University College London

Masters Thesis for MSc in Computational Statistics and Machine Learning

Three Variable Kernel Independence Testing with Time Series

Author:
Paul Kishan Rubenstein

Supervisor:
Dr. Arthur Gretton

September 2015

Acknowledgements

To Arthur Gretton for supervising me and to Kacper Chwialkowski for a lot of help:

Cheers!

As a society, we must ask ourselves: do we wish that the labour of skilled individuals with great technical expertise be used for the development of technology that facilitates killing and destruction?

"...we must guard against the acquisition of unwarranted influence, whether sought or unsought, by the military-industrial complex. The potential for the disastrous rise of misplaced power exists and will persist."

Dwight D. Eisenhower

Abstract

We apply a Wild Bootstrap method to the Lancaster interaction statistic to detect dependencies between three time series.

The Wild Bootstrap is a method to resample a test statistic given some observed data, subject to certain conditions on both the test statistic and the observed data. The main contribution of this thesis is to prove that the Lancaster interaction satisfies the conditions on the test statistic under the null hypothesis of its statistical test. Furthermore, we present a novel proof that the same is true of the Hilbert Schmidt Independence Criterion (HSIC) statistic.

We demonstrate with this method that the Lancaster interaction is sensitive to dependences between three variables in certain cases that HSIC is not - these are cases in which any two variables interact weakly, but all three share a strong mutual dependency.

For accompanying code, see

`https://github.com/paruby/CSML-Thesis-code-repo`

Contents

1 Introduction

A basic problem in machine learning and statistics is independence testing: Given pairs of observations $\mathcal{D} = \{(X_i, Y_i), \quad i = 1, \ldots, n\}$, are the Xs and Ys independent? The Hilbert Schmidt Independence Criterion (HSIC) [1][2] is a kernel methods approach to answering this question in the case that the observations (X, Y) are drawn *iid* from a distribution $\mathbb{P}_{XY}$. More detail will be given about the test in the following section but in short, HSIC uses the properties of kernels to measure a 'distance' between empirical estimates for the joint distribution $\mathbb{P}_{XY}$ and the product of marginals, $\mathbb{P}_X\mathbb{P}_Y$.

HSIC has been recently extended in two new ways: a test for two variables with time series data [3]; and a test for three variables with *iid* data[4]. This project is an effort to combine these two extensions. See Figure 1 for a graphical representation of this.

First, rather than considering two random variables, we may be interested in three. With three random variables, there are more complicated forms of dependence (or independence) that can exist than with two variables. One question that can be asked is whether one variable is independent of the other two, or if they are all mutually independent - equivalently, we may ask: *"does the joint distribution $\mathbb{P}_{XYZ}$ factorise into a product of marginals in some way?"* The Lancaster interaction [4][5] is non-zero if the answer to the preceding question is *no*, and so using it we can design a statistical test for which rejection of the null hypothesis implies that the joint distribution does *not* factorise. This has applications to conditional independence testing, in which two variables X and Y may be independent when considered only together but become dependent when conditioned on a third, Z. This is equivalent to saying that $\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ but that $\mathbb{P}_{XYZ}$ does not factorise. Such relationships between three variables are known as *V-structures*. Their detection is an important part of causal inference. See eg [6] and [7].

Second, rather than considering *iid* data, we may be interested in time series data. Any frequentist statistical hypothesis test is composed of two parts. We must first construct a test statistic - that is, a function of the observed data. Having

done this, we must see where the value of our test statistic lies in comparison to the distribution of the statistic under the null hypothesis (the *null distribution*). If the value is more 'extreme' than a prearranged threshold, we may reject the null hypothesis.

In general, the null distribution is dependent on the distributions of the underlying variables from which our observations are drawn. Since this is not something we are privy to, we must estimate the null distribution from the existing observations (this is often referred to as *bootstrapping*). The existing methods to do this for HSIC and Lancaster fail when the data are not *iid*.

In a recent paper [8], a method called the *Wild Bootstrap* is presented that extends work previously done by Shao [9]. This is a bootstrap method that can be applied to certain types of statistical tests to simulate samples of the test statistic under the null hypothesis, provided that the observed data are drawn from a process satisfying certain conditions (τ-dependence and stationarity). It has already been shown in [3] that HSIC satisfies the conditions required on the test statistic to use the Wild Bootstrap.

The main contribution of this thesis is to show that, under certain conditions on the observed data, the *Wild Bootstrap* method can be applied to Lancaster statistic. In addition, the proof of this is easily adapted to provide a new, simpler proof that the Wild Bootstrap can be used with HSIC. A second, more minor contribution is to show that the power of the Lancaster test described in [4] can be improved - in trying to account for multiple testing error rates, the authors of this paper use conservative thresholds for p-values. It is shown in this thesis that the thresholds can be relaxed, thus increasing power, while still guaranteeing the desired Type I error.

This thesis is structured as follows. In *Section 2*, we discuss in detail the HSIC and Lancaster statistics, and the concepts required to understand the Wild Bootstrap.

In *Section 3*, the main result of this work is presented - namely, that the Lancaster statistic satisfies the conditions required to be able to use the Wild Bootstrap to resample the statistic under the null hypothesis. A simpler, adapted version of this proof is first given to show that HSIC satisfies the conditions, after which the proof for the Lancaster statistic is given. It is not a new result that the Wild Bootstrap

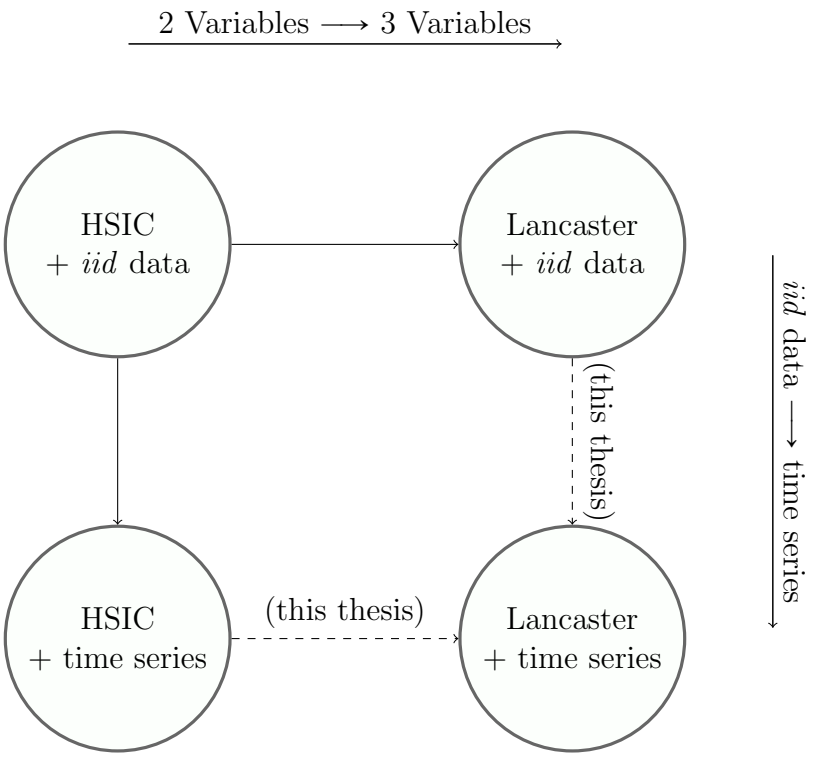

Figure 1: A diagram to show how the work in this thesis relates to existing work.

can be applied to the HSIC statistic; the novel proof given here is, however, arguably simpler than the existing proof given in [10], and demonstrates the main ideas of the proof for the Lancaster statistic in a less algebraically involved setting. This 'preview' will aid the reader in following the proof for the Lancaster statistic.

In *Section 4*, we compare the performance of the Lancaster test developed with two HSIC-based tests on artificial and real forex data. We find that Lancaster outperforms the two HSIC-based tests in situations for which three variables exhibit weak pairwise dependences, but a strong joint (three-way) dependence. In cases for which strong pairwise interactions are present, the Lancaster test does not perform as well.

Section 5 concludes the thesis with a discussion of the results, closing remarks and directions for future research.

2 Background

In this section, the background theory necessary to understand the result of this thesis is introduced. The author suggests that the reader mentally divide this theory into two categories: understanding the HSIC and Lancaster statistical tests in the case that our observations are drawn *iid*; and understanding the conditions that must be met to be able to use the Wild Bootstrap to adapt these tests when our observations are not *iid*.

We will first give a very brief introduction to the theory of *reproducing kernel Hilbert spaces (RKHSs)*. For further information, the interested reader may consult [11], [12],[13], [14] and [15]. Our main objective here is to introduce the *kernel mean embedding* [16], a method that, subject to certain conditions, injectively embeds measures into a Hilbert space and so induces a metric on probability distributions. This method is the basis of our statistical tests.

Next the Hilbert-Schmidt Independence Criterion (HSIC), a statistical test for detecting dependence between two random variables, is introduced for the *iid* case. Then, the Lancaster interaction, which can be viewed as a generalisation of HSIC to three random variables, is introduced, also for the *iid* case.

We then set the stage for describing how we may deal with non-*iid* data. We will give a basic formal definition for certain types of time series. Then we will describe a common class of statistics known as V-statistics. Having done this, we will be able to describe the Wild Bootstrap.

2.1 Kernel basics

Definition 1 (Kernel)**.** *Let $\mathcal{X}$ be a non-empty set. A* kernel on $\mathcal{X}$ *is a function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ for which there exists a Hilbert space $\mathcal{H}$ and a* feature map *$\phi : \mathcal{X} \longrightarrow \mathcal{H}$ such that, for all $x, y \in \mathcal{H}$*

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

Definition 2 (Gram matrix)**.** *Let k be a kernel on $\mathcal{X}$ and suppose that $\mathcal{D} = \{X_1, \ldots, X_n\}$*

is a set of samples from $\mathcal{X}$. The Gram matrix is an $n \times n$ matrix with entries

$$K_{ij} = k(X_i, X_j), \qquad i, j \in \{1, \ldots, n\}$$

Remark (Notation)**.** *Throughout this thesis, we will use the convention that lower case letters are used for kernels and the corresponding upper case letters are used for the corresponding Gram matrices.*

Definition 3 (Reproducing kernel Hilbert space)**.** *Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a* reproducing kernel *of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), if*

(i) $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$

(ii) $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$

Remark (Feature maps)**.** *There may be many possible choices of feature maps to represent any given kernel. We call $\phi(x) = k(\cdot, x)$ as written above the* canonical feature map*. In general, it may not be useful (or even practically possible) to explicitly represent ϕ. Indeed, part of the power of kernels is that we can implicitly work in a high (possibly infinite) dimensional feature space without ever having to explicitly represent it.*

The above definitions say that, given a reproducing kernel Hilbert space, we can take any point $x \in \mathcal{X}$ and embedd it in the RKHS as $\phi(x) = k(\cdot, x)$. This then allows us to evaluate any function $f \in \mathcal{H}$ at x by taking an inner product: $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$.

Under certain conditions, we can extend this notion to embedding probability distributions. This allows us to calculate the *expectation* of a function by taking its inner product with respect to the element in $\mathcal{H}$ corresponding to the embedding of the distribution.

Definition 4 (Mean embedding)**.** *Let $\mathbb{P}_X$ be a probability measure on $\mathcal{X}$ and let $X \sim \mathbb{P}_X$ be a random variable on $\mathcal{X}$. The* mean embedding *$\mu_{\mathbb{P}_X}$ of $\mathbb{P}_X$ is an element of $\mathcal{H}$ satisfying*

$$\mathbb{E}_{X\sim\mathbb{P}_X} f(X) = \langle f, \mu_{\mathbb{P}_X} \rangle$$

for all $f \in \mathcal{H}$

It is not obvious from this definition whether $\mu_{\mathbb{P}_X}$ exists. The following theorem provides a sufficient condition for this:

Theorem 2.1. *Suppose that* k *is measurable and that* $\mathbb{E}_{X\sim\mathbb{P}_X}\sqrt{k(X,X)} < \infty$. *Then* $\mu_{\mathbb{P}_X} \in \mathcal{H}$.

Consequently, if k *is* bounded *(ie* $\exists C$ *s.t.* $k(x,y) \leqslant C \ \forall x, y \in \mathcal{X}$*) then the mean embedding exists for any probability distribution on* $\mathcal{X}$

For a proof of this, see [17]. Throughout this thesis, we will assume that all kernels under consideration are bounded, and hence mean embeddings always exist.

Remark. *For intuition, we can think of the mean embedding of a distribution as the expectation of an* $\mathcal{H}$*-valued random variable* $\phi(X)$*:*

$$\mu_{\mathbb{P}_X} = \mathbb{E}_{X\sim\mathbb{P}_X}[k(\cdot, X)] = \mathbb{E}_{X\sim\mathbb{P}_X}\phi(X)$$

Remark (Distances between probability distributions)**.** *Observe that the kernel mean embedding induces a map*

$$\begin{aligned} \{\textit{Probability distributions on } \mathcal{X}\} &\longrightarrow \mathcal{H} \\ \mathbb{P} &\mapsto \mu_{\mathbb{P}} \end{aligned}$$

Provided this is injective, this induces a metric on $\{$*Probability distributions on* $\mathcal{X}\}$*:*

$$d(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|$$

Kernels for which this map is injective are known as characteristic*[18]. It is not within the scope of this thesis to go into more detail, but it suffices to say that the Gaussian kernel (defined below) is characteristic.*

Definition 5 (Gaussian kernel)**.** *Suppose that $\mathcal{X} \subseteq \mathbb{R}^m$ for some $m \in \mathbb{N}$ The Gaussian kernel with bandwidth parameter $\sigma \in \mathbb{R}_{>0}$ is*

$$k_\sigma(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2})$$

Theorem 2.2. *The Gaussian kernel is indeed a kernel which is moreover characteristic.*

For proof, see Proposition 4.10 and Theorem 4.47 of [14].

Remark. *Many kernel statistical tests, and in particular the ones that will be considered in this thesis, can be viewed as essentially exploiting this 'metric on probability distributions'. For example, the MMD [19] is a test to see whether two samples have come from the same distribution. This uses the fact that $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|^2 = 0 \iff \mathbb{P} = \mathbb{Q}$, and then constructs a test working with finite samples based on this. We are interested here in independence testing - for example, does $\|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}_X \mathbb{P}_Y}\|^2 = 0$? - and so we must build some more theory to be able to consider kernels on more than one random variable.*

We now follow closely [2]

Definition 6 (Hilbert-Schmidt norm)**.** *Suppose that $\mathcal{F}$ and $\mathcal{G}$ are separable Hilbert spaces with orthonormal bases $\{u_i\}$ and $\{v_i\}$ respectively, and suppose that $\mathcal{C} : \mathcal{G} \longrightarrow \mathcal{F}$ is a linear operator. Then, provided that the sum converges, the Hilbert-Schmidt (HS) norm of $\mathcal{C}$ is defined as*

$$\|\mathcal{C}\|_{HS}^2 := \sum_{ij} \langle Cv_i, u_j \rangle_{\mathcal{F}}^2$$

Definition 7 (Hilbert-Schmidt operators)**.** *A linear operator $\mathcal{C} : \mathcal{G} \longrightarrow \mathcal{F}$ is called a Hilbert-Schmidt operator if its HS norm exists. The set of Hilbert-Schmidt operators from $\mathcal{G}$ to $\mathcal{F}$, denoted $HS(\mathcal{G}, \mathcal{F})$, is a separable Hilbert space with inner product*

$$\langle C, D \rangle_{HS} := \sum_{ij} \langle Cv_i, u_j \rangle_{\mathcal{F}} \langle Dv_i, u_j \rangle_{\mathcal{F}}$$

Definition 8 (Tensor product)**.** *Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$. The tensor product operator $f \otimes g : \mathcal{G} \longrightarrow \mathcal{F}$ is defined as*

$$(f \otimes g)h := f\langle g, h\rangle_{\mathcal{G}} \quad \forall h \in \mathcal{G}$$

Remark. *Note that for $f, a \in \mathcal{F}$ and $g, b \in \mathcal{G}$*

$$\begin{aligned}
\langle f \otimes g, a \otimes b\rangle_{HS} &= \sum_{ij} \langle (f \otimes g)v_i, u_j\rangle\langle (a \otimes b)v_i, u_j\rangle \\
&= \sum_{ij} \langle \langle g, v_i\rangle f, u_j\rangle\langle \langle b, v_i\rangle a, u_j\rangle \\
&= \sum_{ij} \langle g, v_i\rangle\langle f, u_j\rangle\langle b, v_i\rangle\langle a, u_j\rangle \\
&= \langle f, a\rangle_{\mathcal{F}}\langle g, b\rangle_{\mathcal{G}}
\end{aligned}$$

And in particular,

$$\begin{aligned}
\|f \otimes g\|^2_{HS} &= \langle f \otimes g, f \otimes g\rangle_{HS} \\
&= \langle f, f\rangle_{\mathcal{F}}\langle g, g\rangle_{\mathcal{G}}
\end{aligned}$$

and so $f \otimes g \in HS(\mathcal{G}, \mathcal{F})$

Remark (Kernels on pairs of variables)**.** *Suppose that X and Y are (not necessarily independent) random variables taking value in $\mathcal{X}$ and $\mathcal{Y}$ respectively. We consider the pair (X, Y) to be a random variable taking value in $\mathcal{X} \times \mathcal{Y}$ with joint distribution $\mathbb{P}_{XY}$.*

Suppose that k is a kernel on $\mathcal{X}$ with canonical feature map ϕ and RKHS $\mathcal{F}$ and l is a kernel on $\mathcal{Y}$ with canonical feature map ψ and RKHS $\mathcal{G}$. We can construct from these a new kernel z on $\mathcal{X} \times \mathcal{Y}$ via

$$
\begin{aligned}
z((x_1, y_1), (x_2, y_2)) &= \langle \phi(x_1) \otimes \psi(y_1), \phi(x_2) \otimes \psi(y_2) \rangle_{HS(\mathcal{G},\mathcal{F})} \\
&= \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{F}} \langle \psi(y_1), \psi(y_2) \rangle_{\mathcal{G}} \\
&= k(x_1, x_2) l(y_1, y_2)
\end{aligned}
$$

Using this, we can embed the joint distribution $\mathbb{P}_{XY}$ *via*

$$
\begin{aligned}
\mu_{\mathbb{P}_{XY}} &= \mathbb{E}_{(X,Y)\sim\mathbb{P}_{XY}}[\phi(X) \otimes \psi(Y)] \\
&= \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)]
\end{aligned}
$$

and the product of marginals

$$
\begin{aligned}
\mu_{\mathbb{P}_X\mathbb{P}_Y} &= \mathbb{E}_{(X,Y)\sim\mathbb{P}_X\mathbb{P}_Y}[\phi(X) \otimes \psi(Y)] \\
&= \mathbb{E}_X\mathbb{E}_Y[\phi(X) \otimes \psi(Y)] \\
&= \mathbb{E}_X[\phi(X)] \otimes \mathbb{E}_Y[\psi(Y)] \\
&= \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_Y}
\end{aligned}
$$

For shorthand, we write μ_{XY} *instead of* $\mu_{\mathbb{P}_{XY}}$*,* μ_X *instead of* $\mu_{\mathbb{P}_X}$*, and* μ_Y *instead of* $\mu_{\mathbb{P}_Y}$*.*

Throughout this thesis, we will assume that all kernels, including those defined over pairs and triples of variables, are characteristic.

