A Kernel Test for Three-Variable Interactions with Random Processes (ICML 2016)

Abstract

Explain what this is all about, and the main contributions:

- Applied Wild Bootstrap to Lancaster test statistic
- Main theoretical challenge was to show that the conditions required to apply WB are satisfied by Lancaster
- This was done in a novel way - rather than using the Hoeffding decomposition, we come up with a new method which is simpler, (but requires an extra condition on the timeseries?)
- We also show that the power of the Lancaster test described in Arthur's original paper can be improved - we show that they used conservative p-values

1. Introduction

- Describe three variable interaction. It is particularly useful for cases in which any pairwise interaction is weak, but that the three variables interact strongly together.

- Test consists of two parts - calculating the test statistic, and bootstrapping the statistic to sample from the null in order to calculate the p-value threshold.

- When using time series, the difficult part is the bootstrapping because shuffling the indices breaks the temporal dependence structure.

- In [Leucht], they give a method for bootstrapping a certain class of statistics.

- The main contributions of this paper are the folllowing:
  - To show that the Lancaster test statistic is such a statistic
  - This is done using a new style of technique which in particular gives a significantly simpler proof that HSIC is also such a statistic (and thus simplifies the proofs used in [HSIC+time series])
  - To show that the multiple testing corrections used in [Lancaster] are too conservative, and therefore that we can improve test power by using a more relaxed correction.

This work combines the works of [HSIC + time series] and [Lancaster interaction] to give a non-parametric test for three variable interactions in which the samples are drawn from random processes.

2. Background

In this section we briefly introduce the theory and definitions required to understand the statement and proof of our main result.

2.1. Kernel Mean Embedding (and HSIC?)

Given an integrally strictly positive definite kernel k on a set $\mathcal{Z}$, the mapping induced by k from $\mathcal{M}(\mathcal{Z})$, the set of signed measures on $\mathcal{Z}$, to the RKHS $\mathcal{H}_k$ of k via $m \mapsto \int k(x,\cdot)dm(x)$ is injective. Given a finite sample $z_1, \dots, z_n$ drawn from a probability distribution $\mathbb{P}_z$, the mean embedding $\mu_{\mathbb{P}_z}$ can be estimated as $\hat{\mu}_{\mathbb{P}_z} = \frac{1}{n}\sum_{i=1}^{n} k(z_i,\cdot)$. This idea is exploited in the construction of certain statistical tests including two sample independence testing (HSIC) - in this case, we wish to understand whether $\mathbb{P}_{XY}$ factorises as $\mathbb{P}_X\mathbb{P}_Y$ based on finite samples (X_i, Y_i) drawn from $\mathbb{P}_{XY}$. We can consider distance between the empirical embeddings of the two measures via $\|\hat{\mu}_{\mathbb{P}_{XY}} - \hat{\mu}_{\mathbb{P}_X\mathbb{P}_Y}\|^2$. We can then bootstrap this statistic to generate samples of it under the null hypothesis that $\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ to calculate a threshold distance over which we would reject the null hypothesis and conclude that the distribution does not factorise

2.1.1. Notation?

Maybe put all notation here? Need to define gram matrices, empirically centred gram matrices and population centred gram matrices.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Throughout this paper we will stick to the convention that X, Y and Z are random variables taking value in $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, on which we define k, l and m respectively to be kernels. We will assume that our kernels are characteristic and bounded. We describe some notation relevant to the kernel k; similar notation holds for the other two kernels.

