have dominated convergence Theorem, justified since $P(nB_n(h) < x|\mathcal{F})$ are bounded by 1. The difference $n(B_n(h) - V_n(h))$ is

$$n\left(B_n(h) - \binom{m}{2} B_n(h_2)\right) + \binom{m}{2}\left(nB_n(h_2) - V\right) + \left(\binom{m}{2} V - nV_n(h)\right)$$

By Lemma 9 and Lemma 5 respectively, both

$$n(B(h) - \binom{m}{2} B(h_2)), n(V_n(h) - n\binom{m}{2} V_n(h_2))$$

converge to zero in mean square. We check assumptions: since Z_t is tau mixing and h is Lipschitz continuous, by Lemma 11 all self products of components and Z_t, $(h_c \times h_c, Z_t)$ for $c > 0$, are Δ type of order $\tau(r)$, of order at least $o(r^{-4})$ (since $\sum_{r=1}^{n} r^2 \sqrt{\tau(r)} < \infty$). Since h is one degenerate, first and zero component h_0, h_1 are equal to zero (and so are $B(h_0), B(h_1)$).

This shows that $nB_n(h_2)$ converges weakly to V. □

C.2 Proof of Theorem 2

Proof. Using Hoeffding decomposition we write the core h as a sum of the components h_c ,

$$\begin{aligned} nV_n(h) =& nV_n(h_m) + \binom{m}{1} nV_n(h_{m-1}) + ... \\ &+ \binom{m}{m-2} nV_n(h_2) + \binom{m}{m-1} nV_n(h_1) + h_0. \end{aligned}$$

By the Lemma 4, for $c \geq 1$, $V_n(h_c)$ converges to zero in probability. The sum associated with h_1 is

$$V_n(h_1) = \frac{1}{n} \sum_{i=1}^{N} h_1(Z_i).$$

By Lemma 11 $(h_1 \times h_1, Z_t)$ is Δ type of order $o(r^{-4})$. Using Lemma 4 we get the growth rate of $E(V_n(h_1))^2 = O(\frac{1}{n})$, thus $V_n(h_1)$ converges in mean square to zero. □

C.3 Proof of Theorem 3

Proof. We show that the second non central moment of B_1 converges to 0. The second non central moment is

$$\begin{aligned} EB_1 &= E\frac{1}{n^{2m}} \sum_{i \in N^{2m}} W_{i_1} W_{i_2} W_{i_{m+1}} W_{i_{m+2}} Eh(Z_{i_1}, ..., Z_{i_m}) h(Z_{i_{m+1}}, ..., Z_{i_{2m}}) \\ &= \frac{1}{n^{2m}} \sum_{i \in N^{2m}} EW_{i_1} W_{i_2} W_{i_{m+1}} W_{i_{m+2}} Eh(\cdots) h(\cdots) \\ &\leq CE\frac{1}{n^4} \sum_{i \in N^4} |EW_{i_1} W_{i_2} W_{i_{m+1}} W_{i_{m+2}}| \\ &= CE \left(\frac{1}{n} \sum_{i=1}^{n} W_i \right)^4. \end{aligned}$$

The inequality in the third line follows from the fact that correlations of the bootstrap process W_i are positive (Bootstrap assumption) and

$$C = \sup_n \sup_{i \in N^m} Eh(Z_{i_1}, ..., Z_{i_m}) h(Z_{i_{m+1}}, ..., Z_{i_2 m}),$$

is finite. By Lemma 6

$$\frac{1}{n} \sum_{i=1}^{n} W_i \to 0,$$

and therefore $EC \left(\frac{1}{n} \sum_{i=1}^{n} W_i \right)^4 \to 0$.

We now prove that $o(n)B_2(h)$ converges to zero. Using Hoeffding decomposition we write core h as a sum of components h_c and h_0

$$nB_2(h) = \frac{1}{n^{m-1}} \sum_{i \in N^m} \Big[h_0 \tilde{W}_{i_1} \tilde{W}_{i_2} + \sum_{1 \leq j \leq m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_j}) \tag{22}$$

$$\sum_{1 \leq j_1 < j_2 \leq m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_2(Z_{i_{j_1}}, Z_{i_{j_2}}) + ... + \tilde{W}_{i_1} \tilde{W}_{i_2} h_m(Z_{i_1}, ..., Z_{i_m}) \Big]. \tag{23}$$

We examine terms of the above sum starting form the one with h_0 - it is equal to zero

$$\frac{1}{n^{m-1}} \sum_{i \in N^m} h_0 \tilde{W}_{i_1} \tilde{W}_{i_2} \overset{L.8}{=\!=} \frac{1}{n} h_0 \sum_{i \in N^2} \tilde{W}_{i_1} \tilde{W}_{i_2} = \frac{1}{n} h_0 \left(\sum_{i=1} \tilde{W}_i \right)^2 \overset{L.7}{=\!=} 0.$$

Term with h_1 is zero as well, to see that fix j and consider

$$T_j = \frac{1}{n^{m-1}} \sum_{i \in N^m} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_j}).$$

If $j = 1$ then

$$T_1 \overset{L.8}{=\!=} \frac{1}{n} \sum_{i \in N^2} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_1}) = \frac{1}{n} \left(\sum_{i=1}^{n} \tilde{W}_i h_1(Z_i) \right) \left(\sum_{i=1} \tilde{W}_i \right) \overset{L.7}{=\!=} 0.$$

If $j = 2$ the same reasoning holds and if $j > 2$

$$T_j \overset{L.8}{=\!=} \frac{1}{n^2} \sum_{i \in N^3} \tilde{W}_{i_1} \tilde{W}_{i_2} h_1(Z_{i_3}) = \frac{1}{n} \left(\sum_{i=1}^{n} h_1(Z_i) \right) \left(\sum_{i=1} \tilde{W}_i \right)^2 \overset{L.7}{=\!=} 0.$$

By Lemma 9, since $B(h_0) = B(h_1) = 0$, $(nB(h) - \binom{m}{2} nB(h_2)) \to 0$ in mean square and the only term that remains is

$$T_n = \frac{1}{n} \sum_{i,j \in N} \tilde{W}_i \tilde{W}_j h_2(Z_i, Z_j)$$

Now we can use the Lemma 4 to show that $o(1)T_n$ converges to zero. □

D Various comments

D.1 Time complexity

The original HSIC and MMD tests for i.i.d. data, the computational cost of the wild bootstrap approach scales quadratically in the number of samples, and linearly in the number of bootstrap iterations (in the i.i.d. case, these were permutations of the data). The main alternative approaches are the lagged bootstrap of [8], which has the same scaling with data and number of bootstraps, and the spectrogram approach of [4] (note, however, that both these alternative approaches apply only to the independence testing case). The cost of [4] is comparable to our approach, however the statistical power of [4] was much weaker on the data we examined.