2.2 Hilbert-Schmidt Independence Criterion (HSIC)

We are now in a position to discuss HSIC. Again, this will only be a brief overview. For a more in depth explanation, see [1] and [2]. The problem we are interested in

solving is the following:

Problem 1. *Suppose we are given a set of* iid *samples* $\{(X_i, Y_i), i = 1, \ldots, n\}$ *of random variables* X *and* Y *taking value in* $\mathcal{X}$ *and* $\mathcal{Y}$ *respectively. Can we tell if* X *and* Y *are independent?*

Equivalently, we may ask: does the joint distribution on (X, Y) *factorise into the product of marginals? ie, does* $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$*?*

Let us fix some notation. We assume that k is a kernel on $\mathcal{X}$ with associated feature map ϕ, and that l is a kernel on $\mathcal{Y}$ with associated feature map ψ

In [2], the *cross-covariance operator is defined* to be

$$\begin{aligned} C_{XY} &:= \mathbb{E}_{XY}[(\phi(X) - \mu_X) \otimes (\psi(Y) - \mu_Y)] \\ &= \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] - \mu_X \otimes \mu_Y \\ &= \mu_{XY} - \mu_X \otimes \mu_Y \end{aligned}$$

Observe that, since we assume the kernel on (X, Y) to be characteristic, $C_{XY} = 0 \iff \mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$. We can exploit this to construct a statistical test for independence. We first define:

Definition 9 (HSIC)**.** $HSIC[\mathbb{P}_{XY}] = \|C_{XY}\|^2_{HS}$

Observe that $HSIC[\mathbb{P}_{XY}]$ can be written in terms of inner products of mean embeddings:

$$
\begin{aligned}
HSIC[\mathbb{P}_{XY}] &= \|C_{XY}\|^2_{HS} \\
&= \|\mu_{XY} - \mu_X \otimes \mu_Y\|^2 \\
&= \langle \mu_{XY}, \mu_{XY} \rangle - 2\langle \mu_{XY}, \mu_X \otimes \mu_Y \rangle + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle \\
&= \langle \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] \rangle \\
&\quad - 2\langle \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)], \mathbb{E}_X[\phi(X)] \otimes \mathbb{E}_Y[\phi(Y)] \rangle \\
&\quad + \langle \mathbb{E}_X[\phi(X)] \otimes \mathbb{E}_Y[\phi(Y)], \mathbb{E}_X[\phi(X)] \otimes \mathbb{E}_Y[\phi(Y)] \rangle
\end{aligned}
$$

In general we do not have access to $\mathbb{P}_{XY}$ and so we cannot evaluate $HSIC[\mathbb{P}_{XY}]$ (indeed, if we did we would not need to construct a complex statistical test to determine whether or not $\mathbb{P}_{XY}$ factorises!). Suppose that we are given independent samples $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, n\}$ drawn from $\mathbb{P}_{XY}$. We define the following quantities, which we can think of as estimating the mean embeddings as follows[1]:

$$
\begin{aligned}
\tilde{\mu}_{XY} &:= \frac{1}{n} \sum_i \phi(X_i) \otimes \psi(Y_i) \approx \mathbb{E}_{XY}[\phi(X) \otimes \psi(Y)] = \mu_{XY} \\
\tilde{\mu}_X &:= \frac{1}{n} \sum_i \phi(X_i) \approx \mathbb{E}_X \phi(X) = \mu_X \\
\tilde{\mu}_Y &:= \frac{1}{n} \sum_i \phi(Y_i) \approx \mathbb{E}_Y \phi(Y) = \mu_X
\end{aligned}
$$

This gives rise to a (biased) estimate which can be expressed in terms of the Gram matrices K and L:

[1]This is a useful intuition for understanding HSIC, but to avoid having to consider technicalities we will not make this notion precise.

$$
\begin{aligned}
HSIC_b[\mathcal{D}] &= \|\tilde{\mu}_{XY} - \tilde{\mu}_X \otimes \tilde{\mu}_Y\|^2 \\
&= \langle \tilde{\mu}_{XY}, \tilde{\mu}_{XY} \rangle - 2\langle \tilde{\mu}_{XY}, \tilde{\mu}_X \otimes \tilde{\mu}_Y \rangle + \langle \tilde{\mu}_X \otimes \tilde{\mu}_Y, \tilde{\mu}_X \otimes \tilde{\mu}_Y \rangle \\
&= \langle \frac{1}{n} \sum_i [\phi(X_i) \otimes \psi(Y_i)], \frac{1}{n} \sum_j [\phi(X_j) \otimes \psi(Y_j)] \rangle \\
&\quad - 2\langle \frac{1}{n} \sum_i [\phi(X_i) \otimes \psi(Y_i)], [\frac{1}{n} \sum_j \phi(X_j)] \otimes [\frac{1}{n} \sum_r \phi(Y_r)] \rangle \\
&\quad + \langle [\frac{1}{n} \sum_i \phi(X_i)] \otimes [\frac{1}{n} \sum_j \phi(Y_j)], [\frac{1}{n} \sum_r \phi(X_r)] \otimes [\frac{1}{n} \sum_s \phi(Y_s)] \rangle \\
&= \frac{1}{n^2} \sum_{ij} \langle \phi(X_i), \phi(X_j) \rangle \langle \psi(Y_i), \psi(Y_j) \rangle \\
&\quad - \frac{2}{n^3} \sum_{ijr} \langle \phi(X_i), \phi(X_j) \rangle \langle \psi(Y_i), \psi(Y_r) \rangle \\
&\quad + \frac{1}{n^4} \sum_{ijrs} \langle \phi(X_i), \phi(X_r) \rangle \langle \psi(Y_j), \psi(Y_s) \rangle \\
&= \frac{1}{n^2} \sum_{ij} K_{ij} L_{ij} - \frac{2}{n^3} \sum_{ijr} K_{ij} L_{ir} + \frac{1}{n^4} \sum_{ijrs} K_{ir} L_{js}
\end{aligned}
$$

This empirical estimate of $HSIC$ converges to its population (ie true) value at a rate of $O(1/\sqrt{n})$ [20][2].

We wish to use this function of the observations as test statistic. The null hypothesis in our test, $\mathcal{H}_0$, is that $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$. The alternative hypothesis $\mathcal{H}_1$, is that $\mathbb{P}_{XY} \neq \mathbb{P}_X \mathbb{P}_Y$. Thus, under $\mathcal{H}_0$, $HSIC_b[\mathcal{D}] \longrightarrow 0$ and under $\mathcal{H}_1$, $HSIC_b[\mathcal{D}] \longrightarrow c$ for some $c > 0$ as the number of data grow.

$HSIC_b[\mathcal{D}]$ is a random variable and thus has some distribution. For any finite sample size we need to be able to estimate the distribution under the null hypothesis in order to be able to find a threshold value of the test statistic at which we would reject the null hypothesis. We do this using a *permutation bootstrap* method to simulate draws of the statistic from its null distribution.

Observe that samples (X_i, Y_i) are *iid* for different i, and therefore X_i will be independent of Y_j for $i \neq j$ even if X and Y are dependent. By applying a permutation π to the indices of one of the variables, we can therefore simulate a sample from the distribution $\mathbb{P}_X \mathbb{P}_Y$:

$$\mathcal{D}' = \{(X_i, Y_{\pi(i)}),\ \ i = 1, \ldots, n\}$$

By simulating a large number of draws in this way, we can estimate the null distribution of the test statistic and calculate a threshold value, given a desired Type I error α.

In summary, the statistical test described above, HSIC, gives us a procedure for which rejection of the null implies that $\mathbb{P}_{XY}$ does not factorise. Since $HSIC_b[\mathcal{D}] \longrightarrow 0 \iff \mathcal{H}_0$ holds, the probability of falsely accepting the null hypothesis tends to zero for any fixed (non-factorising) distribution, and so HSIC is consistent.

2.3 Lancaster statistic

The Lancaster test is a natural generalisation of HSIC to three random variables, however as we will see it lacks consistency. We are interested in answering the following:

Problem 2. *Suppose we are given a set of* iid *samples* $\{(X_i, Y_i, Z_i), i = 1, \ldots, n\}$ *of random variables* X*,* Y *and* Z *taking value in* $\mathcal{X}$*,* $\mathcal{Y}$ *and* $\mathcal{Z}$ *respectively. Can we tell if any of the variables are independent of the others?*

Equivalently, we may ask: does the joint distribution on (X, Y, Z) *factorise in some way? (For example, does* $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$*?)*

In an ideal world, we would like to be able to construct a statistical test that returns a definitive *yes* or *no* to the above question, subject to diminishing Type I and II errors as the number of observations grows.

The Lancaster test does not quite do this. Instead, it is a test for which rejection of the null hypothesis implies that the joint distribution does not factorise. If the

null hypothesis is not rejected, we cannot conclude whether the joint distribution factorises or not, as we will see shortly.

Before we describe the test, let us first fix some notation. We suppose that X, Y and Z are random variables taking value in $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ respectively. We suppose that k, l and m are kernels on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ with canonical feature maps ϕ, ψ and ω respectively. We are given *iid* samples $\mathcal{D} = \{(X_i, Y_i, Z_i), i = 1, \ldots, n\}$. The Gram matrices with respect to these data are written K, L and M.

The Lancaster interaction measure is a signed measure, defined as follows:

$$\Delta_L P = \mathbb{P}_{XYZ} - \mathbb{P}_{XY}\mathbb{P}_Z - \mathbb{P}_{XZ}\mathbb{P}_Y - \mathbb{P}_X\mathbb{P}_{YZ} + 2\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z$$

Claim 2.1. *If any variable is independent of the other two, then the Lancaster interaction vanishes. That is,*

$$(X, Y) \perp\!\!\!\perp Z \vee (X, Z) \perp\!\!\!\perp Y \vee (Y, Z) \perp\!\!\!\perp X \implies \Delta_L P = 0$$

Proof: By symmetry, it suffices to consider the case $(X, Y) \perp\!\!\!\perp Z$.

In this case, $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$. By marginalising out X or Y, we obtain $\mathbb{P}_{YZ} = \mathbb{P}_Y\mathbb{P}_Z$ and $\mathbb{P}_{XZ} = \mathbb{P}_X\mathbb{P}_Z$. Thus

$$\begin{aligned}
\Delta_L P &= \mathbb{P}_{XYZ} - \mathbb{P}_{XY}\mathbb{P}_Z - \mathbb{P}_{XZ}\mathbb{P}_Y - \mathbb{P}_X\mathbb{P}_{YZ} + 2\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z \\
&= \mathbb{P}_{XY}\mathbb{P}_Z - \mathbb{P}_{XY}\mathbb{P}_Z - \mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z - \mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z + 2\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z \\
&= 0
\end{aligned}$$

■

Unfortunately, the reverse implication does not hold - see Table 1 for an example of three binary variables with non-factorising joint distribution but vanishing Lancaster interaction[2].

Generalising the tensor product notation introduced in the previous section, we can define the kernel $k \otimes l \otimes m$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ with canonical feature map $\phi \otimes \psi \otimes \omega$ as:

[2]Note that this Table has been lifted directly from [4], and is not my own work.

Table 1: An example of three binary variables for which the joint distribution does not factorise, yet whose Lancaster interaction is zero. Note that this table is taken from [4] as is not my own work.

$P(0,0,0) = 0.2$	$P(0,0,1) = 0.1$
$P(0,1,0) = 0.1$	$P(0,1,1) = 0.1$
$P(1,0,0) = 0.1$	$P(1,0,1) = 0.1$
$P(1,1,0) = 0.1$	$P(1,1,1) = 0.2$

$$\begin{aligned} k \otimes l \otimes m((x_1, y_1, z_1), (x_2, y_2, z_2)) &= \langle \phi(x_1) \otimes \psi(y_1) \otimes \omega(z_1), \phi(x_2) \otimes \psi(y_2) \otimes \omega(z_2) \rangle \\ &= \langle \phi(x_1), \phi(x_2) \rangle \langle \psi(y_1), \psi(y_2) \rangle \langle \omega(z_1), \omega(z_2) \rangle \\ &= k(x_1, x_2) l(y_1, y_2) m(z_1, z_2) \end{aligned}$$

As such, we can define the mean embedding of the Lancaster interaction measure. For simplicity, we will refer to the embedded version of the signed measure also as $\Delta_L P$:

$$\begin{aligned} \Delta_L P &= \mu_{\mathbb{P}_{XYZ}} - \mu_{\mathbb{P}_{XY}\mathbb{P}_Z} - \mu_{\mathbb{P}_{XZ}\mathbb{P}_Y} - \mu_{\mathbb{P}_X\mathbb{P}_{YZ}} + 2\mu_{\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z} \\ &= \mathbb{E}_{(X,Y,Z)\sim\mathbb{P}_{XYZ}} \phi(X) \otimes \psi(Y) \otimes \omega(Z) - \mathbb{E}_{(X,Y,Z)\sim\mathbb{P}_{XY}\mathbb{P}_Z} \phi(X) \otimes \psi(Y) \otimes \omega(Z) \\ &\quad - \mathbb{E}_{(X,Y,Z)\sim\mathbb{P}_{XZ}\mathbb{P}_Y} \phi(X) \otimes \psi(Y) \otimes \omega(Z) - \mathbb{E}_{(X,Y,Z)\sim\mathbb{P}_X\mathbb{P}_{YZ}} \phi(X) \otimes \psi(Y) \otimes \omega(Z) \\ &\quad + 2\mathbb{E}_{(X,Y,Z)\sim\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z} \phi(X) \otimes \psi(Y) \otimes \omega(Z) \\ &= \mathbb{E}_{XYZ} \phi(X) \otimes \psi(Y) \otimes \omega(Z) - \mathbb{E}_{XY}\mathbb{E}_Z \phi(X) \otimes \psi(Y) \otimes \omega(Z) \\ &\quad - \mathbb{E}_{XZ}\mathbb{E}_Y \phi(X) \otimes \psi(Y) \otimes \omega(Z) - \mathbb{E}_X\mathbb{E}_{YZ} \phi(X) \otimes \psi(Y) \otimes \omega(Z) \\ &\quad + 2\mathbb{E}_X\mathbb{E}_Y\mathbb{E}_Z \phi(X) \otimes \psi(Y) \otimes \omega(Z) \end{aligned}$$

The squared norm $\|\Delta_L P\|^2_{\phi\otimes\psi\otimes\omega}$ is calculated by taking the inner product $\langle \Delta_L P, \Delta_L P \rangle$. This can therefore be expressed in terms of expectations of

$$\langle \phi(X), \phi(X') \rangle \langle \psi(Y), \psi(Y') \rangle \langle \omega(Z), \omega(Z') \rangle$$

with respect to different factorisations of the joint.

Note that without knowing the distribution $\mathbb{P}_{XYZ}$, the we cannot calculate the squared norm. Given the observations $\mathcal{D}$, we can empirically estimate each of the $\mu_{\mathbb{P}}$ with an appropriate average over points in feature space. We can then use these empirical estimates to get an empirical estimate of the squared norm $\|\Delta_L P\|^2_{\phi\otimes\psi\otimes\omega}$. The empirical estimates we use are:

$$\begin{aligned}
\hat{\mu}_{\mathbb{P}_{XYZ}} &= \frac{1}{n}\sum_i \phi(X_i)\otimes\psi(Y_i)\otimes\omega(Z_i) \\
\hat{\mu}_{\mathbb{P}_{XY}\mathbb{P}_Z} &= \frac{1}{n^2}\sum_{ij} \phi(X_i)\otimes\psi(Y_i)\otimes\omega(Z_j) \\
\hat{\mu}_{\mathbb{P}_{XZ}\mathbb{P}_Y} &= \frac{1}{n^2}\sum_{ij} \phi(X_i)\otimes\psi(Y_j)\otimes\omega(Z_i) \\
\hat{\mu}_{\mathbb{P}_X\mathbb{P}_{YZ}} &= \frac{1}{n^2}\sum_{ij} \phi(X_i)\otimes\psi(Y_j)\otimes\omega(Z_j) \\
\hat{\mu}_{\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z} &= \frac{1}{n^3}\sum_{ijk} \phi(X_i)\otimes\psi(Y_j)\otimes\omega(Z_k)
\end{aligned}$$

Observe that the inner products between any two of these empirical mean embeddings can be expressed in terms of the Gram matrices. For example,

$$\begin{aligned}
\langle \hat{\mu}_{\mathbb{P}_{XYZ}}, \hat{\mu}_{\mathbb{P}_{XY}\mathbb{P}_Z} \rangle &= \langle \frac{1}{n}\sum_i \phi(X_i)\otimes\psi(Y_i)\otimes\omega(Z_i), \frac{1}{n^2}\sum_{jk}\phi(X_j)\otimes\psi(Y_j)\otimes\omega(Z_k)\rangle \\
&= \frac{1}{n^3}\sum_{ijk}\langle \phi(X_i)\otimes\psi(Y_i)\otimes\omega(Z_i), \phi(X_j)\otimes\psi(Y_j)\otimes\omega(Z_k)\rangle \\
&= \frac{1}{n^3}\sum_{ijk}\langle \phi(X_i),\phi(X_j)\rangle\langle\psi(Y_i),\psi(Y_j)\rangle\langle\omega(Z_i),\omega(Z_k)\rangle \\
&= \frac{1}{n^3}\sum_{ijk} K_{ij}L_{ij}M_{ik}
\end{aligned}$$

In this manner we can express $\|\Delta_L P\|^2_{\phi\otimes\psi\otimes\omega}$ as a sum of 15 separate terms[3], each written in terms of the Gram matrices.

Remark. *To summarise what we have done thus far: we have constructed a function of the observed data (a* test statistic*) which estimates a quantity that, assuming that one of the variables X, Y or Z is independent of the other two, is zero.*

Therefore, if our test statistic is 'large' then we can conclude that none of the variables are independent of the others; that is to say that the joint distribution $\mathbb{P}_{XYZ}$ does not factorise.

To determine whether our test statistic is 'large', we need to compare its value to the distribution of the statistic under the null distribution. As with HSIC, we use a permutation bootstrap *method to simulate samples under the null hypothesis.*

Recall that the null hypothesis, $\mathcal{H}_0$, is that $\mathbb{P}_{XYZ}$ factorises in some way. Since there are three ways in which this can happen (namely $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z, \mathbb{P}_{XZ}\mathbb{P}_Y$ or $\mathbb{P}_X\mathbb{P}_{YZ}$, noting that the case $\mathbb{P}_{XYZ} = \mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z$ is subsumed by each of these), we must test the three sub-hypotheses separately. If we reject all of the sub-hypotheses, then we reject the whole null hypothesis $\mathcal{H}_0$. Otherwise, we fail to reject the null. In this case, we cannot conclude anything about whether or not the joint factorises.

[3]There are $\binom{5}{2} = 10$ combinations inner products of different terms, and 5 inner products of each term with itself.

To test each sub-hypothesis, we must estimate the corresponding null distribution. By symmetry, it suffices to explain how to estimate the distribution of the statistic under the assumption that $(X, Y) \perp\!\!\!\perp Z$. Similarly to HSIC, we perform a permutation bootstrap procedure to generate simulated samples. By applying a permutation π to the indices of the Z_i, we generate a bootstrapped sample for which the dependence between (X, Y) and Z is broken:

$$\mathcal{D}' = \{(X_i, Y_i, Z_{\pi(i)}), \;\; i = 1, \ldots, n\}$$

By generating many samples in this way, we can estimate any desired quantile for the threshold value of the statistic required to reject the null hypothesis.

As a footnote, observe finally that in the event that $\mathbb{P}_{XYZ} = \mathbb{P}_X \mathbb{P}_Y \mathbb{P}_Z$, the population Lancaster statistic will also be zero. Hence we can use this statistic to create a test for *total independence*. We can simulate draws from the distribution of the statistic under the assumption of total independence by permuting the indices of two variables: for distinct permutations π and σ, let

$$\mathcal{D}' = \{(X_i, Y_{\pi(i)}, Z_{\sigma(i)}), \;\; i = 1, \ldots, n\}$$

We will not explore this further in this thesis.

2.4 Resampling and the Wild Bootstrap

In the descriptions of the procedures for both the HSIC and Lancaster statistical tests above, recall that in order to simulate samples of the test statistic under the null distribution, we permuted the indices of one of the variables. In the case that our observations are not *iid* but are rather drawn from a process with some temporal dependence, this procedure will fail to simulate samples from the correct distribution. Indeed, suppose we take a draw from such a process $(X_i)_{i=1}^n$ and scramble the indices to get a new, simulated draw $(X_{\pi(i)})_{i=1}^n$. The temporal dependence between consecutive X_is will be broken and so the simulated draw will not have the same statistical properties as the true sample.

It follows that we were to try using the permutation methods with non-*iid* data, we may incorrectly set the threshold value of the test statistic required to reject the null hypothesis for any given specificity since we would in fact not be simulating samples from the null distribution of the test statistic[4].

Instead, we need a more sophisticated bootstrap resampling method. The Wild Bootstrap [8] is a scheme that can be used to simulate samples under the null distribution, provided that certain conditions are satisfied. These conditions fall into two categories: first, conditions on the underlying process from which the observations are drawn; second, conditions on the test statistic itself.

Before discussing the Wild Bootstrap itself, it is first necessary to provide background information to understand these conditions. In the following subsections, we will first introduce some formal concepts relating to time series, then the concept of a V-statistic, after which we will be well-equipped to discuss the Wild Bootstrap.

2.4.1 Timeseries

In this thesis we are concerned with extending the HSIC and Lancaster tests to situations in which the *iid* assumption on the observations does not hold. We consider time series, ie data drawn from a random process in which successive observations are dependent on previous observations. There are various formalisations of this 'memory' or mixing. Here we consider the two which are relevant to this thesis. For more information about mixing, see [21][22][23].

Definition 10. *A process* $(X_t)_t$ *is* τ-mixing *if* $\tau(r) \longrightarrow 0$ *as* $r \longrightarrow \infty$*, where*

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

where

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

[4]This is explored further in Example 4 in the Experiments section

Definition 11. *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* $A_i \in \mathcal{A}_1^n$ *and* $B_j \in \mathcal{A}_{n+m}^{\infty}$ *and* $\mathcal{A}_b^c = \sigma(X_b, X_{b+1}, \ldots, X_c)$

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 2.1. *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* $(X_t)_t$*)*

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

Thus

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

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. ■

2.4.2 V-statistics (and U-statistics)

For an in-depth introduction to V- and U-statistics, see [24].

Suppose that $X_1, X_2, \ldots$ are drawn *iid* from a distribution $\mathbb{P}$, and that $\theta = \theta(\mathbb{P})$ is a function of the distribution for which there is an unbiased estimator, where h is a symmetric function of the observations [24]:

$$\theta(\mathbb{P}) = \mathbb{E}_{X_i \sim \mathbb{P}}[h(X_1, \ldots, X_m)]$$

We call h a *core*, and we call its number of arguments its *degree*. Given observations $\mathcal{S}_n = \{X_1, X_2, \ldots, X_n\}$ with $n \geqslant m$, the corresponding U-statistic is denoted

$$U = U(h, \mathcal{S}_n) = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, X_{i_2}, \ldots, X_{i_m})$$

where c runs over each of the $\binom{n}{m}$ choices of m distinct elements $\{i_1, i_2, \ldots, i_m\}$ from $\{1, 2, \ldots, n\}$

V-statistics are closely related. Instead of summing over distinct observations X_i, we sum over all combinations of the observations of size m with replacement. That is,

$$V = V(h, \mathcal{S}_n) = \frac{1}{n^m} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \ldots \sum_{i_m=1}^{n} h(X_{i_1}, X_{i_2}, \ldots, X_{i_m})$$

U is an unbiased estimator, and moreover it is the *minimum variance unbiased estimator* for θ given the observations $\mathcal{S}_n$ [24].

V-statistics are not unbiased, though they asymptotically approach their U-statistic counterparts, with convergence that is dependent on properties of the core h.

In particular, we draw attention to the fact that both $U(h, \mathcal{S}_n)$ and $V(h, \mathcal{S}_n)$ converge to the expectation of the core h as $n \longrightarrow \infty$.