Associated with the kernel k is a Hilbert space $\mathcal{H}_k$ of functions on $\mathcal{X}$ and a feature map $\phi_X : \mathcal{X} \longrightarrow \mathcal{H}_k$ such that $k(x,x') = \langle \phi_X(x), \phi_X(x') \rangle$. Given observations $\{X_i\}_{i=1}^n$, we write K to be the *Gram matrix* with entries $K_{ij} = k(X_i, X_j)$. We write $\mu_X := \mathbb{E}_X k(X, \cdot)$ which we call the *mean embedding* of the random variable X. When k is bounded we can think of μ_X as the expectation of the $\mathcal{H}_k$-valued random variable $\phi_X(X)$. We write $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n k(X_i, \cdot)$ and remark that $\hat{k}(x,x') = \langle \phi_X(x) - \hat{\mu}_X, \phi_X(x') - \hat{\mu}_X \rangle$ is a kernel with feature map $\bar{\phi}_X(X) = \phi_X(X) - \mu_X$. We denote by $\hat{K}$ the Gram matrix with respect to $\hat{k}$ and call this the *empirically centred Gram matrix*. We note also that $\bar{k}(x,x') = \langle \phi_X(x) - \mu_X, \phi_X(x') - \mu_X \rangle$ is a kernel with feature map $\hat{\phi}_X(X) = \phi_X(X) - \hat{\mu}_X$. We denote by $\bar{K}$ the Gram matrix with respect to $\bar{k}$ and call this the *population centred Gram matrix*. We note further here that if k and l are kernels on $\mathcal{X}$ and $\mathcal{Y}$, then $k \otimes l$ is a kernel on $\mathcal{X} \times \mathcal{Y}$. We write $C_{XY} = \mathbb{E}_{XY}\bar{\phi}_X(X) \otimes \bar{\phi}_Y(Y)$ called the *population centred covariance operator* and define $\bar{C}_{XY} = \frac{1}{n}\sum_{i=1}^n \bar{\phi}_X(X_i) \otimes \bar{\phi}_Y(Y_i)$ to be its empirical counterpart. Note that we can consider C_{XY} to be an operator $\mathcal{H}_l \longrightarrow \mathcal{H}_k$, or as an element of the Hilbert space $\mathcal{H}_{k\otimes l}$

2.2. Lancaster

The above ideas of injectively embedding measures into a Hilbert space can be extended from the two variable case to consider properties of three or more variables. The Lancaster statistic on the triple of variables (X, Y, Z) is defined as the signed measure $\Delta_L P = \mathbb{P}_{XYZ} - \mathbb{P}_{XY}\mathbb{P}_Z - \mathbb{P}_{XZ}\mathbb{P}_Y - \mathbb{P}_X\mathbb{P}_{YZ} + 2\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z$. 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

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

Given a finite sample $(X_i, Y_i, Z_i)_{i=1}^n$, the mean embedding of the Lancaster interaction can be empirically estimated as $\Delta_L \hat{P} = \hat{\mu}_{\mathbb{P}_{XYZ}} - \hat{\mu}_{\mathbb{P}_{XY}\mathbb{P}_Z} - \hat{\mu}_{\mathbb{P}_{XZ}\mathbb{P}_Y} - \hat{\mu}_{\mathbb{P}_X\mathbb{P}_{YZ}} + 2\hat{\mu}_{\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z}$. We use the squared RKHS norm of this quantity as a test statistic to test the following hypothesis:

$\mathcal{H}_0 : \mathcal{H}_X \ \vee\ \mathcal{H}_Y \ \vee\ \mathcal{H}_Z$

$\mathcal{H}_1 : \mathbb{P}_{XYZ}$ does not factorise in any way

Given kernels k, l and m on $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ respectively, $k \otimes l \otimes m$ defines a kernel on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. We write K, L and M to denote the gram matrices of each kernel with respect to the observations, where for example $K_{ij} = k(X_i, X_j)$. We further write $\tilde{K}, \tilde{L}$ and $\tilde{M}$ for the empirically centred gram matrices, where for example $\tilde{K}_{ij} = k(X_i, X_j) - \frac{1}{n}\sum_i k(X_i, X_j) - \frac{1}{n}\sum_j k(X_i, X_j) + \frac{1}{n^2}\sum_{ij} k(X_i, X_j)$. We can then write [Lancaster]

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

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

The next part of the statistical test is to find threshold values of the statistic beyond which we would reject the null hypothesis. In the case that the observations are drawn *iid*, this can be done using a permutation bootstrap method. Since our null hypothesis is a composite of three 'sub-hypotheses', we must test each of them separately. We reject the composite null hypothesis if and only if we reject all three of the components. For more information on the details of the bootstrapping method, see [Lancaster].