We say that $nV(h, \mathcal{S}_n)$ is a *normalised* V-statistic. In this thesis we will restrict ourselves to considering V-statistics of degree two, that is $h = h(X_1, X_2)$. We say that such a core is *degenerate* if, for any x_1 and for $X_2 \sim \mathbb{P}$, $\mathbb{E}_{X_2}[h(x_1, X_2)] = 0$. If a V-statistic has a degenerate core, we say that it is a *degenerate V-statistic*.

Note that if $V(h, \mathcal{S}_n)$ is degenerate, then $V \longrightarrow \mathbb{E}_{X_1,X_2}[h(X_1, X_2)] = 0$ as $n \longrightarrow \infty$. It can be shown that a normalised V-statistic with a degenerate core converges to a random variable [24].

Observe that if $\mathbb{E}_{X_1,X_2}[h(X_1, X_2)] \neq 0$, then $nV(h, \mathcal{S}_n) \longrightarrow \infty$ as $n \longrightarrow \infty$.

Relevant to the above two statements is the fact that many kernel test statistics can be viewed as normalised V-statistics which, under the null hypothesis, are degenerate. If under the alternative hypothesis the test statistic diverges, then the test is consistent.

The main result of this thesis is to show that, under the null hypothesis, they are asymptotically equal to degenerate V-statistics.

2.4.3 The Wild Bootstrap

We are now equipped to discuss the Wild Bootstrap. Recall that the problem we face with using time series for HSIC and Lancaster is that the permutation methods to resample from the null distribution of the test statistics fails. The Wild Bootstrap provides a method to sample from the correct null distribution.

Suppose we have a test statistic that is a normalised V-statistic nV with core h of degree two, and suppose further that [8][10]:

(1) *Conditions on the observations:* Our observations $\mathcal{D}_n = (S_t)_{t=1}^n$ are drawn from a strictly stationary, τ-dependent process with $\sum_{r=1}^{\infty} r^2 \sqrt{\tau(r)} < \infty$

(2) *Conditions on the test statistic*:

   (i) h is a bounded kernel on $\mathcal{S}$, where each S_i takes value in $\mathcal{S}$

   (ii) h is degenerate as a core.

   (iii) h is Lipschitz continuous.

Our normalised V-statistic can be written

$$nV(h, \mathcal{D}_n) = \frac{1}{n} \sum_{ij} h(S_i, S_j)$$

We define the following two bootstrapped statistics [10]:

$$nV_{b1}(h, \mathcal{D}_n) := \frac{1}{n} \sum_{ij} W_{i,n} W_{j,n} h(S_i, S_j)$$
$$nV_{b2}(h, \mathcal{D}_n) := \frac{1}{n} \sum_{ij} \tilde{W}_{i,n} \tilde{W}_{j,n} h(S_i, S_j)$$

where $(W_{t,n})_{1 \leqslant t \leqslant n}$ is known as the *auxiliary wild bootstrap process* and $\tilde{W}_{t,n} = W_{t,n} - \frac{1}{n} \sum_i W_{i,n}$. We suppose that this process satisfies the following assumption [10]:

(3) *Conditions on bootstrap process*: $(W_{t,n})_{1 \leqslant t \leqslant n}$ is a row-wise strictly stationary triangular aray independent of all S_t such that $\mathbb{E}W_{t,n} = 0$ and $\sup_n \mathbb{E}|W_{t,n}^{2+\sigma}| < \infty$ for some $\sigma > 0$. The autocovariance of the process is given by $cov(W_{s,n}, W_{t,n}) = \rho(|s-t|/l_n)$ for some function ρ, such that $\lim_{u \longrightarrow 0} \rho(u) = 1$ and $\sum_{r=1}^{n-1} \rho(|r|/l_n) = O(l_n)$. The sequence (l_n) is taken such that $l_n = o(n)$ and $lim_{n \longrightarrow \infty} l_n = \infty$. The variables $W_{t,n}$ are τ-weakly dependent with coefficients $\tau(r) \leqslant C\zeta^{\frac{r}{l_n}}$ for $r = 1, ..., n$, $\zeta \in (0, 1)$ and $C < \infty$

A simple example of a process satisfying these properties is given by [10][8]:

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

where $W_{0,n}$ and $\epsilon_1, \ldots, \epsilon_n$ are independent standard normal random variables. [8] demonstrates the following result.

Theorem 2.3 (Leucht)**.** *Assume that conditions* (1), (2) *and* (3) *above hold. Then* nV, nV_{b1} *and* nV_{b2} *converge to the same distribution. In particular,*

$$nV,\ \ nV_{b1},\ \ nV_{b2} \xrightarrow{d} Z := \sum_k \lambda_k Z_k^2$$

where $(Z_k)_k$ *is a sequence of centred, jointly normal random variables with* $cov(Z_j, Z_k) = \sum_{r=-\infty}^{\infty} cov(\Phi_j(X_0), \Phi_k(X_r))$*, and* $(\lambda_k)_k$ *and* $(\Phi_k)_k$ *are the sequences of non-zero eigenvalues and the corresponding eigenfunctions of* $\mathbb{E}[h(x, X_0)\Phi(X_0)] = \lambda\Phi(x)$

The particular distribution to which the test statistic and the bootstrapped statistics converge is not particularly important. The important thing about this theorem is that firstly, the test statistic does converge to some random variable, and that secondly we can simulate samples from this distribution using the bootstrapped statistics. Thus, provided that our observations satisfy (1) and that under the null hypothesis our test statistic satisfies (2), we now have a method to estimate the null distribution of our test statistic.

In this thesis we are primarily concerned with condition (2) above: the aim is essentially to show that the test statistics for HSIC and Lancaster satisfy this. However, before we continue it is worth briefly considering conditions (1) and (3) as they suggest further directions for research.

When can we say with confidence that observations have indeed been drawn from a process satisfying (1)?

As stated in the explanation of (3), it is relatively simple to write down a process that satisfies (3). However, there will exist many other processes that one could consider - indeed, even in the one written down, the choice of l_n is open. As mentioned in [8], in a similar method by Shao [9], the parameter choice of l_n was not found to be very important. Further investigation into this or of other choices of bootstrap processes would be interesting, though in a different direction to the nature of the work done in this thesis.

2.5 Summary

- HSIC is a statistical test for detecting dependence between two random variables.

- Lancaster is a statistical test for detecting dependence between three random variables. If we reject the null hypothesis, we conclude that there is no factorisation of the joint; if we do not reject the null hypothesis, we cannot conclude anything.

- When the observed data are drawn *iid*, it is possible to resample from the null distributions for both HSIC and Lancaster using permutation resampling. When the observed data are *not iid*, this method is not valid.

- The Wild Bootstrap is a method that enables simulated samples of a test statistic to be drawn under certain conditions. These are:

  (i) The observed data must be drawn from a τ-mixing process

  (ii) The test statistic must be a V-statistic with a core that is a degenerate, Lipschitz continuous kernel

The objective of the following section is to show that, under their null hypotheses, Lancaster and HSIC satisfy condition (ii) above.

3 Main theoretical result

In this section, a result is proved that implies the Wild Bootstrap can be applied to the Lancaster test statistic. This is the major new theoretical contribution of this work. Furthermore, this proof can be adapted to give a simpler proof that the Wild Bootstrap can be applied to HSIC. This is not new knowledge, however the previous proof given in [3] was longer and relied on more advanced theory of U- and V- statistics, including the Hoeffding decomposition [24]. The reader should note that although the new proof given here is shorter, some of the complexity of the argument is simply deferred to the proof of a Hilbert space Central Limit Theorem for time series [25].

The result which will be proved is that, under their respective null hypotheses, the normalised HSIC and Lancaster statistics are asymptotically V-statistics with cores that are degenerate kernels.

The layout of this section is as follows. First, notation will be set up. Second, the result will be proved for HSIC. Third, the result will be proved for Lancaster. The proof ideas are essentially the same for the two statistics; the main difference is that the proof for Lancaster involves considering the asymptotic properties of many more terms than HSIC and is for this reason significantly longer, though not particularly more conceptually involved.

3.1 Notation

Let k be a kernel with canonical feature map ϕ. Given observations $X_i, \quad i = 1, 2, \dots n$ let K be the gram matrix such that $K_{ij} = k(X_i, X_j) = \langle \phi(X_i), \phi(X_j) \rangle$.

Let $\tilde{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$ and let $\tilde{K}_{ij} = \langle \phi(X_i) - \tilde{\mu}_X, \phi(X_j) - \tilde{\mu}_X \rangle$. We call $\tilde{K}$ an *empirically* centred gram matrix.

Let $\mu_X = \mathbb{E}_X \phi(X)$ and let $\bar{K}_{ij} = \langle \phi(X_i) - \mu_X, \phi(X_j) - \mu_X \rangle$. We call $\bar{K}$ an *population* centred gram matrix.

Note that we cannot in general explicitly represent the feature mapped observations, and even when we can it may not be desirable to do so. We are, however, able to write the empirically centred gram matrix $\tilde{K}$ just in terms of values of the

original matrix K, and the population centred gram matrix $\bar{K}$ in terms of K and expectations of the kernel k. Indeed, by simply expanding the sum:

$$\begin{aligned}
\tilde{K}_{ij} &= \langle \phi(X_i) - \tilde{\mu}_X, \phi(X_j) - \tilde{\mu}_X \rangle \\
&= \langle \phi(X_i) - \frac{1}{n}\sum_k \phi(X_k), \phi(X_j) - \frac{1}{n}\sum_l \phi(X_l) \rangle \\
&= \langle \phi(X_i), \phi(X_j) \rangle - \frac{1}{n}\sum_k \langle \phi(X_k), \phi(X_j) \rangle - \frac{1}{n}\sum_l \langle \phi(X_i), \phi(X_l) \rangle + \frac{1}{n^2}\sum_k \sum_l \langle \phi(X_k), \phi(X_l) \rangle \\
&= K_{ij} - \frac{1}{n}\sum_k K_{kj} - \frac{1}{n}\sum_l K_{il} + \frac{1}{n^2}\sum_{kl} K_{kl}
\end{aligned}$$

For notational ease, we here introduce the following notational convention. For any matrix A, define

$$\begin{aligned}
A_{i+} &= \sum_j A_{ij} \\
A_{+j} &= \sum_i A_{ij} \\
A_{++} &= \sum_{ij} A_{ij}
\end{aligned}$$

If additionally A is symmetric, we write

$$(A_+)_{ij} = A_{+j} = A_{j+}$$

Hence we can write

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{n}K_{+j} - \frac{1}{n}K_{i+} + \frac{1}{n^2}K_{++}$$

Observe that this gives a method to convert an empirically-uncentred gram matrix into a centred one. For each entry, subtract the corresponding row and column averages and add the whole matrix average.

Remark. *The average value of any row or column of an empirically centred gram matrix is zero*

We can expand the population centred $\bar{K}$ in a similar fashion, expressing it in terms of expectations of the original kernel function k. In the following, X and X' are independent copies of the original random variable.

$$\begin{aligned}
\bar{K}_{ij} &= \langle \phi(X_i) - \mu_X, \phi(X_j) - \mu_X \rangle \\
&= \langle \phi(X_i) - \mathbb{E}_X \phi(X), \phi(X_j) - \mathbb{E}_{X'} \phi(X') \rangle \\
&= \langle \phi(X_i), \phi(X_j) \rangle - \mathbb{E}_X \langle \phi(X), \phi(X_j) \rangle - \mathbb{E}_{X'} \langle \phi(X_i), \phi(X') \rangle + \mathbb{E}_X \mathbb{E}_{X'} \langle \phi(X), \phi(X') \rangle \\
&= k(X_i, X_j) - \mathbb{E}_X k(X, X_j) - \mathbb{E}_{X'} k(X_i, X') + \mathbb{E}_X \mathbb{E}_{X'} k(X, X')
\end{aligned}$$

Similarly, we define the population centred kernel $\bar{k}$ similar to the above.

$$\bar{k}(X_i, X_j) = k(X_i, X_j) - \mathbb{E}_X k(X, X_j) - \mathbb{E}_{X'} k(X_i, X') + \mathbb{E}_X \mathbb{E}_{X'} k(X, X')$$

We define the *population centred* feature map $\bar{\phi}(X) := \phi(X) - \mu_X$. Note that since $\bar{k}$ corresponds to an inner product between points mapped under the feature map $\bar{\phi}$, $\bar{k}$ is indeed a valid kernel.

Remark. *$\bar{k}$ is a degenerate kernel. Indeed, observe also that for any x,*

$$\begin{aligned}\mathbb{E}_{X*}\bar{k}(x, X^*) &= \mathbb{E}_{X*}k(x, X^*) - \mathbb{E}_X\mathbb{E}_{X*}k(X, X^*) - \mathbb{E}_{X'}k(x, X') + \mathbb{E}_X\mathbb{E}_{X'}k(X, X') \\ &= 0\end{aligned}$$

We denote by $\circ$ the Hadamard product, such that for any two matrices K and L of equal dimension, $(K \circ L)_{ij} = K_{ij}L_{ij}$.

Finally, throughout this thesis we will use at most 3 random variables and their associated feature maps and gram matrices. The random variables will always be referred to as X, Y and Z. The associated kernels are denoted k, l and m respectively. The associated gram matrices are denoted K, L and M respectively. The associated feature maps are denoted ϕ, ψ and ω respectively.

3.2 HSIC

Recall that given samples $X_i, Y_i, i = 1, \ldots, n$, kernels k on $\mathcal{X}$ and l on $\mathcal{Y}$ and associated gram matrices K and L, the biased statistic for HSIC is defined by

$$\begin{aligned}HSIC_b &= \frac{1}{n^2}\sum_{ij} K_{ij}L_{ij} - \frac{2}{n^3}\sum_{ijr} K_{ij}L_{jr} + \frac{1}{n^4}\sum_{ijrs} K_{ij}L_{rs} \\ &= \frac{1}{n^2}(K \circ L)_{++} - \frac{2}{n^3}(KL)_{++} + \frac{1}{n^4}K_{++}L_{++}\end{aligned}$$

In this section it will be shown that under the null hypothesis that X and Y are independent, the normalised statistic for HSIC is asymptotically a degenerate V-statistic. This degenerate V-statistic can then have the Wild Bootstrap applied to

it to resample under the null hypothesis. We will first show that

$$nHSIC_b = \frac{1}{n}(\bar{K} \circ \bar{L})_{++} - \frac{2}{n^2}(\bar{K}\bar{L})_{++} + \frac{1}{n^3}\bar{K}_{++}\bar{L}_{++} \qquad (\dagger)$$

then show that the latter two terms in the above equation tend to 0 as $n \longrightarrow \infty$. Finally we will show that $\frac{1}{n}(\bar{K} \circ \bar{L})_{++}$ is a normalised degenerate V-statistic.

Claim 3.1. *Let α and β be fixed elements of the Hilbert Spaces in which $\phi(X)$ and $\psi(Y)$ take value respectively. Define $\phi'(X) = \phi(X) - \alpha$ and $\psi'(Y) = \psi(Y) - \beta$.*

$$\frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++} = \frac{1}{n^2}(K' \circ L')_{++} - \frac{2}{n^3}(K'L')_{++} + \frac{1}{n^4}K'_{++}L'_{++}$$

where $K'_{ij} = \langle \phi'(X_i), \phi'(X_j) \rangle$ and $L'_{ij} = \langle \psi'(Y_i), \psi'(Y_j) \rangle$.
In particular, taking $\alpha = \mu_X$ and $\beta = \mu_Y$, observe that

$$\frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++} = \frac{1}{n^2}(\bar{K} \circ \bar{L})_{++} - \frac{2}{n^3}(\bar{K}\bar{L})_{++} + \frac{1}{n^4}\bar{K}_{++}\bar{L}_{++}$$

And taking $\alpha = 0$ and $\beta = 0$,

$$\frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++} = HSIC_b$$

and hence $(\dagger)$ is true.

Proof: Given feature maps ϕ and ψ and observations $\mathcal{D} = \{X_i, Y_i, i = 1, \ldots, n\}$ define

$$T(\phi, \psi, \mathcal{D}) = \frac{1}{n^2}\sum_{ij}\Big\langle \phi(X_i) - \frac{1}{n}\sum_k \phi(X_k), \phi(X_j) - \frac{1}{n}\sum_k \phi(X_k)\Big\rangle \times \Big\langle \psi(Y_i) - \frac{1}{n}\sum_k \psi(Y_k), \psi(Y_j) - \frac{1}{n}\sum_k \psi(Y_k)\Big\rangle$$

Note that by definition of $\tilde{K}$ and $\tilde{L}$,

$$T(\phi, \psi, \mathcal{D}) = \frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++}$$

By Claim 6.1 in the Appendix,

$$T(\phi, \psi, \mathcal{D}) = \frac{1}{n^2}(K \circ L)_{++} - \frac{2}{n^3}(KL)_{++} + \frac{1}{n^2}K_{++}L_{++} \qquad (*)$$

Next, observe that

$$\begin{aligned}\phi'(X_i) - \frac{1}{n}\sum_k \phi'(X_k) &= \phi(X_i) - \alpha - \frac{1}{n}\sum_k \big\{\phi(X_k) - \alpha\big\} \\ &= \phi(X_i) - \frac{1}{n}\sum_k \phi(X_k)\end{aligned}$$

A similar result holds for ψ', hence

$$T(\phi', \psi', \mathcal{D}) = \frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++}$$

But by $(*)$, we also have that

$$T(\phi', \psi', \mathcal{D}) = \frac{1}{n^2}(K' \circ L')_{++} - \frac{2}{n^3}(K'L')_{++} + \frac{1}{n^2}K'_{++}L'_{++}$$

Putting the two equalities together, we see that the claim is true.

■

We can now write

$$nHSIC_b = \frac{1}{n}(\tilde{K} \circ \tilde{L})_{++} = \frac{1}{n}(\bar{K} \circ \bar{L})_{++} - \frac{2}{n^2}(\bar{K}\bar{L})_{++} + \frac{1}{n^3}\bar{K}_{++}\bar{L}_{++}$$

It remains to show the following three things.

Claim 3.2. *Under the null hypothesis that* $\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$,

(i) $\quad \frac{1}{n^3}\bar{K}_{++}\bar{L}_{++} \longrightarrow 0$

(ii) $\frac{1}{n^2}(\bar{K}\bar{L})_{++} \longrightarrow 0$

(iii) $\frac{1}{n}(\bar{K} \circ \bar{L})_{++}$ *is a degenerate normalised V-statistic*

Having proved this, it is then clear that $nHSIC_b$ is asymptotically equal to a degenerate normalised V-statistic.

To prove these claims, let us first set up some more notation.

First, recall that we denote the *population centered* feature maps $\bar{\phi} = \phi - \mu_X$ and $\bar{\psi} = \psi - \mu_Y$. Note that the gram matrices $\bar{K}$ and $\bar{L}$ represent inner products of the data with respect to $\bar{\phi}$ and $\bar{\psi}$. In a similar fashion, write $\bar{\mu}_X = \tilde{\mu}_X - \mu_X$ and $\bar{\mu}_Y = \tilde{\mu}_Y - \mu_Y$

Next, define the *population centred* empirical covariance function

$$\bar{C}_{XY} = \frac{1}{n}\sum_i \bar{\phi}(X_i) \otimes \bar{\psi}(Y_i)$$

This is an empirical estimator for the true population centred covariance function

$$C_{XY} = \mathbb{E}_{XY}[\bar{\phi}(X) \otimes \bar{\psi}(Y)] = \mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}_X} \otimes \mu_{\mathbb{P}_Y}$$

Lemma 3.1. *Assume that* $(X_i, Y_i)_i$ *is* β*-mixing with coefficients* $\beta_{XY}(m)$ *satisfying* $\sum_{m=1}^{\infty}(\beta_{XY}(m))^{\frac{\delta}{2+\delta}}$ *for some* $\delta > 0$.

Then

$$\|\bar{C}_{XY} - C_{XY}\| = O(\frac{1}{\sqrt{n}})$$
$$\|\tilde{\mu}_X - \mu_X\| = O(\frac{1}{\sqrt{n}})$$
$$\|\tilde{\mu}_Y - \mu_Y\| = O(\frac{1}{\sqrt{n}})$$

Proof: We exploit Theorem 1.1 from [25]. Using the language of this paper, $\bar{\phi}(X_i) \otimes \bar{\psi}(Y_i)$ is a 1-approximating functional of $(X_i, Y_i)_i$ (this follows straightforwardly from the

definition of 1-approximating functionals given since $\bar{\phi}(X_i) \otimes \bar{\psi}(Y_i) = (\bar{\phi} \otimes \bar{\psi})(X_i, Y_i)$ is a function of (X_i, Y_i)).

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

$$\mathbb{E}\|\bar{\phi}(X_1) \otimes \bar{\psi}(Y_1)\|^{2+\delta} < C^{2+\delta} < \infty \ \ \forall \delta > 0$$

Thus condition (1) is satisfied.

We can take $f_m = \bar{\phi}(X_0) \otimes \bar{\psi}(Y_0) \ \ \forall m$ and so acheive $a_m = 0 \ \ \forall m$, thus condition (2) is satisfied.

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

Thus, by Theorem 1.1 in [25]

$$\sqrt{n}(\bar{C}_{XY} - C_{XY}) \overset{n}{\rightsquigarrow} N$$

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

$$\|\bar{C}_{XY} - C_{XY}\| = O(\frac{1}{\sqrt{n}})$$

Note that a similar arguments hold to prove that conditions (1) and (2) are satisfied by the sub-processes $(X_i)_i$ and $(Y_i)_i$. By Lemma 2.1, satisfaction of condition (3) by $(X_i, Y_i)_i$ implies that any subprocess also satisfies condition (3). It thus follows that $\|\tilde{\mu}_X - \mu_X\| = O(\frac{1}{\sqrt{n}})$ and $\|\tilde{\mu}_Y - \mu_Y\| = O(\frac{1}{\sqrt{n}})$ ■

Under the null hypothesis, $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$ and thus

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

And therefore, noting also that $\tilde{\mu}_X - \mu_X = \bar{\mu}_X$ and $\tilde{\mu}_Y - \mu_Y = \bar{\mu}_Y$ we have that

$$\begin{aligned}\|\bar{C}_{XY}\| &= O(\frac{1}{\sqrt{n}}) \\ \|\bar{\mu}_X\| &= O(\frac{1}{\sqrt{n}}) \\ \|\bar{\mu}_Y\| &= O(\frac{1}{\sqrt{n}})\end{aligned}$$

Now we can prove the claims.

Proof: (i)

$$\begin{aligned}\frac{1}{n^3}\bar{K}_{++}\bar{L}_{++} &= \frac{1}{n^3}\sum_{i,j}\Big\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\Big\rangle\sum_{k,l}\Big\langle\bar{\psi}(Y_k),\bar{\psi}(Y_l)\Big\rangle \\ &= n\Big\langle\frac{1}{n}\sum_i\bar{\phi}(X_i),\frac{1}{n}\sum_j\bar{\phi}(X_j)\Big\rangle\Big\langle\frac{1}{n}\sum_k\bar{\psi}(Y_k),\frac{1}{n}\sum_l\bar{\psi}(Y_l)\Big\rangle \\ &= n\Big\langle\bar{\mu}_X,\bar{\mu}_X\Big\rangle\Big\langle\bar{\mu}_Y,\bar{\mu}_Y\Big\rangle \\ &= n\|\bar{\mu}_X\|^2\|\bar{\mu}_Y\|^2 \\ &= nO(\frac{1}{n})O(\frac{1}{n}) \\ &= O(\frac{1}{n}) \longrightarrow 0\end{aligned}$$

■

Proof: (ii)

$$\begin{aligned}
\frac{1}{n^2}(\bar{K}\bar{L})_{++} &= \frac{1}{n^2}\sum_{ijk}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_j),\bar{\psi}(Y_k)\rangle \\
&= \frac{1}{n^2}\sum_{ijk}\langle\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j),\bar{\phi}(X_i)\otimes\bar{\psi}(Y_k)\rangle \\
&= n\langle\frac{1}{n}\sum_{j}[\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)],[\frac{1}{n}\sum_{i}\bar{\phi}(X_i)]\otimes[\frac{1}{n}\sum_{k}\bar{\psi}(Y_k)]\rangle \\
&= n\langle\bar{C}_{XY},\bar{\mu}_X\otimes\bar{\mu}_Y\rangle \\
&= O(\frac{1}{\sqrt{n}})
\end{aligned}$$

■

Proof: (iii)

$$\frac{1}{n}(\bar{K}\circ\bar{L})_{++} = \frac{1}{n}\sum_{ij}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_i),\bar{\psi}(Y_j)\rangle$$

Letting $S_i = (X_i, Y_i)$, observe that this is a normalised V-statistic with core

$$h(S_i,S_j) = \langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_i),\bar{\psi}(Y_j)\rangle$$

Under the null hypothesis the joint factorises as $P_{XY} = P_X P_Y$. The expectation operator $\mathbb{E}_{XY}$ therefore factorises as

$$\mathbb{E}_{XY} = \mathbb{E}_X\mathbb{E}_Y$$

Thus

$$\begin{aligned}
\mathbb{E}_{S_j}h(s_i,S_j) &= \mathbb{E}_{X_jY_j}\langle\bar{\phi}(x_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(y_i),\bar{\psi}(Y_j)\rangle \\
&= \langle\bar{\phi}(x_i),\mathbb{E}_{X_j}\bar{\phi}(X_j)\rangle\langle\bar{\psi}(y_i),\mathbb{E}_{Y_j}\bar{\psi}(Y_j)\rangle \\
&= \langle\bar{\phi}(x_i),0\rangle\langle\bar{\psi}(y_i),0\rangle = 0
\end{aligned}$$

■

To conclude, we have proven that the normalised test statistic for HSIC asymptotically approaches a degenerate normalised V-statistic, and therefore satisfies the conditions required to use the wild bootstrap.

3.3 Lancaster

The notation used in this section will be the same as for HSIC, except that now we have three variables rather than two. The new variable will be denoted Z, its kernel function m, its associated feature map ω and its Gram matrix M. Use of over-head bars and tildes represent centering with respect to expectations and sample averages, as with HSIC.

As shown in [4], the Lancaster test statistic can be expressed as:

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

Our aim, as before with HSIC, is to show that, after normalisation, this is asymptotically equal to a normalised V-statistic with a degenerate core. We will follow a similar approach - first expand it to express it as a sum of terms, then show that all but one of the terms goes to 0 and that the remaining term is a normalised V-statistic with a degenerate core.

Claim 3.3. *Let α, β and γ be fixed elements of the Hilbert Spaces in which $\phi(X)$, $\psi(Y)$ and $\omega(Z)$ take value respectively. Define $\phi'(X) = \phi(X) - \alpha$, $\psi'(Y) = \psi(Y) - \beta$ and $\omega'(Z) = \omega(Z) - \gamma$. Then*

$$\begin{aligned}\|\Delta_L\hat{P}\|^2 = \frac{1}{n^2}(K'\circ L'\circ M')_{++} &- \frac{2}{n^3}((K'\circ L')M')_{++} &&- \frac{2}{n^3}((K'\circ M')L')_{++}\\ - \frac{2}{n^3}((M'\circ L')K')_{++} &+ \frac{1}{n^4}(K'\circ L')_{++}M'_{++} &&+ \frac{1}{n^4}(K'\circ M')_{++}L'_{++}\\ + \frac{1}{n^4}(L'\circ M')_{++}K'_{++} &+ \frac{2}{n^4}(M'K'L')_{++} &&+ \frac{2}{n^4}(K'L'M')_{++}\\ + \frac{2}{n^4}(K'M'L')_{++} &+ \frac{4}{n^4}tr(K'_+\circ L'_+\circ M'_+) &&- \frac{4}{n^5}(K'L')_{++}M'_{++}\\ - \frac{4}{n^5}(K'M')_{++}L'_{++} &- \frac{4}{n^5}(L'M')_{++}K'_{++} &&+ \frac{4}{n^6}K'_{++}L'_{++}M'_{++}\end{aligned}$$

where $K'_{ij} = \langle\phi'(X_i), \phi'(X_j)\rangle$, $L'_{ij} = \langle\psi'(Y_i), \psi'(Y_j)\rangle$ *and* $M'_{ij} = \langle\omega'(Z_i), \omega'(Z_j)\rangle$.

In particular, taking $\alpha = \mu_X$, $\beta = \mu_Y$ *and* $\gamma = \mu_Z$, *we can replace* K', L' *and* M' *by* $\bar{K}, \bar{L}$ *and* $\bar{M}$ *respectively in the above equation*

Proof: The proof of this claim mirrors that of Claim 3.1.

Given feature maps ϕ, ψ and ω, and observations $\mathcal{D} = \{X_i, Y_i, Z_i, i = 1, \ldots, n\}$ define

$$\begin{aligned}T(\phi,\psi,\omega,\mathcal{D}) = \frac{1}{n^2}\sum_{i,j}&\Big\langle\phi(X_i) - \frac{1}{n}\sum_k\phi(X_k), \phi(X_j) - \frac{1}{n}\sum_k\phi(X_k)\Big\rangle\\ &\times\Big\langle\psi(Y_i) - \frac{1}{n}\sum_k\psi(Y_k), \psi(Y_j) - \frac{1}{n}\sum_k\psi(Y_k)\Big\rangle\\ &\times\Big\langle\omega(Z_i) - \frac{1}{n}\sum_k\omega(Z_k), \omega(Z_j) - \frac{1}{n}\sum_k\omega(Z_k)\Big\rangle\end{aligned}$$

By definition of $\tilde{K}$, $\tilde{L}$ and $\tilde{M}$, observe that

$$T(\phi,\psi,\omega,\mathcal{D}) = \frac{1}{n^2}\big(\tilde{K}\circ\tilde{L}\circ\tilde{M}\big)_{++} = \|\Delta_L\hat{P}\|^2$$

By expanding the inner products in the definition of T, it is straightforward (but

very tedious) algebra to show that

$$\begin{aligned}
T(\phi,\psi,\omega,\mathcal{D}) = \frac{1}{n^2}(K\circ L\circ M)_{++} &- \frac{2}{n^3}((K\circ L)M)_{++} &&- \frac{2}{n^3}((K\circ M)L)_{++}\\
- \frac{2}{n^3}((M\circ L)K)_{++} &+ \frac{1}{n^4}(K\circ L)_{++}M_{++} &&+ \frac{1}{n^4}(K\circ M)_{++}L_{++}\\
+ \frac{1}{n^4}(L\circ M)_{++}K_{++} &+ \frac{2}{n^4}(MKL)_{++} &&+ \frac{2}{n^4}(KLM)_{++}\\
+ \frac{2}{n^4}(KML)_{++} &+ \frac{4}{n^4}tr(K_+\circ L_+\circ M_+) &&- \frac{4}{n^5}(KL)_{++}M_{++}\\
- \frac{4}{n^5}(KM)_{++}L_{++} &- \frac{4}{n^5}(LM)_{++}K_{++} &&+ \frac{4}{n^6}K_{++}L_{++}M_{++}
\end{aligned}$$

(for a proof see Claim 6.2 in the appendix.)

Next, observe that

$$\begin{aligned}
\phi'(X_i) - \frac{1}{n}\sum_k \phi'(X_k) &= \phi(X_i) - \alpha - \frac{1}{n}\sum_k \big\{\phi(X_k) - \alpha\big\}\\
&= \phi(X_i) - \frac{1}{n}\sum_k \phi(X_k)
\end{aligned}$$

Similar results hold for ϕ' and ω', hence

$$T(\phi',\psi',\omega',\mathcal{D}) = \frac{1}{n^2}\big(\tilde{K}\circ\tilde{L}\circ\tilde{M}\big)_{++} = \|\Delta_L\hat{P}\|^2$$

But we also have that

$$
\begin{aligned}
T(\phi', \psi', \omega', \mathcal{D}) = \frac{1}{n^2}(K' \circ L' \circ M')_{++} &- \frac{2}{n^3}((K' \circ L')M')_{++} &- \frac{2}{n^3}((K' \circ M')L')_{++} \\
- \frac{2}{n^3}((M' \circ L')K')_{++} &+ \frac{1}{n^4}(K' \circ L')_{++}M'_{++} &+ \frac{1}{n^4}(K' \circ M')_{++}L'_{++} \\
+ \frac{1}{n^4}(L' \circ M')_{++}K'_{++} &+ \frac{2}{n^4}(M'K'L')_{++} &+ \frac{2}{n^4}(K'L'M')_{++} \\
+ \frac{2}{n^4}(K'M'L')_{++} &+ \frac{4}{n^4}tr(K'_+ \circ L'_+ \circ M'_+) &- \frac{4}{n^5}(K'L')_{++}M'_{++} \\
- \frac{4}{n^5}(K'M')_{++}L'_{++} &- \frac{4}{n^5}(L'M')_{++}K'_{++} &+ \frac{4}{n^6}K'_{++}L'_{++}M'_{++}
\end{aligned}
$$

And hence the result follows. ■

Observe that up to symmetries $K \leftrightarrow L \leftrightarrow M$, there are 7 different terms here.

Before we go further, we need to introduce yet more notation. Similar to $\bar{C}_{XY}$ and C_{XY} in the HSIC section, we define

$$\bar{C}_{XYZ} = \frac{1}{n}\sum_i \bar{\phi}(X_i) \otimes \bar{\psi}(Y_i) \otimes \bar{\omega}(Z_i)$$

$$C_{XYZ} = \mathbb{E}_{XYZ}[\bar{\phi}(X) \otimes \bar{\psi}(Y) \otimes \bar{\omega}(Z)]$$

By permuting the ordering of $\bar{\phi}(X)$, $\bar{\psi}(Y)$ and $\bar{\omega}(Z)$ we can similarly define C_{ZXY}, C_{XZY} etc.

It is possible to express each of the 7 symmetrically-distinct terms as inner products of covariance operators and mean embeddings:

Claim 3.4. *Let u, v and w be random variables with population centred Gram ma-*

trices $\bar{U}$, $\bar{V}$ *and* $\bar{W}$ *respectively. Then*

$$(i) \qquad \frac{1}{n}(\bar{U} \circ \bar{V} \circ \bar{W})_{++} = n\langle \bar{C}_{uvw}, \bar{C}_{uvw} \rangle$$

$$(ii) \qquad \frac{1}{n^2}((\bar{U} \circ \bar{V})\bar{W})_{++} = n\langle \bar{C}_{uvw}, \bar{C}_{uv} \otimes \bar{\mu}_w \rangle$$

$$(iii) \qquad \frac{1}{n^3}(\bar{U} \circ \bar{V})_{++}\bar{W}_{++} = n\langle \bar{C}_{uv} \otimes \bar{\mu}_w, \bar{C}_{uv} \otimes \bar{\mu}_w \rangle$$

$$(iv) \qquad \frac{1}{n^3}(\bar{U}\bar{V}\bar{W})_{++} = n\langle \bar{C}_{uv} \otimes \bar{\mu}_w, \bar{\mu}_u \otimes \bar{C}_{vw} \rangle$$

$$(v) \qquad \frac{1}{n^3} tr(\bar{U}_+ \circ \bar{V}_+ \circ \bar{W}_+) = n\langle \bar{C}_{uvw}, \bar{\mu}_u \otimes \bar{\mu}_v \otimes \bar{\mu}_w \rangle$$

$$(vi) \qquad \frac{1}{n^3}(\bar{U}\bar{V})_{++}\bar{W}_{++} = n\langle \bar{C}_{uv} \otimes \bar{\mu}_w, \bar{\mu}_u \otimes \bar{\mu}_v \otimes \bar{\mu}_w \rangle$$

$$(vii) \qquad \frac{1}{n^3}\bar{U}_{++}\bar{V}_{++}\bar{W}_{++} = n\langle \bar{\mu}_u \otimes \bar{\mu}_v \otimes \bar{\mu}_w, \bar{\mu}_u \otimes \bar{\mu}_v \otimes \bar{\mu}_w \rangle$$

For proof of Claim 3.4, see page 76 in the appendix.

It follows that we can write the normalised Lancaster statistic as

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

Let us take a moment to remember our aim.

Remark. *We need to show that under the null hypothesis, the normalised Lancaster test statistic asymptotically approaches a normalised degenerate V-statistic. Recall that the Lancaster test has a composite null hypothesis:* $X \perp\!\!\!\perp (Y,Z) \vee Y \perp\!\!\!\perp (X,Z) \vee Z \perp\!\!\!\perp (X,Y)$. *We therefore need to demonstrate the required result under each component of the null hypothesis separately. Since the Lancaster statistic is symmetric in* $X \leftrightarrow Y \leftrightarrow Z$, *we can without loss of generality consider only the case* $Z \perp\!\!\!\perp (X,Y)$*; equivalently* $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$. *Observe further that if any two components are satisfied then there is total independence, ie* $\mathbb{P}_{XYZ} = \mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z$

Theorem 3.1. *Suppose that* $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$. *Then*

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

and this is a normalised degenerate V-statistic.

Proof: $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$ implies that the expectation operator factorises similarly as $\mathbb{E}_{XYZ} = \mathbb{E}_{XY}\mathbb{E}_Z$. Hence

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

Similarly, $C_{XZY} = C_{YZX} = 0$

By marginalising with respect to X or Y, we obtain $\mathbb{P}_{YZ} = \mathbb{P}_Y\mathbb{P}_Z$ and $\mathbb{P}_{XZ} = \mathbb{P}_X\mathbb{P}_Z$ respectively. The expectation operators again factorise similarly, and therefore

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

and

$$
\begin{aligned}
C_{YZ} &= \mathbb{E}_{YZ}[\bar{\psi}(Y) \otimes \bar{\omega}(Z)] \\
&= \mathbb{E}_Y \mathbb{E}_Z[\bar{\psi}(Y) \otimes \bar{\omega}(Z)] \\
&= [\mathbb{E}_Y \bar{\psi}(Y)] \otimes [\mathbb{E}_Z \bar{\omega}(Z)] \\
&= [\mathbb{E}_Y \bar{\psi}(Y)] \otimes 0 \\
&= 0
\end{aligned}
$$

Lemma 3.2. *Assume that $(X_i, Y_i, Z_i)_i$ is β-mixing with coefficients $\beta_{XYZ}(m)$ satisfying $\sum_{m=1}^{\infty} (\beta_{XYZ}(m))^{\frac{\delta}{2+\delta}}$ for some $\delta > 0$. Then*

$$
\begin{aligned}
\|\bar{C}_{XYZ}\| &= O(\frac{1}{\sqrt{n}}) \\
\|\bar{C}_{XZY}\| &= O(\frac{1}{\sqrt{n}}) \\
\|\bar{C}_{YZX}\| &= O(\frac{1}{\sqrt{n}}) \\
\|\bar{C}_{XZ}\| &= O(\frac{1}{\sqrt{n}}) \\
\|\bar{C}_{YZ}\| &= O(\frac{1}{\sqrt{n}}) \\
\|\bar{\mu}_X\| &= O(\frac{1}{\sqrt{n}}) \\
\|\bar{\mu}_Y\| &= O(\frac{1}{\sqrt{n}}) \\
\|\bar{\mu}_Z\| &= O(\frac{1}{\sqrt{n}})
\end{aligned}
$$

Proof: The proof of this is exactly the same as the proof of Lemma 3.1, replacing (X_i, Y_i) and $\bar{\phi}(X_i) \otimes \bar{\psi}(Y_i)$ with (X_i, Y_i, Z_i) and $\bar{\phi}(X_i) \otimes \bar{\psi}(Y_i) \otimes \bar{\omega}(Z_i)$ respectively.

■

Observe that $\|C_{XY}\| \neq 0$, and so $\|\bar{C}_{XY}\| \longrightarrow const$

We will now show that almost all of the terms in the expression for $n\|\Delta_L \hat{P}\|^2$ tend to 0 in the limit $n \longrightarrow \infty$.

Claim 3.5.

$$n\|\Delta_L \hat{P}\|^2 \longrightarrow n\langle \bar{C}_{XYZ}, \bar{C}_{XYZ}\rangle - 2n\langle \bar{C}_{XYZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle + n\langle \bar{C}_{XY} \otimes \bar{\mu}_Z, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle$$

Proof: The proof of this is very simple, but requires familiarity with how to manipulate inner products of tensor products. We demonstrate here only that the term $n\langle \bar{\mu}_X \otimes \bar{C}_{YZ}, \bar{C}_{XY} \otimes \bar{\mu}_Z\rangle$ goes to zero; the proofs for the other terms are essentially the same.

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

All of the other terms are bounded in the same way. We first use Cauchy-Schwartz, then the property that $\|a \otimes b\| = \|a\|\|b\|$, then we use the asymptotic bounds on the previous page.

All of the other terms (except those in the statement of the claim) go to 0 as $n \longrightarrow \infty$. ■

Remark. *The remaining terms would be found to be* $O(1)$ *if subjected to the analysis above.*

Rewriting in terms of Gram matrices, we have thus shown that

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

We will use this to prove the final result

Claim 3.6.

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

and moreover this is a normalised degenerate V-statistic

Proof:

Treating (X, Y) to be a single variable with a kernel given by the feature map $\bar{\phi} \otimes \bar{\psi}$, observe that

$$\frac{1}{n}((\bar{K} \circ \bar{L}) \circ \bar{M})_{++} - \frac{2}{n^2}((\bar{K} \circ \bar{L})\bar{M})_{++} + \frac{1}{n^3}(\bar{K} \circ \bar{L})_{++}\bar{M}_{++}$$

can be thought of as HSIC between (X, Y) and Z. Observe that with this kernel, the expectation of (X, Y) in feature space is

$$\mathbb{E}_{XY}[\bar{\phi}(X) \otimes \bar{\psi}(Y)] = C_{XY}$$

We can therefore recentre the matrix $(\bar{K} \circ \bar{L})_{ij} = \langle \bar{\phi}(X_i), \bar{\phi}(X_j)\rangle\langle \bar{\psi}(Y_i), \bar{\psi}(Y_j)\rangle = \langle \bar{\phi}(X_i) \otimes \bar{\psi}(Y_i), \bar{\phi}(X_j) \otimes \bar{\psi}(Y_j)\rangle$ with respect to the expectation of its entries to get

$$\overline{(\bar{K} \circ \bar{L})}_{ij} = \langle \bar{\phi}(X_i) \otimes \bar{\psi}(Y_i) - C_{XY}, \bar{\phi}(X_j) \otimes \bar{\psi}(Y_j) - C_{XY}\rangle$$

Recall that in Claim 3.1, we proved that HSIC is invariant to recentering the Gram matrices. Using this and Claim 3.2, we obtain

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

It remains to show that $\frac{1}{n}(\overline{(\bar{K}\circ\bar{L})}\circ\bar{M})_{++}$ is a normalised degenerate V-statistic. Let $S_i = (X_i, Y_i, Z_i)$.

$$\frac{1}{n}(\overline{(\bar{K}\circ\bar{L})}\circ\bar{M})_{++} = \frac{1}{n}\sum_{ij}\langle\bar{\phi}(X_i)\otimes\bar{\psi}(Y_i) - C_{XY}, \bar{\phi}(X_j)\otimes\bar{\psi}(Y_j) - C_{XY}\rangle\langle\bar{\omega}(Z_i), \bar{\omega}(Z_j)\rangle$$

This is a normalised V-statistic with core

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

To show that the core is degenerate, we fix any value s_i and take expectations with respect to S_j:

$$\begin{aligned}
\mathbb{E}_{S_j}h(s_i, S_j) &= \mathbb{E}_{X_jY_jZ_j}\langle\bar{\phi}(x_i)\otimes\bar{\psi}(x_i) - C_{XY}, \bar{\phi}(X_j)\otimes\bar{\psi}(Y_j) - C_{XY}\rangle\langle\bar{\omega}(z_i), \bar{\omega}(Z_j)\rangle \\
&= \mathbb{E}_{X_jY_j}\mathbb{E}_{Z_j}\langle\bar{\phi}(x_i)\otimes\bar{\psi}(x_i) - C_{XY}, \bar{\phi}(X_j)\otimes\bar{\psi}(Y_j) - C_{XY}\rangle\langle\bar{\omega}(z_i), \bar{\omega}(Z_j)\rangle \\
&= \langle\bar{\phi}(x_i)\otimes\bar{\psi}(x_i) - C_{XY}, \mathbb{E}_{X_jY_j}[\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)] - C_{XY}\rangle\langle\bar{\omega}(z_i), \mathbb{E}_{Z_j}\bar{\omega}(Z_j)\rangle \\
&= \langle\bar{\phi}(x_i)\otimes\bar{\psi}(x_i) - C_{XY}, 0\rangle\langle\bar{\omega}(z_i), 0\rangle \\
&= 0
\end{aligned}$$

■

To conclude, we have shown that if $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$, then

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

and moreover this is a normalised degenerate V-statistic.