2.3. Time series

In this paper we are extending the existing Lancaster test from the *iid* case to a case in which our observations are drawn from a random process. There are various formalisations of memory or 'mixing' of a random process; of relevance to this paper are the following two:

2.3.1. τ-MIXING

Definition 1. *A process $(X_t)_t$ is τ-mixing if $\tau(r) \longrightarrow 0$ as $r \longrightarrow \infty$, where*

$$\tau(r) = \sup_{l\in\mathbb{N}} \frac{1}{l} \sup_{r \le i_1 \le \ldots \le i_l} \tau(\mathcal{F}_0, (X_{i_1}, \ldots, X_{i_l})) \longrightarrow 0$$

where

$$\tau(\mathcal{M}, X) = \mathbb{E}(\sup_{g\in\Lambda} |\int g(t)\mathbb{P}_{X|\mathcal{M}}(dt) - \int g(t)\mathbb{P}_X(dt)|)$$

2.3.2. β-MIXING

Definition 2. *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 $A_i \in \mathcal{H}_1^n$ and $B_j \in \mathcal{H}_{n+m}^{\infty}$ and $\mathcal{H}_b^c = \sigma(X_b, X_{b+1}, \ldots, X_c)$

The concept of β-mixing will be invoked when applying a central limit theorem in the next section. We will also need the following lemma:

Lemma 1. *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 $(X_t)_t$)*

2.4. V-statistics

A V-statistic of a k-argument, symmetric function f given *iid* observations $\mathcal{S}_n = \{S_1, \ldots, S_n\}$ where each $S_i \sim \mathbb{P}$ is written

$$V(f, \mathcal{S}) = \frac{1}{n^k} \sum_{1 \leq i_1, \ldots, i_k \leq n} f(S_{i_1}, \ldots, S_{i_k})$$

In this case, $V(f, \mathcal{S})$ is a biased (but asymptotically unbiased) estimator of $\mathbb{E}_{S_{i_1},\ldots S_{i_k} \sim \mathbb{P}}[f(S_{i_1}, \ldots, S_{i_k})]$

In this paper we are only concerned with V-statistics for which $k = 2$. We call $nV(f, \mathcal{S})$ *normalised.* We call f the *core* of V and we say that f is *degenerate* if, for any s_1, $\mathbb{E}_{S_2 \sim \mathbb{P}}[f(s_1, S_2)] = 0$ in which case we say that V is a *degenerate V-statistic*

Of relevance to us is the fact that many kernel test statistics can be viewed as normalised V-statistics which, under the null hypothesis, are degenerate. If moreover the test statistics diverge under the alternative hypothesis, the test would be consistent. Our main result is to prove that the Lancaster statistic is asymptotically a degenerate V-statistic.

2.5. Wild Bootstrap

In many frequentist statistical tests, estimates of the test statistic threshold required to achieve a given Type I error ***ie p-value*** are obtained through a bootstrap resampling method. In the case of the Lancaster and HSIC tests with *iid* observations, this is done by permuting the time indices of one of the variables to simulate samples from the distribution in which the permuted variable is independent of the other(s). However, this procedure relies on the *iid* assumption of the data generating process - if, in fact, subsequent samples are *not* independent of previous samples, then permuting the order of the time indices destroys any backward dependence.

When our test statistics are normalised V-statistics which are degenerate under the null hypothesis, the Wild Bootstrap is a method that comes to our rescue. Rather than directly providing a way to generate new samples, it directly resamples the test statistic subject to certain conditions, which can be categorised as concerning: (1) Appropriate τ-mixing of the process from which our observations are drawn; (2) The core of the V-statistic. If these conditions are met by the statistic $nV(f, \mathcal{S}_n)$, then [Wild Bootstrap] tell us that a random matrix W can be drawn simply such that the bootstrapped statistic $nV_b(f, \mathcal{S}_n) = \frac{1}{n} \sum_{i,j,p,q} W_{ij} f(S_j, S_p) W_{pq}$ is distributed according to the null distribution of nV. The condition on $V(f, \mathcal{S})$ that is of crucial importance to this paper is that f must be a degenerate core.

2.6. Hilbert spaced random variable CLT

Should we actually state the theorem here? We should include the proof that our situation satisfies the conditions of the theorem regardless though, but maybe in the supplementary section.