Now observe that if, further, it holds that $\mathbb{P}_{XYZ} = \mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z$, all of the above analysis still holds but additionally we have that $C_{XY} = 0$ and hence $(\bar{K} \circ \bar{L})$ is already centred in expectation. Thus

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

and this is a normalised degenerate V-statistic.

■

4 Experiments

This section is laid out as follows. We first describe two HSIC-based tests for detecting joint dependence of three random variables - similar to the Lancaster test, rejection of the null in these tests tells us that the joint distribution $\mathbb{P}_{XYZ}$ does not factorise. If we fail to reject the null, they are non-informative.

We then discuss briefly multiple testing corrections. We describe an improvement to the Holm-Bonferroni correction [26] proposed in [4] for the Lancaster test.

We next compare the performance of the Lancaster test to the two HSIC-based tests described above on three artificial datasets. The first two datasets are ones for which the joint distribution does not factorise. In the third, the distribution does factorise - we use this to see how the specificities of the tests compare. We then show that the Wild Bootstrap is indeed necessary when in the non-*iid* case, by comparing its performance in controlling Type I errors to that of the permutation bootstrap in a fourth artificial dataset.

Finally, we perform each of the tests on real forex data.

4.1 Using HSIC for three-way independence testing

We introduce briefly two different ways of using HSIC to test for dependence between three variables X, Y and Z. We will compare them with Lancaster in the remainder of this section.

4.1.1 The 'Pairwise HSIC' test

Observe that the following statements hold by marginalising out X, Y or Z from the left hand equations as appropriate.

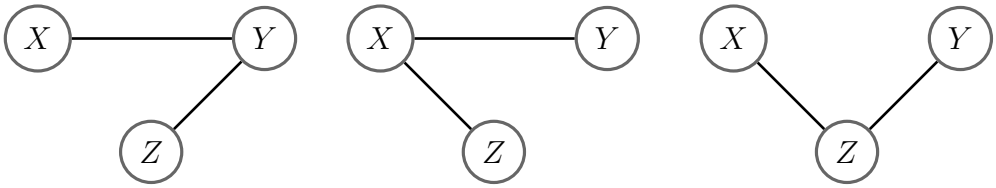

Figure 2: If any two edges are present in a graphical model on three nodes then the graph is connected, and thus the joint distribution $\mathbb{P}_{XYZ}$ does not factorise.

$$\begin{aligned}
\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z &\implies [\mathbb{P}_{XZ} = \mathbb{P}_X\mathbb{P}_Z] \wedge [\mathbb{P}_{YZ} = \mathbb{P}_Y\mathbb{P}_Z] \\
\mathbb{P}_{XYZ} = \mathbb{P}_{XZ}\mathbb{P}_Y &\implies [\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y] \wedge [\mathbb{P}_{YZ} = \mathbb{P}_Y\mathbb{P}_Z] \\
\mathbb{P}_{XYZ} = \mathbb{P}_X\mathbb{P}_{YZ} &\implies [\mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y] \wedge [\mathbb{P}_{XZ} = \mathbb{P}_X\mathbb{P}_Z]
\end{aligned}$$

The contrapositives of these statements are, in order

$$\begin{aligned}
[\mathbb{P}_{XZ} \neq \mathbb{P}_X\mathbb{P}_Z] \vee [\mathbb{P}_{YZ} \neq \mathbb{P}_Y\mathbb{P}_Z] &\implies \mathbb{P}_{XYZ} \neq \mathbb{P}_{XY}\mathbb{P}_Z \\
[\mathbb{P}_{XY} \neq \mathbb{P}_X\mathbb{P}_Y] \vee [\mathbb{P}_{YZ} \neq \mathbb{P}_Y\mathbb{P}_Z] &\implies \mathbb{P}_{XYZ} \neq \mathbb{P}_{XZ}\mathbb{P}_Y \\
[\mathbb{P}_{XY} \neq \mathbb{P}_X\mathbb{P}_Y] \vee [\mathbb{P}_{XZ} \neq \mathbb{P}_X\mathbb{P}_Z] &\implies \mathbb{P}_{XYZ} \neq \mathbb{P}_X\mathbb{P}_{YZ}
\end{aligned}$$

Thus, if the left hand sides of all of the above statements hold, we can conclude that $\mathbb{P}_{XYZ}$ does not factorise. All three statements will hold provided that any two pairs of variables are dependent. This can be graphically interpreted by Figure 2, which shows that if any two of the three possible edges in a graph on three nodes are present, the graph is connected (and thus the joint probability distribution on the three nodes does not factorise).

This gives us a statistical test for testing joint dependence. We test each of the three hypotheses:

$$\begin{aligned}\mathcal{H}_{XY}: &\quad \mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y \\ \mathcal{H}_{XZ}: &\quad \mathbb{P}_{XZ} = \mathbb{P}_X\mathbb{P}_Z \\ \mathcal{H}_{YZ}: &\quad \mathbb{P}_{YZ} = \mathbb{P}_Y\mathbb{P}_Z\end{aligned}$$

using HSIC. If we reject any two of them, we conclude that the distribution does not factorise. In the remainder of this section, we will call this the Pairwise HSIC test.

4.1.2 The '3-way HSIC' test

Recall that in the Lancaster test, our null hypothesis $\mathcal{H}_0$ is a *composite* hypothesis. Let us denote by $\mathcal{H}_X$ the hypothesis that $X \perp\!\!\!\perp (Y, Z)$, and define similarly $\mathcal{H}_Y$ and $\mathcal{H}_Z$. Then

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

and so we reject $\mathcal{H}_0$ if and only if we reject each of $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$. Instead of using the Lancaster statistic to test each of these sub-hypotheses, we could use HSIC:

We can consider (Y, Z) to be a single random variable. We can thus test $\mathcal{H}_X$ (ie whether (Y, Z) is independent of X or not) using HSIC. We can similarly test $\mathcal{H}_Y$ and $\mathcal{H}_Z$. If we reject all three of the sub-hypotheses, we reject $\mathcal{H}_0$.

We will refer to this test as the 3-way HSIC test throughout the remainder of this section.

4.2 Multiple testing correction

When performing multiple hypothesis tests, we must take into consideration the fact that the probability of falsely rejecting one of the hypotheses grows as we increase

the number of tests. In both the Lancaster and HSIC-based tests, our null hypothesis consists of multiple sub-hypotheses. In order to control the overall Type-I error rate, we must consider carefully the threshold p-values we use for each of the sub-tests.

4.2.1 Holm-Bonferroni

Given a family of multiple hypotheses $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_m$, the *Holm-Bonferroni* method [26] is a way to control the *family-wise error rate*. This is the probability that one or more of the hypotheses are falsely rejected. To ensure a family-wise error rate of at most α, we perform the following procedure.

1. Let $p_1, \ldots, p_m$ be the p-values corresponding to each hypothesis.
2. Sort the p-values from lowest to highest. Write them as $p_{(1)}, \ldots, p_{(m)}$ and let $\mathcal{H}_{(1)}, \ldots, \mathcal{H}_{(m)}$ be the corresponding hypotheses.
3. Let k be the minimal index such that $p_{(k)} > \frac{\alpha}{m+1-k}$
4. Reject the hypotheses $\mathcal{H}_{(1)}, \ldots, \mathcal{H}_{(k-1)}$ and do not reject the hypotheses $\mathcal{H}_{(k)}, \ldots, \mathcal{H}_{(m)}$. If $k = 1$, we reject all of the hypotheses.

4.2.2 Multiple correction for Pairwise HSIC test

Observe that we are testing three hypotheses. If we reject any two of the hypotheses, we conclude that the joint distribution $\mathbb{P}_{XYZ}$ does not factorise. How can we bound the Type I error rate of our overall test? Let us consider the following example.

Example (See Figure 3)**.** *Suppose that $\mathbb{P}_{XYZ} = \mathbb{P}_{XY}\mathbb{P}_Z$, and that $\mathbb{P}_{XY}$ does not factorise. First note that we should reject $\mathcal{H}_{XY}$ but not reject the null hypothesis overall. Suppose however that we falsely reject one of $\mathcal{H}_{YZ}$ or $\mathcal{H}_{XZ}$. In this case the overall result would be to incorrectly conclude that $\mathbb{P}_{XYZ}$ must not factorise.*

Therefore, if our aim is to bound the probability of a Type I error overall by α, then we should apply the Holm-Bonferroni method to account for multiple testing error, since the family-wise error rate (ie the probability of incorrectly rejecting at

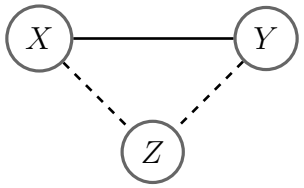

Figure 3: If we correctly detect the edge XY, but falsely detect edges XZ or YZ, then we would incorrectly conclude that the joint distribution does not factorise.

least one of the hypotheses) is what we need to bound, as illustrated by the above example.

4.2.3 Multiple correction for Lancaster and 3-way HSIC test

Recall that in the Lancaster and 3-way HSIC tests, our null hypothesis $\mathcal{H}_0$ is a *composite* hypothesis:

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

where $\mathcal{H}_X$ the hypothesis that $X \perp\!\!\!\perp (Y, Z)$, and $\mathcal{H}_Y$ and $\mathcal{H}_Z$ are defined similarly.

We perform statistical tests for each of $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$, and we reject $\mathcal{H}_0$ if and only if we reject each of its components.

We wish to control the Type I error of the composite test - that is, we wish to control the probability of falsely rejecting $\mathcal{H}_0$. Denote by A_* the event that we reject $\mathcal{H}_*$. Then

$$\mathbb{P}(A_0) = \mathbb{P}(A_X \wedge A_Y \wedge A_Z) \leqslant \min\{\mathbb{P}(A_X), \mathbb{P}(A_Y), \mathbb{P}(A_Z)\}$$

Whether or not there exists a better bound than this that holds in absolute generality is not clear, because the events A_X, A_Y and A_Z are not independent. If $\mathcal{H}_0$ is true, then at least one of $\mathcal{H}_X$, $\mathcal{H}_Y$ and $\mathcal{H}_Z$ must be true. Therefore, if we use a threshold p-value of α in each statistical test separately, then WLOG assuming that $\mathcal{H}_X$ is true, we see that

$$\mathbb{P}(A_0) \leqslant \min\{\mathbb{P}(A_X), \mathbb{P}(A_Y), \mathbb{P}(A_Z)\} \leqslant \mathbb{P}(A_X) = \alpha$$

To conclude: we can bound the Type I error rate by α by setting the Type I error rate for each constituent test individually to be α. We refer to this correction as the *'naive' correction* throughout the remainder of this thesis.

In [4], it is suggested that one use the Holm-Bonferroni method to control the Type I error rate. This results in a worse test power than the 'naive' correction, since it provides overly conservative bounds on the Type I error rate. Indeed, for a rejection of the null to occur using the Holm-Bonferroni correction, the sorted p-values of the hypotheses would have to be lower than $[\frac{\alpha}{3}, \frac{\alpha}{2}, \alpha]$, compared to only $[\alpha, \alpha, \alpha]$ using the method described above.

See Example 3 below for an empirical comparison of the Type I error rates of the two methods on artificial data for which the ground truth is that $\mathcal{H}_0$ is true.

4.3 Results

4.3.1 Example 1: Artificial data

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

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

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

Data were generated according to this definition with varying values for the dependence coefficient. For each value of the dependence coefficient, 500 datasets were generated, each consisting of 2000 consecutive observations of the variables. We

ran the Wild Bootstrap with 250 bootstrapping procedures and we used a Gaussian kernel with bandwidth parameter 1 on each of X, Y and Z.

In this example, the ground truth is that Z is dependent on both X and Y separately, as well as on them both jointly (ie dependent on (X, Y)). The results are presented in Figure 4. Observe that the HSIC-based test is able to detect the dependence more easily than the Lancaster test when the interaction is weak, and that when using the 'naive' correction, the Lancaster test has a higher test power than when using the Holm-Bonferroni correction.

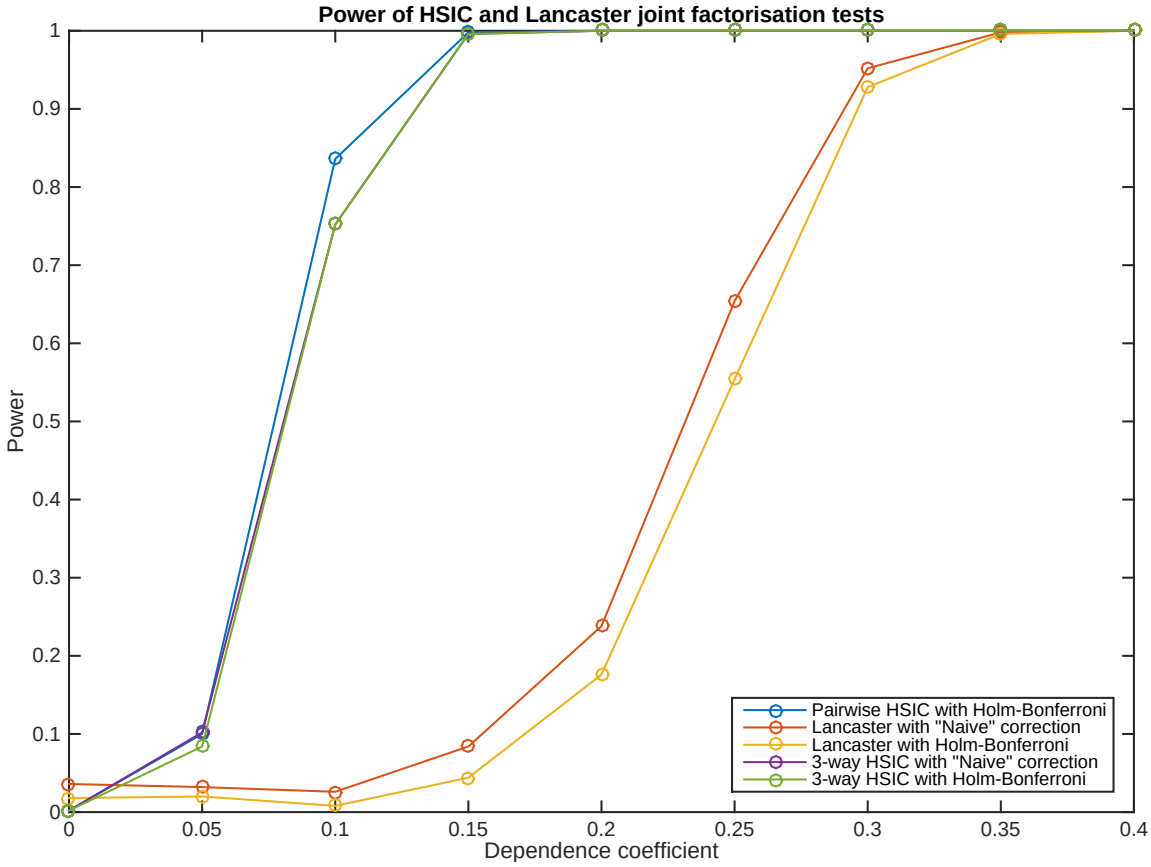

Figure 4: Performance of Lancaster and HSIC-based joint dependence tests on data from Example 1. Observe that the HSIC-based tests outperform the Lancaster tests, and that Lancaster with 'naive' correction performs better than with Holm-Bonferroni.

4.3.2 Example 2: Artificial data

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

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

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

Data were generated according to this definition with varying values for the dependence coefficient. For each value of the dependence coefficient, 500 datasets were generated, each consisting of 2000 consecutive observations of the variables. We ran the Wild Bootstrap with 250 bootstrapping procedures and we used a Gaussian kernel with bandwidth parameter 1 on each of X, Y and Z.

In contrast to the dataset in Example 1, Z is dependent on the process (X, Y) but is independent of X and Y when considered separately. Indeed, observe that the marginal distributions of X and Y are both normal distributions with mean 0, and thus $\text{sign}(X_tY_t)$ is either -1 or 1 with equal probability, conditioned upon neither or exactly one of X_t or Y_t.

The results are presented in Figure 5. Observe that Pairwise HSIC is unable to correctly identify that the distribution does not factorise. This is because it only looks at the variables pairwise, and so is unable to detect the three-way dependence that Lancaster can detect.

3-way HSIC should in principle be able to detect the dependence, but when the dependence coefficient is low the interactions between X and (Y, Z), as well as

between Y and (X, Z), may be very weak.

As before, Lancaster with the 'naive' correction outperforms Lancaster with Holm-Bonferroni.

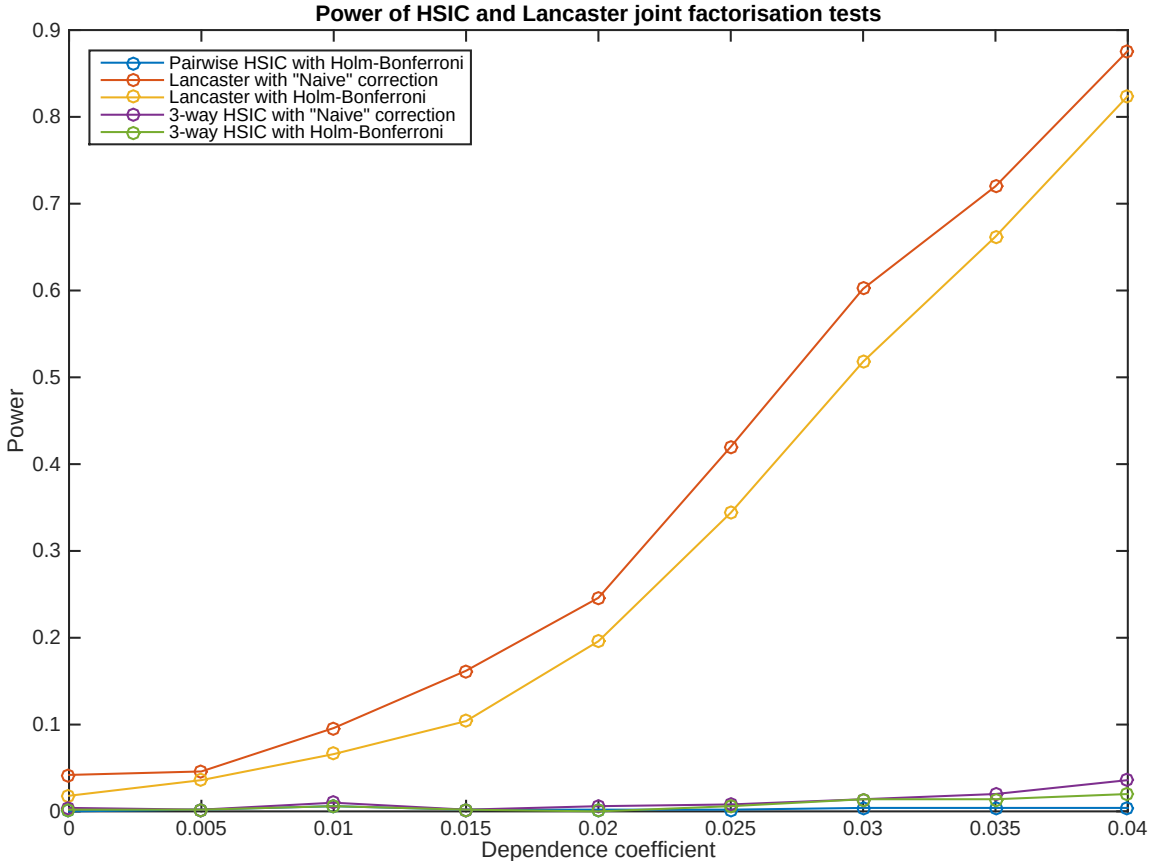

Figure 5: Performance of Lancaster and HSIC-based joint dependence tests on data from Example 2. Observe that the Pairwise HSIC does not detect any dependence at all. In the range of dependence coefficients considered, 3-way HSIC appears to be unable to detect dependence whereas Lancaster performs well. Note again that Lancaster with 'naive' correction performs better than with Holm-Bonferroni.

4.3.3 Example 3: Artificial data

The purpose of this example is to understand how the tests behave when the null hypothesis is true, and how well we can control the Type I error rates.

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

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

where $X_0, Y_0, Z_0, \epsilon_t, \eta_t$ and ζ_t are *iid* $\mathcal{N}(0,1)$ random variables. Observe that $(X, Z) \perp\!\!\!\perp Y$, and so the null hypothesis is true.