- Kernel mean embedding
- Lancaster
- Time series
  - τ-mixing
  - β-mixing
  - Lemma that sub-processes of β-mixing processes are β-mixing
- V-statistics
- Hilbert space valued random variable central limit theorem

3. Lancaster Interaction for Random Processes

(Following kacper's paper...)

In this section we construct the Lancaster Interaction test for random processes. The major difficulty in doing so is showing that the test statistic asymptotically satisfies the conditions of the Wild Bootstrap under the null hypothesis of the test, and therefore the Wild Bootstrap can be used to resample the test statistic and provide consistent thresholds for desired p-values.

The approach taken in this paper can also be applied to the HSIC test statistic to give a simpler proof that the Wild Bootstrap can be used for HSIC+timeseries than that given in [Kacper].

Lemma 2. *Suppose that (X_i) is β-mixing with coefficients $\beta(m)$ satisfying $\sum_{m=1}^{\infty} \beta(m)^{\frac{\delta}{2+\delta}} < \infty$ and that k is a bounded kernel on $\mathcal{X}$. Then $\|\hat{\mu}_X - \mu_X\|_k = O(n^{-\frac{1}{2}})$*

Theorem 1. *Suppose that $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$ and that $(X_i, Y_i, Z_i)_{i=1}^n$ are drawn from a process that is both:*

- *β-mixing with coefficients $\beta(m)$ satisfying $\sum_{m=1}^{\infty} \beta(m)^{\frac{\delta}{2+\delta}} < \infty$*
- *τ-mixing with coefficients $\tau(m)$ satisfying $\sum_{m=1}^{\infty} m^2 \sqrt{\tau(m)} < \infty$*

. Then, as $n \longrightarrow \infty$,

$$n\|\Delta_L \hat{P}\|^2 \longrightarrow \frac{1}{n}\left(\overline{\left(\bar{K}\circ\bar{L}\right)}\circ\bar{M}\right)_{++}$$

and this is a normalised degenerate V-statistic.

Corollary 1. *Suppose in addition to the above that W is drawn from a process satisfying the conditions of [Wild Bootstrap]. Then asymptotically,*

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

has the same distribution as $n\|\Delta_L \hat{P}\|^2$.

We can therefore use this to generate samples of the test statistic $n\|\Delta_L \hat{P}\|^2$ under the null hypothesis $\mathcal{H}_Z$. Using these samples we can select a threshold value of the test statistic such that the Type I error is bounded by whatever α we choose. By symmetry, we can use a similar procedure to test $\mathcal{H}_X$ and $\mathcal{H}_Y$.

4. Multiple testing correction

In the Lancaster test, we use a composite null hypothesis which requires us to test each of the three hypotheses $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$ separately. We reject the null hypothesis $\mathcal{H}_0$ if and only if we reject all three of the components. In [Lancaster], it is suggested that the Holm-Bonferroni correction be used to account for multiple testing. 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. WLOG 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 $\mathbb{P}(\mathcal{A}_0) \leq \alpha$.

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, is therefore more conservative than necessary and thus loses on test power compared to the 'correction' proposed here.

5. Experiments

5.1. Artificial data

5.2. Real data

Maybe check this out for some data? `https://stat.duke.edu/~mw/ts_data_sets.html`

6. Proofs

Proof of Lemma 2:

Proof: We exploit Theorem 1.1 from (**?**). 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, $\exists C: \ \|\bar{\phi}(X_i)\| < C$ 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 (**?**)

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

where N is a Hilbert space valued Gaussian random variable. Thus

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

■

Proof of Theorem 1

Proof:

By writing

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

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

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

Each of these terms can be expressed as inner products between empirical estimates of population centred covariance operators and tensor products of mean embeddings, and rewriting them as such gives