4000 datasets were generated according to this definition, each consisting of 2000 consecutive observations of the variables. We ran the Wild Bootstrap with 250 bootstrapping procedures and we used a Gaussian kernel with bandwidth parameter 1 on each of X, Y and Z. For each test, the p-values were recorded. For any desired Type I error rate α, the proportion of tests that would have resulted in a rejection of the null could then be calculated. This is presented in Figure 6.

Ideally, we would have an empirical Type I error that is very close to, but bounded by, α. In this case, we have a good understanding of the Type I error and therefore can choose a threshold in an informed way to control the tradeoff between sensitivity and specificity. It is 'bad' if the empirical Type I error is drastically less than the desired level, as this means that the sensitivity of the test is lower than it would otherwise be[5].

Observe that for 3-way HSIC, the empirical Type I error rate is almost exactly as desired. For Pairwise-HSIC, the empirical Type I error rate is slightly less than the desired level. The Lancaster tests have a much lower empirical Type I error than desired. Using the 'naive' correction gives a considerable improvement over

[5]It is 'even worse' if the empirical Type I error is not bounded by the desired level. In this case, it is not possible to control the false positive rate! It is for this reason that the Wild Bootstrap is needed, rather than using the permutation bootstrap that works in the *iid case*. See Example 4.

the Holm-Bonferroni correction, though it is still significantly lower than the desired error.

4.3.4 Example 4: Artificial data

The purpose of this example is to demonstrate that the Wild Bootstrap is actually needed when resampling from the null distribution.

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

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

where $X_0, Y_0, Z_0, \epsilon_t, \eta_t$ and ζ_t are *iid* $\mathcal{N}(0,1)$ random variables and a, called the *dependence coefficient*, determines how temporally dependent the processes are. Observe that each process is independent of the others and so the null hypothesis is true.

We performed the Lancaster test using both the Wild Bootstrap and simple permutation bootstrap (used in the *iid* case) methods to sample from the null distribution. We used a fixed desired false positive rate $\alpha = 0.05$ with sample of size 1000, with 500 experiments run for each value of a. Figure 7 shows the false positive rates for these two methods for varying a. It shows that as the processes become more dependent, the false positive rate for the permutation method becomes very large, and is not bounded by the fixed α, whereas the false positive rate for the Wild Bootstrap method is bounded by α.

4.3.5 Example 5: Forex data

We performed the tests on exchange rates CHF/USD, CHF/GBP and USD/CAD from 01/01/1990 until 31/12/1999. We tried to answer two questions using this data:

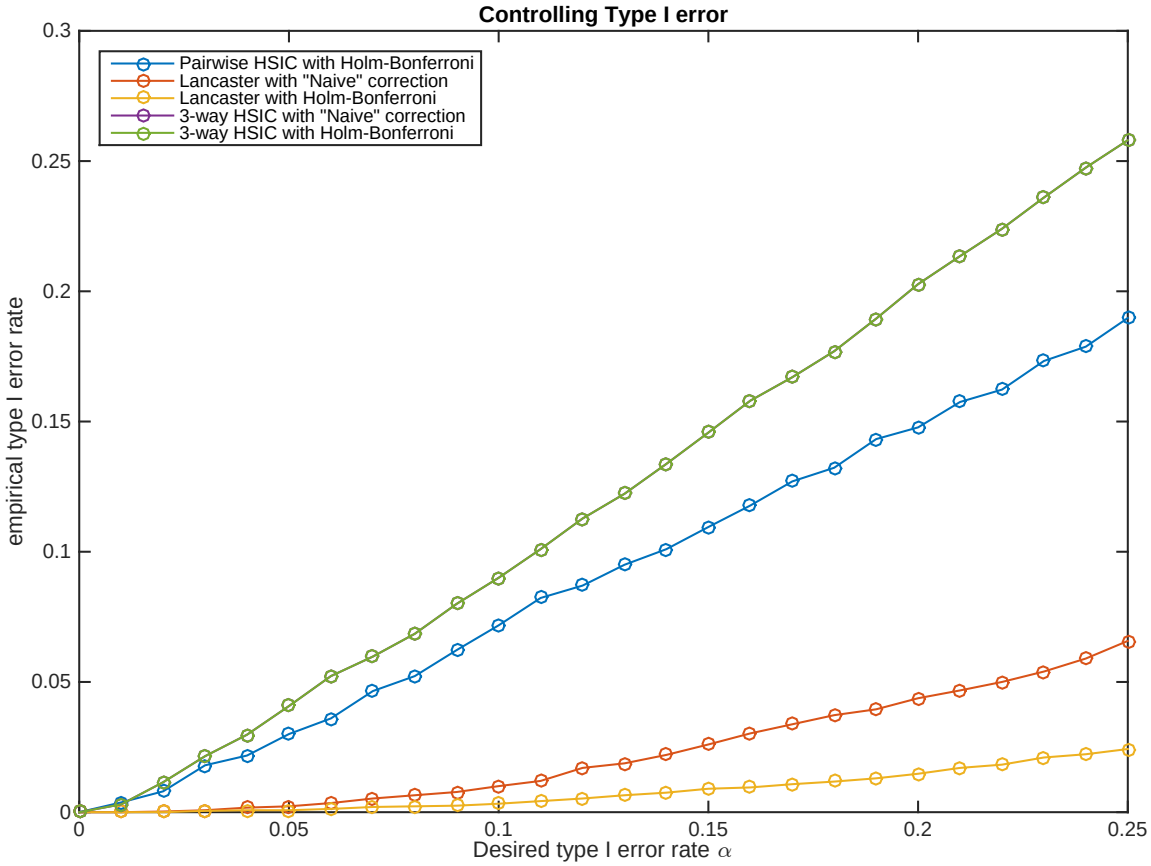

Figure 6: Performance of Lancaster and HSIC-based joint dependence tests on data from Example 3. Here the null hypothesis is true, and we measure the empirical Type I error rate as a function of the desired Type I error rate. Observe that 3-way HSIC achieves the desired Type I error rate more or less exactly. Note that using Holm-Bonferroni or the 'naive' correction seemingly makes no difference for 3-way HSIC - indeed, the purple line is not visible as it is perfectly obscured by the green line. Pairwise HSIC is also close to the desired error. Note that for Lancaster, using the 'naive' rather than Holm-Bonferroni correction results in a Type I error rate that is closer to the desired rate, thus implying that this correction gives a better test power.

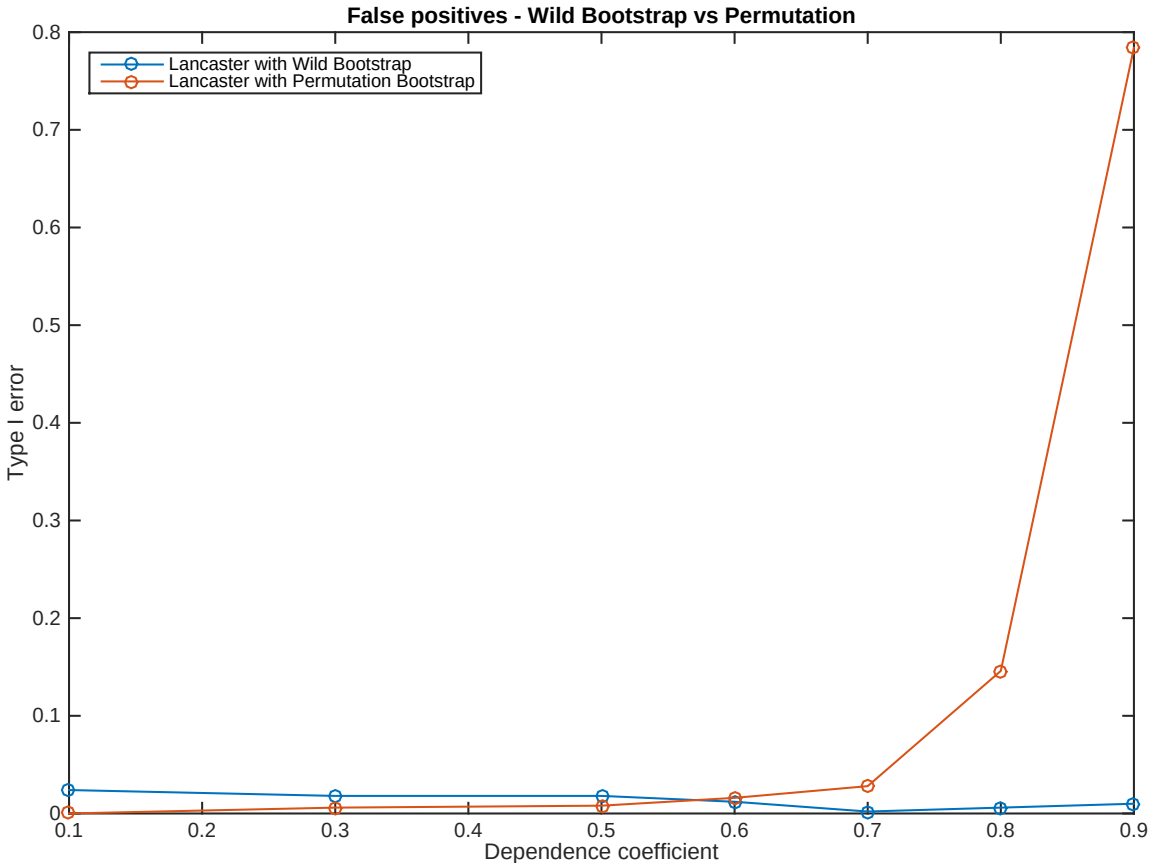

Figure 7: Empirical Type I errors of the Lancaster test when using the Wild Bootstrap and simple permutation bootstrap methods. Observe that when the dependence coefficient is large, a very large false positive rate is obtained when using the permutation bootstrap. This implies that the Wild Bootstrap is indeed actually needed - the permutation bootstrap does not work in this example.

(i) Are the exchange rates themselves dependent?

(ii) Are the fluctuations within each time series dependent?

We performed two different types of preprocessing before running the tests.

To answer (i), we took logarithms of each datum, then centred each time series with respect to its mean and then scaled each time series to have unit variance. We refer to these processed time series as the normalised time series. Figure 8 displays the normalised time series. Observe that these time series do not appear to be stationary, and so application of the Wild Bootstrap may not be valid.

To answer (ii), we first took logarithms of each datum. We then smoothened each time series by taking a 5-day running average, and subtracted these from the original time series. We then centred and scaled these 'fluctuation time series' to each have zero mean and unit variance. Figure 9 shows the fluctuation time series after this procedure.

For the output of the code when run on these datasets, see the appendix. For the normalised data, both the HSIC-based tests and the Lancaster test (with both 'naive' and Holm-Bonferroni corrections) rejected the null hypothesis (ie the joint distribution does not factorise). For the fluctuation data, the HSIC tests rejected the null hypothesis, however Lancaster failed to reject the null hypothesis.

4.3.6 Example 6: Forex data

We performed the tests on exchange rates GBP/USD, USD/HRK and GBP/HRK for 4534 working days days from 09/09/1996. We tried to answer the same questions as in Example 5 and so processed the data in the same way. Figure 10 displays the normalised time series. Figure 11 displays the fluctuation time series. Note that the time series in Figure 10 do not appear to be stationary and so application of the Wild Bootstrap may not be valid.

For the output of the code when run on these datasets, see the appendix. For both sets of processed data, all of the tests reject the null hypothesis (ie they detect that the distribution does not factorise and thus the variables are all dependent).

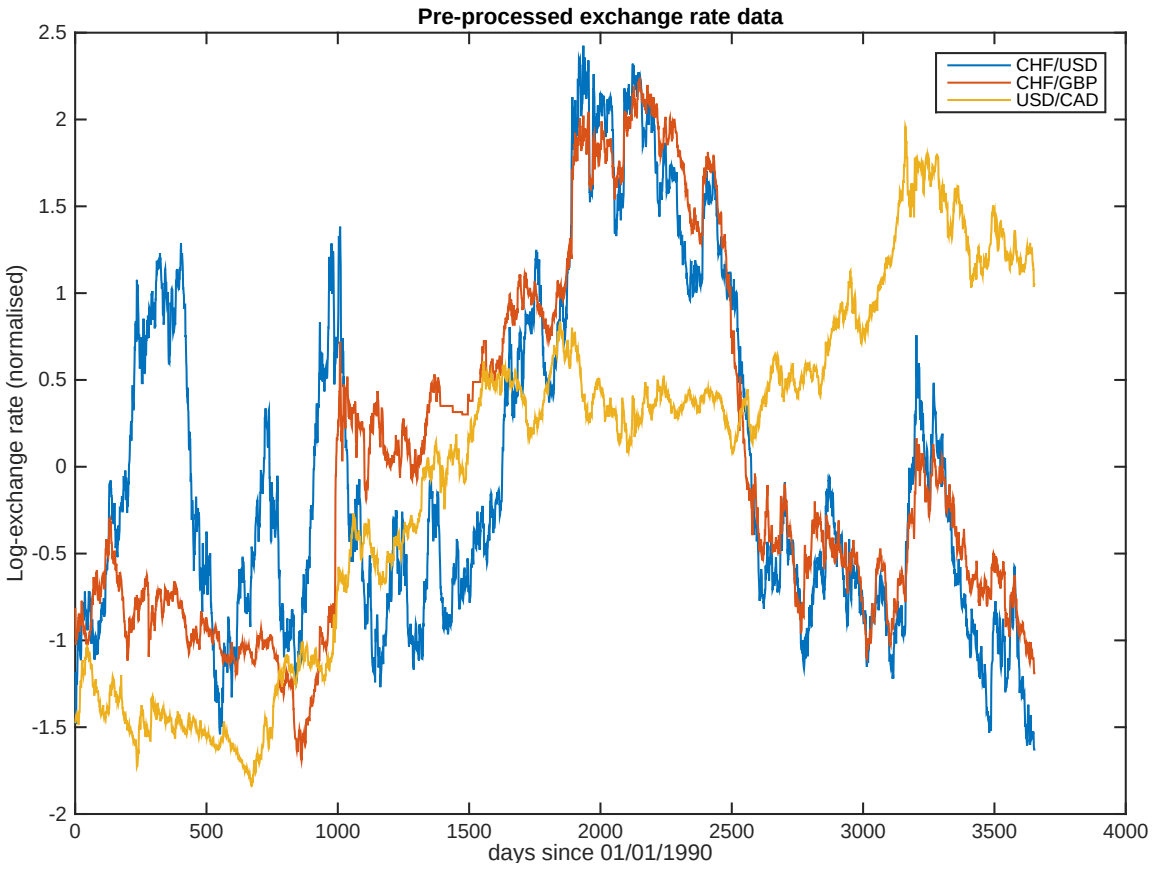

Figure 8: Normalised time series for data in Example 5. Observe that these time series do not appear to be stationary, and so application of the Wild Bootstrap may not be valid.

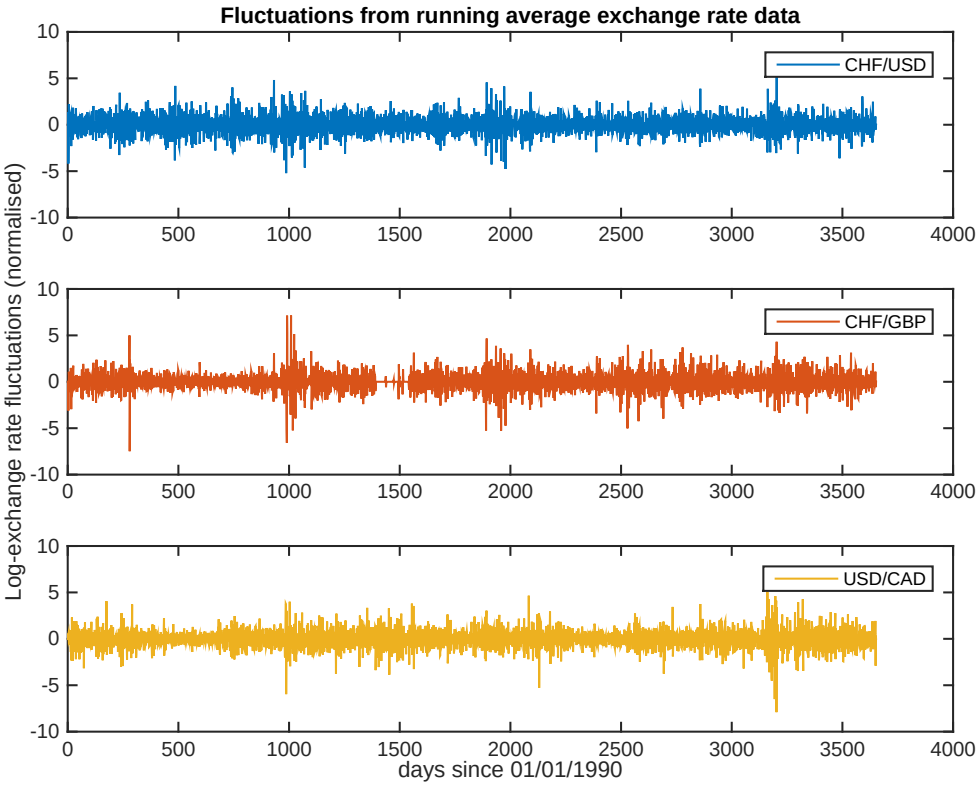

Figure 9: Fluctuation time series for data in Example 5.

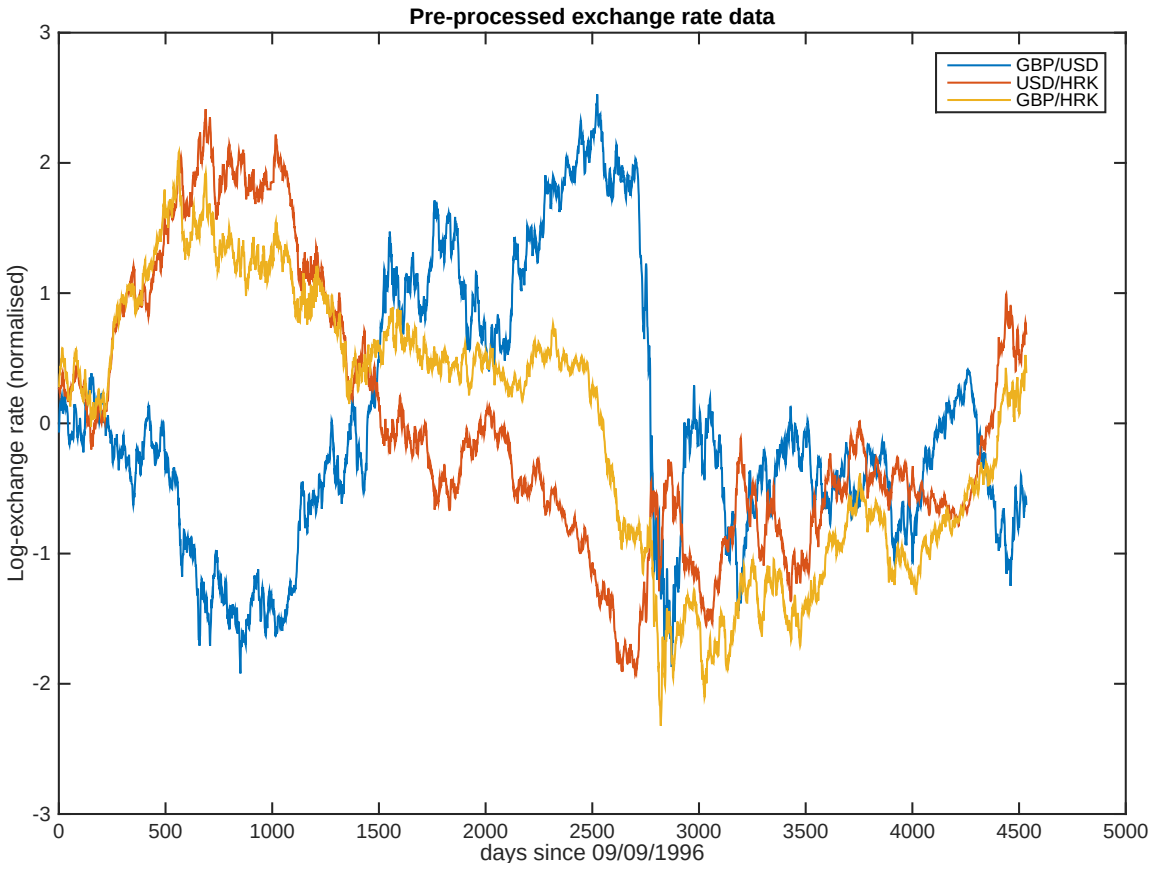

Figure 10: Normalised time series for data in Example 6. Observe that these time series do not appear to be stationary, and so application of the Wild Bootstrap may not be valid.

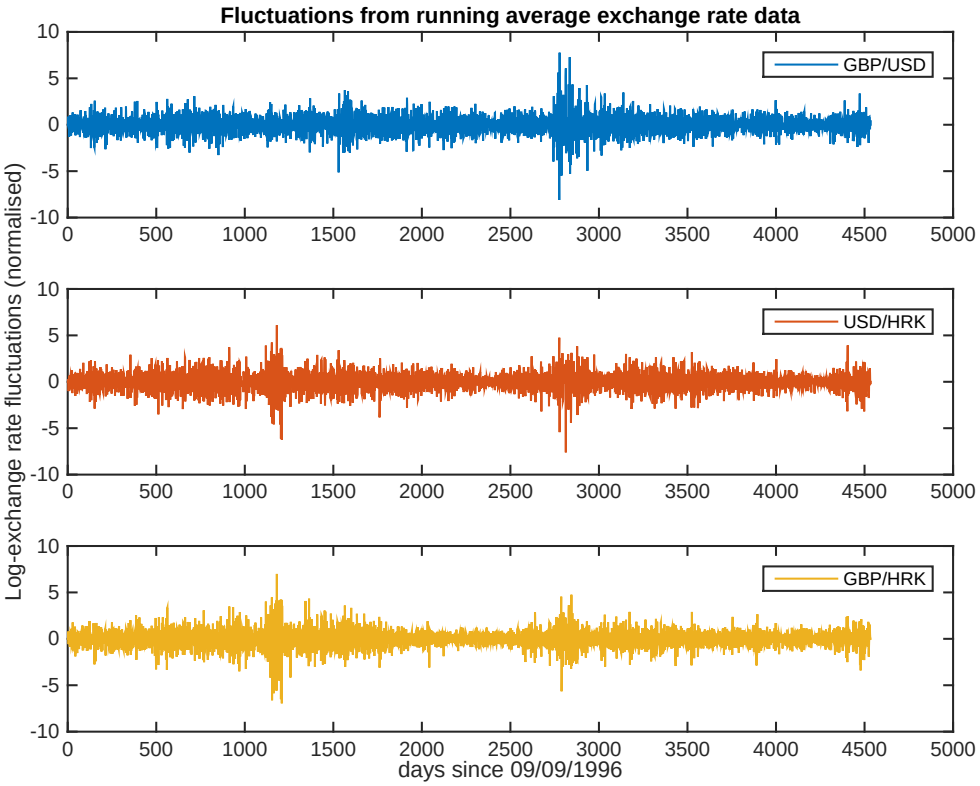

Figure 11: Fluctuation time series for data in Example 6.

4.4 Discussion of results

Let us note first that the Lancaster test can detect dependence in some circumstances for which the HSIC-based tests fail to detect dependence, such as in Example 2. This is an example of a situation in which Z is weakly dependent on X and Y separately, but strongly dependent on the pair (X, Y).

In cases for which the pairwise dependence is strong, it appears that the HSIC-based tests have a better power than the Lancaster test, as demonstrated by Example 1.

There is a confounding factor here though - as demonstrated by Example 3, the Type I error of the Lancaster test has a much cruder bound than those for Pairwise HSIC and 3-way HSIC. It may be possible that the lower power of the Lancaster test is due, at least in part, to the fact that we are setting our thresholds for the p-values to be too low with the result that we fail to reject too many cases in which the null hypothesis in fact does not hold.

Examples 1, 2 and 3 together serve to demonstrate that the 'naive' multiple testing correction is better than the previously proposed Holm-Bonferroni correction in [4]. This improvement should hold in the *iid* case considered in [4] too.

Looking at the real data analysed in Examples 5 and 6, we see that the Lancaster test did not allow us to draw any conclusions beyond what we learned from the HSIC-based tests, however it is possible that it may be useful beyond HSIC in other circumstances. One 'problem' with financial data is that there are many confounding factors, and so it is in general quite hard to find pairs of variables that are marginally independent of one another (or weakly dependent) and yet (strongly) dependent via a third - these are the situations in which the Lancaster test is particularly useful.

It is worth nothing that the Lancaster test performs well on the data in Example 6 - here there are good reasons to believe that there is a very strong 3-way dependence between the three random variables under consideration, beyond the dependence found between pairs of variables. Indeed, since the three variables cover all exchange rates between three currencies, any two should determine the third (since trading, for example, GBP $\longrightarrow$ USD $\longrightarrow$ HRK $\longrightarrow$ GBP should result in no net loss or gain,

else there would be opportunity for arbitrage).

It should be noted that it is not clear whether the time series used in Examples 5 and 6 satisfy the conditions required of the Wild Bootstrap. 'By eye' inspection suggests that stationarity does not hold in the normalised time series. Whether or not the β- and τ-mixing conditions hold in either the normalised time series or the fluctation timeseries is unclear.

5 Conclusions and directions for further research

In this thesis we present a kernel statistical test of dependence between three stationary random processes that satisfy β- and τ-mixing assumptions. The null hypothesis of this test is that the joint distribution $\mathbb{P}_{XYZ}$ factorises in some way, so that rejection of the null implies that $\mathbb{P}_{XYZ}$ does not factorise. This uses the Lancaster interaction as its test statistic, and uses the Wild Bootstrap to resample the statistic under the null distribution.

In order to show that use of the Wild Bootstrap results in samples from the correct null distribution, we prove that the normalised Lancaster statistic is a degenerate V-statistic under the null hypothesis. This is the main contribution of this thesis. The same proof idea is used to give a new proof that the Wild Bootstrap can be used with HSIC when the observations are drawn from random processes satisfying the same conditions as stated above.

When performing the Lancaster statistical test, we test multiple hypotheses and so consider multiple testing corrections. A minor contribution of this thesis is to show that the existing corrections used were more conservative than necessary, and a new, better correction is provided resulting in a greater test power.

Comparing the performance of the Lancaster test with HSIC-based tests on artificial data shows that, when the three variables interact weakly when considered pairwise, but strongly when considered all together, the Lancaster test outperforms the HSIC-based tests. However, when strong pairwise interactions are present, the HSIC-based tests considered are more able to identify joint dependence. An interesting finding that may account for some of the relatively-better-power of the HSIC-based tests was that even with the better multiple testing correction, the empirical bound on the Type I errors was found to be very severe. Thus, some power of the Lancaster test may be lost due to in practice having a more extreme test statistic threshold than necessary for any given Type I error bound.

Moving forward, there are many questions that this research has raised. We list some here.

- Recall that the statement of the theorem of the Wild Bootstrap lists three

sets of conditions. They concern: (1) Conditions on the observations; (2) Conditions on the test statistic; (3) Conditions on the bootstrapping process. In this thesis we have proved that the Lancaster and HSIC test statistics satisfy (2), but what about the other two conditions? How can we tell if observations satisfy (1)? What happens if we choose a different bootstrap process satisfying (3)?

- Recall that we use a Hilbert space Central Limit Theorem for random processes in our proof. Currently we assume that they are β-mixing, in addition to the τ-mixing already needed for the wild bootstrap. Is it possible to relax the conditions on the processes?

- Experiment 3 shows that we are bounding the Type I error rate too severely, possibly at the expense of test power. Can we better understand how to bound the false positive rate?

- When we fail to reject the null hypothesis of the Lancaster test, we cannot conclude anything about the distribution. This is because the Lancaster statistic can be zero *even when the joint distribution does not factorise*. Better understanding these 'counterexamples' and possibly even characterising them might help us to create a consistent test of joint dependence on three variables.

Table 2: V-statistic estimates of $\langle\langle \nu, \nu' \rangle\rangle_{k \otimes l}$ in the two-variable case. Note that this table has been copied exactly from [4], and is not the original work of the author of this thesis

$\nu \backslash \nu'$	P_{XY}	$P_X P_Y$
P_{XY}	$\frac{1}{n^2} (K \circ L)_{++}$	$\frac{1}{n^3} (KL)_{++}$
$P_X P_Y$		$\frac{1}{n^4} K_{++} L_{++}$

6 Appendix

6.1 Proofs

Claim 6.1. *Using the same setup as in Claim 3.1,*
$T(\phi, \psi, \mathcal{D}) = \frac{1}{n^2}(K \circ L)_{++} - \frac{2}{n^3}(KL)_{++} + \frac{1}{n^2} K_{++} L_{++}$

There are two ways to prove this. The first relies on the Lancaster paper [4], and is relatively short. The second is from direct manipulation of the definition of T and is only straightforward, but extremely tedious, algebra.
Proof: (i) *Using the results from Lancaster paper* Noting first by definition of T, and then by *Section 4.1* in [4],

$$\begin{aligned} T(\phi, \psi, \omega, \mathcal{D}) &= \frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++} \\ &= \|\hat{P}_{XY} - \hat{P}_X \hat{P}_Y\|^2_{k \otimes l} \\ &= \langle \hat{P}_{XY}, \hat{P}_{XY} \rangle - 2\langle \hat{P}_{XY}, \hat{P}_X \hat{P}_Y \rangle + \langle \hat{P}_X \hat{P}_Y, \hat{P}_X \hat{P}_Y \rangle \end{aligned}$$

Each of these inner products can be expressed in terms of the Gram matrices K and L. *Table 1* in [4], which has been exactly copied here as Table 2 for convenience, gives us each of these expressions. Substituting yields the desired result.

■

Proof: (ii) *A direct proof*

$$
\begin{aligned}
&T(\phi,\psi,\mathcal{D}) \\
&= \frac{1}{n^2}\sum_{ij}\Big\langle \phi(X_i) - \frac{1}{n}\sum_k \phi(X_k), \phi(X_j) - \frac{1}{n}\sum_k \phi(X_k)\Big\rangle \\
&\qquad\qquad \times \Big\langle \psi(Y_i) - \frac{1}{n}\sum_k \psi(Y_k), \psi(Y_j) - \frac{1}{n}\sum_k \psi(Y_k)\Big\rangle \\
&= \frac{1}{n^2}\sum_{ij}\Big\{\langle \phi(X_i),\phi(X_j)\rangle - \frac{1}{n}\sum_k \langle \phi(X_i),\phi(X_k)\rangle \\
&\qquad\qquad - \frac{1}{n}\sum_k \langle \phi(X_k),\phi(X_j)\rangle + \frac{1}{n^2}\sum_{kl}\langle \phi(X_k),\phi(X_l)\rangle\Big\} \\
&\quad \times \Big\{\langle \psi(Y_i),\psi(Y_j)\rangle - \frac{1}{n}\sum_k \langle \psi(Y_i),\psi(Y_k)\rangle - \frac{1}{n}\sum_k \langle \psi(Y_k),\psi(Y_j)\rangle + \frac{1}{n^2}\sum_{kl}\langle \psi(Y_k),\psi(Y_l)\rangle\Big\} \\
&= \frac{1}{n^2}\sum_{ij}\Big\{K_{ij} - \frac{1}{n}K_{i+} - \frac{1}{n}K_{+j} + \frac{1}{n^2}K_{++}\Big\}\Big\{L_{ij} - \frac{1}{n}L_{i+} - \frac{1}{n}L_{+j} + \frac{1}{n^2}L_{++}\Big\} \\
&= \frac{1}{n^2}\sum_{ij}\Big\{K_{ij}L_{ij} - \frac{1}{n}K_{ij}L_{i+} - \frac{1}{n}K_{ij}L_{+j} + \frac{1}{n^2}K_{ij}L_{++} \\
&- \frac{1}{n}K_{i+}L_{ij} + \frac{1}{n^2}K_{i+}L_{i+} + \frac{1}{n^2}K_{i+}L_{+j} - \frac{1}{n^3}K_{i+}L_{++} \\
&- \frac{1}{n}K_{+j}L_{ij} + \frac{1}{n^2}K_{+j}L_{i+} + \frac{1}{n^2}K_{+j}L_{+j} - \frac{1}{n^3}K_{+j}L_{++} \\
&+ \frac{1}{n^2}K_{++}L_{ij} - \frac{1}{n^3}K_{++}L_{i+} - \frac{1}{n^3}K_{++}L_{+j} + \frac{1}{n^4}K_{++}L_{++}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{n^2}\Big\{\sum_{ij}K_{ij}L_{ij}-\frac{1}{n}\sum_i K_{i+}L_{i+}-\frac{1}{n}\sum_j K_{+j}L_{+j}+\frac{1}{n^2}K_{++}L_{++}\\
&-\frac{1}{n}\sum_i K_{i+}L_{i+}+\frac{1}{n}\sum_i K_{i+}L_{i+}+\frac{1}{n^2}K_{++}L_{++}-\frac{1}{n^2}K_{++}L_{++}\\
&-\frac{1}{n}\sum_j K_{+j}L_{+j}+\frac{1}{n^2}K_{++}L_{++}+\frac{1}{n}\sum_j K_{+j}L_{+j}-\frac{1}{n^2}K_{++}L_{++}\\
&+\frac{1}{n^2}K_{++}L_{++}-\frac{1}{n^2}K_{++}L_{++}-\frac{1}{n^2}K_{++}L_{++}+\frac{1}{n^2}K_{++}L_{++}\Big\}
\end{aligned}
$$

$$
=\frac{1}{n^2}(K\circ L)_{++}-\frac{2}{n^3}(KL)_{++}+\frac{1}{n^2}K_{++}L_{++}
$$

where the last equality follows due to the coloured terms cancelling and the middle two black terms being equal by symmetry of K and L. ■

Claim 6.2. *Using the same setup as in Claim 3.3,*

$$
\begin{aligned}
T(\phi,\psi,\omega,\mathcal{D})=&\frac{1}{n^2}(K\circ L\circ M)_{++}-\frac{2}{n^3}((K\circ L)M)_{++}-\frac{2}{n^3}((K\circ M)L)_{++}\\
&-\frac{2}{n^3}((M\circ L)K)_{++}+\frac{1}{n^4}(K\circ L)_{++}M_{++}+\frac{1}{n^4}(K\circ M)_{++}L_{++}\\
&+\frac{1}{n^4}(L\circ M)_{++}K_{++}+\frac{2}{n^4}(MKL)_{++}+\frac{2}{n^4}(KLM)_{++}\\
&+\frac{2}{n^4}(KML)_{++}+\frac{4}{n^4}tr(K_+\circ L_+\circ M_+)-\frac{4}{n^5}(KL)_{++}M_{++}\\
&-\frac{4}{n^5}(KM)_{++}L_{++}-\frac{4}{n^5}(LM)_{++}K_{++}+\frac{4}{n^6}K_{++}L_{++}M_{++}
\end{aligned}
$$

There are two ways to prove this. The first relies on the Lancaster paper [4], and is relatively short. The second is from direct manipulation of the definition of T and is only straightforward, but extremely tedious, algebra. The second proof is included to demonstrate that the result *can* be derived from 'first principles', without the need for advanced theory, though the author would recommend that the first proof is a much better way of thinking about the problem.

Proof: (i) *Using Lancaster interaction paper*

Noting first by definition of T, and then by *Proposition 3* in [4],

$$\begin{aligned} T(\phi, \psi, \omega, \mathcal{D}) &= \frac{1}{n^2}(\tilde{K} \circ \tilde{L} \circ \tilde{M})_{++} \\ &= \|\Delta_L \hat{P}\|^2_{k \otimes l \otimes m} \end{aligned}$$

Next, expanding $\Delta_L \hat{P}$ in terms of empirical embeddings of various factorisations of the joint, as in equation (2) of [4] yields

$$\begin{aligned} \|\Delta_L \hat{P}\|^2_{k \otimes l \otimes m} &= \|\hat{P}_{XYZ} - \hat{P}_{XY}\hat{P}_Z - \hat{P}_{YZ}\hat{P}_X - \hat{P}_{XZ}\hat{P}_Y + 2\hat{P}_X\hat{P}_Y\hat{P}_Z\|^2_{k \otimes l \otimes m} \\ &= \langle \hat{P}_{XYZ}, \hat{P}_{XYZ} \rangle - \langle \hat{P}_{XYZ}, \hat{P}_{XY}\hat{P}_Z \rangle - \langle \hat{P}_{XYZ}, \hat{P}_{YZ}\hat{P}_X \rangle \\ &\quad - \langle \hat{P}_{XYZ}, \hat{P}_{XZ}\hat{P}_Y \rangle + 2\langle \hat{P}_{XYZ}, \hat{P}_X\hat{P}_Y\hat{P}_Z \rangle \\ &\quad - \langle \hat{P}_{XY}\hat{P}_Z, \hat{P}_{XYZ} \rangle + \langle \hat{P}_{XY}\hat{P}_Z, \hat{P}_{XY}\hat{P}_Z \rangle + \langle \hat{P}_{XY}\hat{P}_Z, \hat{P}_{YZ}\hat{P}_X \rangle \\ &\quad + \langle \hat{P}_{XY}\hat{P}_Z, \hat{P}_{XZ}\hat{P}_Y \rangle - 2\langle \hat{P}_{XY}\hat{P}_Z, \hat{P}_X\hat{P}_Y\hat{P}_Z \rangle \\ &\quad - \langle \hat{P}_{YZ}\hat{P}_X, \hat{P}_{XYZ} \rangle + \langle \hat{P}_{YZ}\hat{P}_X, \hat{P}_{XY}\hat{P}_Z \rangle + \langle \hat{P}_{YZ}\hat{P}_X, \hat{P}_{YZ}\hat{P}_X \rangle \\ &\quad + \langle \hat{P}_{YZ}\hat{P}_X, \hat{P}_{XZ}\hat{P}_Y \rangle - 2\langle \hat{P}_{YZ}\hat{P}_X, \hat{P}_X\hat{P}_Y\hat{P}_Z \rangle \\ &\quad - \langle \hat{P}_{XZ}\hat{P}_Y, \hat{P}_{XYZ} \rangle + \langle \hat{P}_{XZ}\hat{P}_Y, \hat{P}_{XY}\hat{P}_Z \rangle + \langle \hat{P}_{XZ}\hat{P}_Y, \hat{P}_{YZ}\hat{P}_X \rangle \\ &\quad + \langle \hat{P}_{XZ}\hat{P}_Y, \hat{P}_{XZ}\hat{P}_Y \rangle - 2\langle \hat{P}_{XZ}\hat{P}_Y, \hat{P}_X\hat{P}_Y\hat{P}_Z \rangle \\ &\quad + 2\langle \hat{P}_X\hat{P}_Y\hat{P}_Z, \hat{P}_{XYZ} \rangle - 2\langle \hat{P}_X\hat{P}_Y\hat{P}_Z, \hat{P}_{XY}\hat{P}_Z \rangle - 2\langle \hat{P}_X\hat{P}_Y\hat{P}_Z, \hat{P}_{YZ}\hat{P}_X \rangle \\ &\quad - 2\langle \hat{P}_X\hat{P}_Y\hat{P}_Z, \hat{P}_{XZ}\hat{P}_Y \rangle + 4\langle \hat{P}_X\hat{P}_Y\hat{P}_Z, \hat{P}_X\hat{P}_Y\hat{P}_Z \rangle \end{aligned}$$

Each of these inner products can be expressed in terms of the Gram matrices K, L and M. *Table 2* in [4], which has been exactly copied here as Table 3 for convenience, gives us each of these expressions. Substituting yields the desired result.

■

Table 3: V-statistic estimates of $\langle\langle \nu, \nu' \rangle\rangle_{k \otimes l \otimes m}$ in the three-variable case. Note that this table has been copied exactly from [4], and is not the original work of the author of this thesis

$\nu \backslash \nu'$	nP_{XYZ}	$n^2 P_{XY} P_Z$	$n^2 P_{XZ} P_Y$	$n^2 P_{YZ} P_X$	$n^3 P_X P_Y P_Z$
nP_{XYZ}	$(K \circ L \circ M)_{++}$	$((K \circ L)\, M)_{++}$	$((K \circ M)\, L)_{++}$	$((M \circ L)\, K)_{++}$	$tr(K_+ \circ L_+ \circ M_+)$
$n^2 P_{XY} P_Z$		$(K \circ L)_{++}\, M_{++}$	$(MKL)_{++}$	$(KLM)_{++}$	$(KL)_{++} M_{++}$
$n^2 P_{XZ} P_Y$			$(K \circ M)_{++}\, L_{++}$	$(KML)_{++}$	$(KM)_{++} L_{++}$
$n^2 P_{YZ} P_X$				$(L \circ M)_{++}\, K_{++}$	$(LM)_{++} K_{++}$
$n^3 P_X P_Y P_Z$					$K_{++} L_{++} M_{++}$