$$\begin{aligned}
n\|\Delta_L \hat{P}\|^2 = \; & n\langle \bar{C}_{XYZ}, \bar{C}_{XYZ}\rangle \\
& - 2n\langle \bar{C}_{XYZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle \\
& - 2n\langle \bar{C}_{XZY}, \bar{C}_{XZ} \otimes \bar{\mu}_Y\rangle \\
& - 2n\langle \bar{C}_{YZX}, \bar{C}_{YZ} \otimes \bar{\mu}_X\rangle \\
& + n\langle \bar{C}_{XY} \otimes \bar{\mu}_Z, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle \\
& + n\langle \bar{C}_{XZ} \otimes \bar{\mu}_Y, \bar{C}_{XZ} \otimes \bar{\mu}_Y\rangle \\
& + n\langle \bar{C}_{YZ} \otimes \bar{\mu}_X, \bar{C}_{YZ} \otimes \bar{\mu}_X\rangle \\
& + 2n\langle \bar{\mu}_Z \otimes \bar{C}_{XY}, \bar{C}_{ZX} \otimes \bar{\mu}_Y\rangle \\
& + 2n\langle \bar{\mu}_X \otimes \bar{C}_{YZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle \\
& + 2n\langle \bar{\mu}_X \otimes \bar{C}_{ZY}, \bar{C}_{XZ} \otimes \bar{\mu}_Y\rangle \\
& + 4n\langle \bar{C}_{XYZ}, \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z\rangle \\
& - 4n\langle \bar{C}_{XY} \otimes \bar{\mu}_Z, \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z\rangle \\
& - 4n\langle \bar{C}_{XZ} \otimes \bar{\mu}_Y, \bar{\mu}_X \otimes \bar{\mu}_Z \otimes \bar{\mu}_Y\rangle \\
& - 4n\langle \bar{C}_{YZ} \otimes \bar{\mu}_X, \bar{\mu}_Y \otimes \bar{\mu}_Z \otimes \bar{\mu}_X\rangle \\
& + 4n\langle \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z, \bar{\mu}_X \otimes \bar{\mu}_Y \otimes \bar{\mu}_Z\rangle
\end{aligned}$$

By assumption, $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$ and thus the expectation operator also factorises similarly. As a consequence,

$$\begin{aligned}
C_{XYZ} &= \mathbb{E}_{XYZ}[\bar{\phi}_X(X) \otimes \bar{\phi}_Y(Y) \otimes \bar{\phi}_Z(Z)] \\
&= \mathbb{E}_{XY}[\bar{\phi}_X(X) \otimes \bar{\phi}_Y(Y)] \otimes \mathbb{E}_Z\bar{\phi}_Z(Z) = 0
\end{aligned}$$

Similarly, C_{XZY}, C_{YZX}, C_{XZ}, C_{YZ} are all 0 in their respective Hilbert spaces. Lemma 1 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 2 to each of the covariance operators at the top of this paragraph we see that each of $\|\bar{C}_{XZY}\|$, $\|\bar{C}_{YZX}\|$, $\|\bar{C}_{XZ}\|$, $\|\bar{C}_{YZ}\|$, $\|\bar{\mu}_X\|$, $\|\bar{\mu}_Y\|$, $\|\bar{\mu}_Z\| = O\left(\frac{1}{\sqrt{n}}\right)$

This can be used to show that

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

since all the other terms go to 0 - we show this 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.

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

By treating $\bar{k} \otimes \bar{l}$ as a kernel on the single variable $T := (X, Y)$, we can perform the same recentering trick as before to show that

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

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 of the above expression tend to 0, and thus $n\|\Delta_L \hat{P}\|^2 \longrightarrow \frac{1}{n}((\overline{\bar{K} \circ \bar{L}}) \circ \bar{M})_{++}$ as required.

To show that this is a normalised degenerate V-statistic observe that, writing $S_i = (X_i, Y_i, Z_i)$ and

$h(S_i, S_j) = \langle \bar{\phi}(X_i) \otimes \bar{\phi}(Y_i) - C_{XY}, \bar{\phi}(X_j) \otimes \bar{\phi}(Y_j) - C_{XY} \rangle \langle \bar{\phi}(Z_i), \bar{\phi}(Z_j) \rangle$, we can write:

$$\frac{1}{n}((\overline{\bar{K} \circ \bar{L}}) \circ \bar{M})_{++} = \frac{1}{n} \sum_{ij} h(S_i, S_j)$$

And thus it is a normalised V-statistic. To show that it is degenerate, fix any s_i and observe that $\mathbb{E}_{S_j} h(s_i, S_j) = 0$

■

Proof of Lemma 1:

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

Thus

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

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. ■

Acknowledgments

cheers!