Proof: (ii) *A direct proof (not recommended!)*

$$
\begin{aligned}
T(\phi, \psi, \omega, \mathcal{D}) = \frac{1}{n^2} \sum_{i,j} & \Big\langle \phi(X_i) - \frac{1}{n} \sum_k \phi(X_k), \phi(X_j) - \frac{1}{n} \sum_k \phi(X_k) \Big\rangle \\
& \times \Big\langle \psi(Y_i) - \frac{1}{n} \sum_k \psi(Y_k), \psi(Y_j) - \frac{1}{n} \sum_k \psi(Y_k) \Big\rangle \\
& \times \Big\langle \omega(Z_i) - \frac{1}{n} \sum_k \omega(Z_k), \omega(Z_j) - \frac{1}{n} \sum_k \omega(Z_k) \Big\rangle \\
= \frac{1}{n^2} \sum_{ij} & \Big\{ \langle \phi(X_i), \phi(X_j) \rangle - \frac{1}{n} \sum_k \langle \phi(X_i), \phi(X_k) \rangle \\
& \quad - \frac{1}{n} \sum_k \langle \phi(X_k), \phi(X_j) \rangle + \frac{1}{n^2} \sum_{kl} \langle \phi(X_k), \phi(X_l) \rangle \Big\} \\
& \times \Big\{ \langle \psi(Y_i), \psi(Y_j) \rangle - \frac{1}{n} \sum_k \langle \psi(Y_i), \psi(Y_k) \rangle \\
& \quad - \frac{1}{n} \sum_k \langle \psi(Y_k), \psi(Y_j) \rangle + \frac{1}{n^2} \sum_{kl} \langle \psi(Y_k), \psi(Y_l) \rangle \Big\} \\
& \times \Big\{ \langle \omega(Z_i), \omega(Z_j) \rangle - \frac{1}{n} \sum_k \langle \omega(Z_i), \omega(Z_k) \rangle \\
& \quad - \frac{1}{n} \sum_k \langle \omega(Z_k), \omega(Z_j) \rangle + \frac{1}{n^2} \sum_{kl} \langle \omega(Z_k), \omega(Z_l) \rangle \Big\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{ij}\Big\{K_{ij} - \frac{1}{n}K_{i+} - \frac{1}{n}K_{+j} + \frac{1}{n^2}K_{++}\Big\} \\
&\quad \times \Big\{L_{ij} - \frac{1}{n}L_{i+} - \frac{1}{n}L_{+j} + \frac{1}{n^2}L_{++}\Big\} \\
&\quad \times \Big\{M_{ij} - \frac{1}{n}M_{i+} - \frac{1}{n}M_{+j} + \frac{1}{n^2}M_{++}\Big\} \\
&= \frac{1}{n^2}\sum_{ij}\Big\{K_{ij}L_{ij}M_{ij} - \frac{1}{n}K_{ij}L_{ij}M_{i+} - \frac{1}{n}K_{ij}L_{ij}M_{+j} + \frac{1}{n^2}K_{ij}L_{ij}M_{++} \\
&- \frac{1}{n}K_{ij}L_{i+}M_{ij} + \frac{1}{n^2}K_{ij}L_{i+}M_{i+} + \frac{1}{n^2}K_{ij}L_{i+}M_{+j} - \frac{1}{n^3}K_{ij}L_{i+}M_{++} \\
&- \frac{1}{n}K_{ij}L_{+j}M_{ij} + \frac{1}{n^2}K_{ij}L_{+j}M_{i+} + \frac{1}{n^2}K_{ij}L_{+j}M_{+j} - \frac{1}{n^3}K_{ij}L_{+j}M_{++} \\
&+ \frac{1}{n^2}K_{ij}L_{++}M_{ij} - \frac{1}{n^3}K_{ij}L_{++}M_{i+} - \frac{1}{n^3}K_{ij}L_{++}M_{+j} + \frac{1}{n^4}K_{ij}L_{++}M_{++} \\
&- \frac{1}{n}K_{i+}L_{ij}M_{ij} + \frac{1}{n^2}K_{i+}L_{ij}M_{i+} + \frac{1}{n^2}K_{i+}L_{ij}M_{+j} - \frac{1}{n^3}K_{i+}L_{ij}M_{++} \\
&+ \frac{1}{n^2}K_{i+}L_{i+}M_{ij} - \frac{1}{n^3}K_{i+}L_{i+}M_{i+} - \frac{1}{n^3}K_{i+}L_{i+}M_{+j} + \frac{1}{n^4}K_{i+}L_{i+}M_{++} \\
&+ \frac{1}{n^2}K_{i+}L_{+j}M_{ij} - \frac{1}{n^3}K_{i+}L_{+j}M_{i+} - \frac{1}{n^3}K_{i+}L_{+j}M_{+j} + \frac{1}{n^4}K_{i+}L_{+j}M_{++} \\
&- \frac{1}{n^3}K_{i+}L_{++}M_{ij} + \frac{1}{n^4}K_{i+}L_{++}M_{i+} + \frac{1}{n^4}K_{i+}L_{++}M_{+j} - \frac{1}{n^5}K_{i+}L_{++}M_{++} \\
&- \frac{1}{n}K_{+j}L_{ij}M_{ij} + \frac{1}{n^2}K_{+j}L_{ij}M_{i+} + \frac{1}{n^2}K_{+j}L_{ij}M_{+j} - \frac{1}{n^3}K_{+j}L_{ij}M_{++} \\
&+ \frac{1}{n^2}K_{+j}L_{i+}M_{ij} - \frac{1}{n^3}K_{+j}L_{i+}M_{i+} - \frac{1}{n^3}K_{+j}L_{i+}M_{+j} + \frac{1}{n^4}K_{+j}L_{i+}M_{++} \\
&+ \frac{1}{n^2}K_{+j}L_{+j}M_{ij} - \frac{1}{n^3}K_{+j}L_{+j}M_{i+} - \frac{1}{n^3}K_{+j}L_{+j}M_{+j} + \frac{1}{n^4}K_{+j}L_{+j}M_{++} \\
&- \frac{1}{n^3}K_{+j}L_{++}M_{ij} + \frac{1}{n^4}K_{+j}L_{++}M_{i+} + \frac{1}{n^4}K_{+j}L_{++}M_{+j} - \frac{1}{n^5}K_{+j}L_{++}M_{++} \\
&+ \frac{1}{n^2}K_{++}L_{ij}M_{ij} - \frac{1}{n^3}K_{++}L_{ij}M_{i+} - \frac{1}{n^3}K_{++}L_{ij}M_{+j} + \frac{1}{n^4}K_{++}L_{ij}M_{++} \\
&- \frac{1}{n^3}K_{++}L_{i+}M_{ij} + \frac{1}{n^4}K_{++}L_{i+}M_{i+} + \frac{1}{n^4}K_{++}L_{i+}M_{+j} - \frac{1}{n^5}K_{++}L_{i+}M_{++} \\
&- \frac{1}{n^3}K_{++}L_{+j}M_{ij} + \frac{1}{n^4}K_{++}L_{+j}M_{i+} + \frac{1}{n^4}K_{++}L_{+j}M_{+j} - \frac{1}{n^5}K_{++}L_{+j}M_{++} \\
&+ \frac{1}{n^4}K_{++}L_{++}M_{ij} - \frac{1}{n^5}K_{++}L_{++}M_{i+} - \frac{1}{n^5}K_{++}L_{++}M_{+j} + \frac{1}{n^6}K_{++}L_{++}M_{++}
\end{aligned}
$$

Summing over the indices (taking care that if an index is not present in a term, summing over this multiplies the term by a factor of n) and cancelling equal terms yields the required result.

■

Proof of Claim 3.4

Proof: To illustrate the symmetry, we used u, v and w and U, V and W in the statement of the claim. However, for notational ease we will prove the claim here using X, Y, Z etc.

$$
\begin{aligned}
(i) \quad \frac{1}{n}(K \circ L \circ M)_{++} &= \frac{1}{n}\sum_{ij}\langle\bar{\phi}(X_i), \bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_i), \bar{\psi}(Y_j)\rangle\langle\bar{\omega}(Z_i), \bar{\omega}(Z_j)\rangle \\
&= \frac{1}{n}\sum_{ij}\langle\bar{\phi}(X_i) \otimes \bar{\psi}(Y_i) \otimes \bar{\omega}(Z_i), \bar{\phi}(X_j) \otimes \bar{\psi}(Y_j) \otimes \bar{\omega}(Z_j)\rangle \\
&= n\Big\langle\frac{1}{n}\sum_{i}\bar{\phi}(X_i) \otimes \bar{\psi}(Y_i) \otimes \bar{\omega}(Z_i), \frac{1}{n}\sum_{j}\bar{\phi}(X_j) \otimes \bar{\psi}(Y_j) \otimes \bar{\omega}(Z_j)\Big\rangle \\
&= n\langle\bar{C}_{XYZ}, \bar{C}_{XYZ}\rangle
\end{aligned}
$$

$$
\begin{aligned}
(ii)\quad \frac{1}{n^2}((K\circ L)M)_{++} &= \frac{1}{n^2}\sum_{ijk}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_i),\bar{\psi}(Y_j)\rangle\langle\bar{\omega}(Z_j),\bar{\omega}(Z_k)\rangle \\
&= \frac{1}{n^2}\sum_{ijk}\langle\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)\otimes\bar{\omega}(Z_j),\bar{\phi}(X_i)\otimes\bar{\psi}(Y_i)\otimes\bar{\omega}(Z_k)\rangle \\
&= n\Big\langle\frac{1}{n}\sum_{j}\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)\otimes\bar{\omega}(Z_j), \\
&\qquad [\frac{1}{n}\sum_{i}\bar{\phi}(X_i)\otimes\bar{\psi}(Y_i)]\otimes[\frac{1}{n}\sum_{k}\bar{\omega}(Z_k)]\Big\rangle \\
&= n\langle\bar{C}_{XYZ},\bar{C}_{XY}\otimes\bar{\mu}_Z\rangle
\end{aligned}
$$

$$
\begin{aligned}
(iii)\quad \frac{1}{n^3}(K\circ L)_{++}M_{++} &= \frac{1}{n^3}\sum_{ijkl}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_i),\bar{\psi}(Y_j)\rangle\langle\bar{\omega}(Z_k),\bar{\omega}(Z_l)\rangle \\
&= \frac{1}{n^3}\sum_{ijkl}\langle\bar{\phi}(X_i)\otimes\bar{\psi}(Y_j)\otimes\bar{\omega}(Z_k),\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)\otimes\bar{\omega}(Z_l)\rangle \\
&= n\Big\langle[\frac{1}{n}\sum_{i}\bar{\phi}(X_i)\otimes\bar{\psi}(Y_i)]\otimes[\frac{1}{n}\sum_{k}\bar{\omega}(Z_k)], \\
&\qquad [\frac{1}{n}\sum_{j}\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)]\otimes[\frac{1}{n}\sum_{l}\bar{\omega}(Z_l)]\Big\rangle \\
&= n\langle\bar{C}_{XY}\otimes\bar{\mu}_Z,\bar{C}_{XY}\otimes\bar{\mu}_Z\rangle
\end{aligned}
$$

$$(iv) \quad \frac{1}{n^3}(KLM)_{++} = \frac{1}{n^3}\sum_{ijkl}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_j),\bar{\psi}(Y_k)\rangle\langle\bar{\omega}(Z_k),\bar{\omega}(Z_l)\rangle$$

$$= \frac{1}{n^3}\sum_{ijkl}\langle\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)\otimes\bar{\omega}(Z_l),\bar{\phi}(X_i)\otimes\bar{\psi}(Y_k)\otimes\bar{\omega}(Z_k)\rangle$$

$$= n\big\langle[\frac{1}{n}\sum_j\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)]\otimes[\frac{1}{n}\sum_l\bar{\omega}(Z_l)],$$

$$[\frac{1}{n}\sum_i\bar{\phi}(X_i)]\otimes[\frac{1}{n}\sum_k\bar{\psi}(Y_k)\otimes\bar{\omega}(Z_k)]\big\rangle$$

$$= n\langle\bar{C}_{XY}\otimes\bar{\mu}_Z,\bar{\mu}_X\otimes\bar{C}_{YZ}\rangle$$

$$(v) \quad \frac{1}{n^3}tr(K_+\circ L_+\circ M_+)_{++} = \frac{1}{n^3}\sum_{ijkl}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_i),\bar{\psi}(Y_k)\rangle\langle\bar{\omega}(Z_i),\bar{\omega}(Z_l)\rangle$$

$$= \frac{1}{n^3}\sum_{ijkl}\langle\bar{\phi}(X_i)\otimes\bar{\psi}(Y_i)\otimes\bar{\omega}(Z_i),\bar{\phi}(X_j)\otimes\bar{\psi}(Y_k)\otimes\bar{\omega}(Z_l)\rangle$$

$$= n\big\langle[\frac{1}{n}\sum_i\bar{\phi}(X_i)\otimes\bar{\psi}(Y_i)\otimes\bar{\omega}(Z_i)],$$

$$[\frac{1}{n}\sum_j\bar{\phi}(X_j)]\otimes[\frac{1}{n}\sum_k\bar{\psi}(Y_k)]\otimes[\frac{1}{n}\sum_l\bar{\omega}(Z_l)]\big\rangle$$

$$= n\langle\bar{C}_{XYX},\bar{\mu}_X\otimes\bar{\mu}_Y\otimes\bar{\mu}_Z\rangle$$

$$(vi)\quad \frac{1}{n^4}(KL)_{++}M_{++} = \frac{1}{n^4}\sum_{ijklp}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_j),\bar{\psi}(Y_k)\rangle\langle\bar{\omega}(Z_l),\bar{\omega}(Z_p)\rangle$$
$$= \frac{1}{n^4}\sum_{ijklp}\langle\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)\otimes\bar{\omega}(Z_l),\bar{\phi}(X_i)\otimes\bar{\psi}(Y_k)\otimes\bar{\omega}(Z_p)\rangle$$
$$= n\big\langle[\frac{1}{n}\sum_j\bar{\phi}(X_j)\otimes\bar{\psi}(Y_j)]\otimes[\frac{1}{n}\sum_l\bar{\omega}(Z_l)],$$
$$[\frac{1}{n}\sum_i\bar{\phi}(X_i)]\otimes[\frac{1}{n}\sum_k\bar{\psi}(Y_k)]\otimes[\frac{1}{n}\sum_p\bar{\omega}(Z_p)]\big\rangle$$
$$= n\langle\bar{C}_{XY}\otimes\bar{\mu}_Z,\bar{\mu}_X\otimes\bar{\mu}_Y\otimes\bar{\mu}_Z\rangle$$

$$(vii)\quad \frac{1}{n^5}K_{++}L_{++}M_{++} = \frac{1}{n^5}\sum_{ijklpq}\langle\bar{\phi}(X_i),\bar{\phi}(X_j)\rangle\langle\bar{\psi}(Y_k),\bar{\psi}(Y_l)\rangle\langle\bar{\omega}(Z_p),\bar{\omega}(Z_q)\rangle$$
$$= \frac{1}{n^5}\sum_{ijklpq}\langle\bar{\phi}(X_i)\otimes\bar{\psi}(Y_k)\otimes\bar{\omega}(Z_p),\bar{\phi}(X_j)\otimes\bar{\psi}(Y_l)\otimes\bar{\omega}(Z_q)\rangle$$
$$= n\big\langle[\frac{1}{n}\sum_i\bar{\phi}(X_i)]\otimes[\frac{1}{n}\sum_k\bar{\psi}(Y_k)]\otimes[\frac{1}{n}\sum_p\bar{\omega}(Z_p)],$$
$$[\frac{1}{n}\sum_j\bar{\phi}(X_j)]\otimes[\frac{1}{n}\sum_l\bar{\psi}(Y_l)]\otimes[\frac{1}{n}\sum_q\bar{\omega}(Z_q)]\big\rangle$$
$$= n\langle\bar{\mu}_X\otimes\bar{\mu}_Y\otimes\bar{\mu}_Z,\bar{\mu}_X\otimes\bar{\mu}_Y\otimes\bar{\mu}_Z\rangle$$

■

6.2 Code output from Example 5

6.2.1 Output for normalised timeseries

```
Lancaster test results
----------------------
x not independent of (y,z) [the distribution does not factorise]
      p-value: 0
y not independent of (x,z) [the distribution does not factorise]
      p-value: 0
z not independent of (x,y) [the distribution does not factorise]
      p-value: 0
total independence rejected [the distribution does not factorise]
      p-value: 0


Pairwise HSIC test results
----------------------
x and y are dependent [the distribution does not factorise]
      p-value: 0
x and z are dependent [the distribution does not factorise]
      p-value: 0
y and z are dependent [the distribution does not factorise]
      p-value: 0


Threeway HSIC test results
----------------------
(X,Y) and Z are dependent [the distribution does not factorise]
      p-value: 0
(X,Z) and Y are dependent [the distribution does not factorise]
      p-value: 0
```

```
(Y,Z) and X are dependent [the distribution does not factorise]
      p-value: 0

Lancaster Performing Holm-Bonferroni correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

Lancaster: Performing "Naive" (but better) correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

Performing pairwise HSIC joint factorisation test
with Holm-Bonferroni multiple correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing Holm-Bonferroni correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing "Naive" (but better) correction
----------------------
Reject null hypothesis: Joint distribution does not factorise
```

6.2.2 Output for fluctuation timeseries

```
Lancaster test results
----------------------
x independence of (y,z) cannot be rejected
      p-value: 0.168
```

```
y independence of (x,z) cannot be rejected
      p-value: 0.142
z independence of (x,y) cannot be rejected
      p-value: 0.128
total independence cannot be rejected
      p-value: 0.168


Pairwise HSIC test results
----------------------
x and y are dependent [the distribution does not factorise]
      p-value: 0
x and z are dependent [the distribution does not factorise]
      p-value: 0.018
y and z are dependent [the distribution does not factorise]
      p-value: 0


Threeway HSIC test results
----------------------
(X,Y) and Z are dependent [the distribution does not factorise]
      p-value: 0
(X,Z) and Y are dependent [the distribution does not factorise]
      p-value: 0
(Y,Z) and X are dependent [the distribution does not factorise]
      p-value: 0

Lancaster Performing Holm-Bonferroni correction
----------------------
Cannot reject null hypothesis: joint distribution may or may not factorise
```

```
Lancaster: Performing "Naive" (but better) correction
----------------------
Cannot reject null hypothesis: joint distribution may or may not factorise

Performing pairwise HSIC joint factorisation test
with Holm-Bonferroni multiple correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing Holm-Bonferroni correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing "Naive" (but better) correction
----------------------
Reject null hypothesis: Joint distribution does not factorise
```

6.3 Code output from Example 6

6.3.1 Output for normalised timeseries

```
Lancaster test results
----------------------
x not independent of (y,z) [the distribution does not factorise]
      p-value: 0
y not independent of (x,z) [the distribution does not factorise]
      p-value: 0
z not independent of (x,y) [the distribution does not factorise]
      p-value: 0
total independence rejected [the distribution does not factorise]
      p-value: 0


Pairwise HSIC test results
----------------------
x and y are dependent [the distribution does not factorise]
      p-value: 0
x and z are dependent [the distribution does not factorise]
      p-value: 0
y and z are dependent [the distribution does not factorise]
      p-value: 0


Threeway HSIC test results
----------------------
(X,Y) and Z are dependent [the distribution does not factorise]
      p-value: 0
(X,Z) and Y are dependent [the distribution does not factorise]
```

```
      p-value: 0
(Y,Z) and X are dependent [the distribution does not factorise]
      p-value: 0

Lancaster Performing Holm-Bonferroni correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

Lancaster: Performing "Naive" (but better) correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

Performing pairwise HSIC joint factorisation test
with Holm-Bonferroni multiple correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing Holm-Bonferroni correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing "Naive" (but better) correction
----------------------
Reject null hypothesis: Joint distribution does not factorise
```

6.3.2 Output for fluctuation timeseries

```
Lancaster test results
----------------------
x not independent of (y,z) [the distribution does not factorise]
```

```
      p-value: 0
y not independent of (x,z) [the distribution does not factorise]
      p-value: 0
z not independent of (x,y) [the distribution does not factorise]
      p-value: 0
total independence rejected [the distribution does not factorise]
      p-value: 0


Pairwise HSIC test results
----------------------
x and y are dependent [the distribution does not factorise]
      p-value: 0
x and z are dependent [the distribution does not factorise]
      p-value: 0
y and z are dependent [the distribution does not factorise]
      p-value: 0


Threeway HSIC test results
----------------------
(X,Y) and Z are dependent [the distribution does not factorise]
      p-value: 0
(X,Z) and Y are dependent [the distribution does not factorise]
      p-value: 0
(Y,Z) and X are dependent [the distribution does not factorise]
      p-value: 0

Lancaster Performing Holm-Bonferroni correction
----------------------
Reject null hypothesis: Joint distribution does not factorise
```

```
Lancaster: Performing "Naive" (but better) correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

Performing pairwise HSIC joint factorisation test
with Holm-Bonferroni multiple correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing Holm-Bonferroni correction
----------------------
Reject null hypothesis: Joint distribution does not factorise

3 way HSIC: Performing "Naive" (but better) correction
----------------------
Reject null hypothesis: Joint distribution does not factorise
```

References

[1] Arthur Gretton et al. "A kernel statistical test of independence". In: *Advances in Neural Information Processing Systems*. 2007, pp. 585–592.

[2] Arthur Gretton et al. "Measuring statistical dependence with Hilbert-Schmidt norms". In: *Algorithmic learning theory*. Springer. 2005, pp. 63–77.

[3] Kacper Chwialkowski and Arthur Gretton. "A kernel independence test for random processes". In: *arXiv preprint arXiv:1402.4501* (2014).

[4] Dino Sejdinovic, Arthur Gretton, and Wicher Bergsma. "A kernel test for three-variable interactions". In: *Advances in Neural Information Processing Systems*. 2013, pp. 1124–1132.

[5] Henry Oliver Lancaster. *Chi-Square Distribution*. Wiley Online Library, 1969.

[6] Kun Zhang et al. "Kernel-based conditional independence test and application in causal discovery". In: *arXiv preprint arXiv:1202.3775* (2012).

[7] Judea Pearl. *Causality*. Cambridge university press, 2009.

[8] Anne Leucht and Michael H Neumann. "Dependent wild bootstrap for degenerate U-and V-statistics". In: *Journal of Multivariate Analysis* 117 (2013), pp. 257–280.

[9] Xiaofeng Shao. "The dependent wild bootstrap". In: *Journal of the American Statistical Association* 105.489 (2010), pp. 218–235.

[10] Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. "A wild bootstrap for degenerate kernel tests". In: *Advances in neural information processing systems*. 2014, pp. 3608–3616.

[11] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. "Kernel methods in machine learning". In: *The annals of statistics* (2008), pp. 1171–1220.

[12] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[13] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[14] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[15] Christopher KI Williams and Carl Edward Rasmussen. "Gaussian processes for machine learning". In: *the MIT Press* 2.3 (2006), p. 4.

[16] Alex Smola et al. "A Hilbert space embedding for distributions". In: *Algorithmic Learning Theory*. Springer. 2007, pp. 13–31.

[17] Bharath K Sriperumbudur et al. "Hilbert space embeddings and metrics on probability measures". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1517–1561.

[18] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. "Universality, characteristic kernels and RKHS embedding of measures". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2389–2410.

[19] Arthur Gretton et al. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[20] Arthur Gretton et al. "Kernel methods for measuring independence". In: *The Journal of Machine Learning Research* 6 (2005), pp. 2075–2129.

[21] Paul Doukhan. *Mixing*. Springer, 1994.

[22] Richard C Bradley et al. "Basic properties of strong mixing conditions. A survey and some open questions". In: *Probability surveys* 2.2 (2005), pp. 107–144.

[23] Jérôme Dedecker et al. "Weak dependence". In: *Weak Dependence: With Examples and Applications*. Springer, 2007, pp. 9–20.

[24] Robert J Serfling. *Approximation theorems of mathematical statistics*. Vol. 162. John Wiley & Sons, 2009.

[25] Herold Dehling, Olimjon Sh Sharipov, and Martin Wendler. "Bootstrap for dependent Hilbert space-valued random variables with application to von Mises statistics". In: *Journal of Multivariate Analysis* 133 (2015), pp. 200–215.

[26] Sture Holm. "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics* (1979), pp. 65–70.