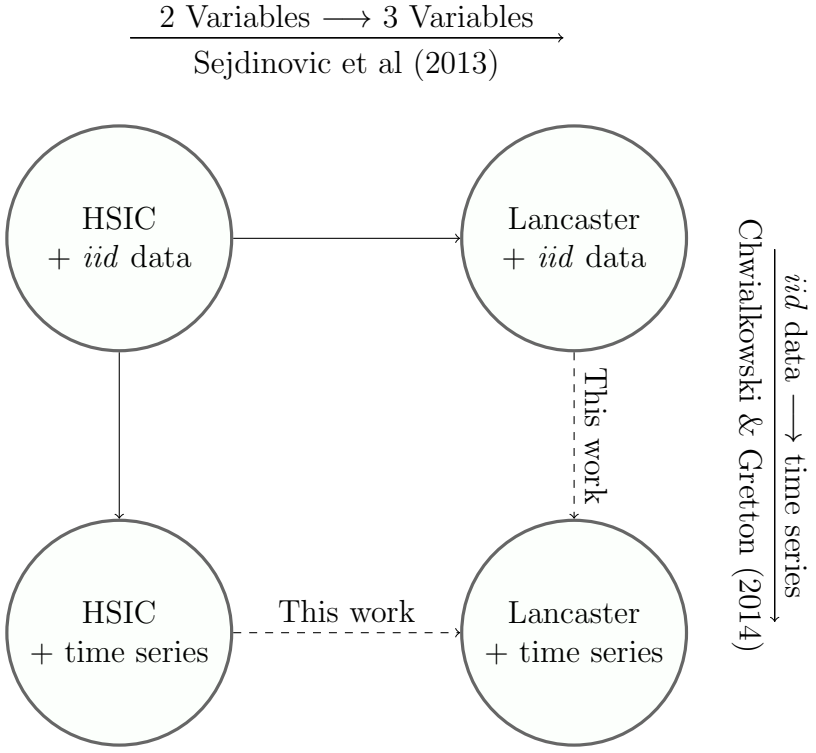
A Kernel Test for Three-Variable Interactions with Random Processes

Paul K Rubenstein

April 10, 2016

Context & Acknowledgements

Work was done with Arthur Gretton and Kacper Chwialkowski while I was a Masters student at UCL



Main Contributions

1. Proved that "wild bootstrap" procedure used for HSIC + timeseries can be applied to the Lancaster 3-variable interaction
2. Manner of proof used is simpler/shorter than existing methods (we use a Hilbert Space CLT rather than Hoeffding decomposition)

Outline

- Background
  - Kernel mean embedding
  - HSIC (2-variable interaction)
  - Lancaster (3-variable interaction)
- The non-iid case
  - Wild Bootstrap overview
  - Technical conditions
  - Our proof (outline)
- Results + open questions

Notation

Random variables	X	Y	Z
Domains	$\mathcal{X}$	$\mathcal{Y}$	$\mathcal{Z}$
Kernels	k	l	m
Gram matrices	K	L	M
RKHS	$\mathcal{F}_k$	$\mathcal{F}_l$	$\mathcal{F}_m$

'$\circ$' is Hadamard/element-wise matrix product: $(K \circ L)_{ij} = K_{ij}L_{ij}$
++ means 'sum all elements': $K_{++} = \sum_{ij} K_{ij}$

Kernel Mean Embedding

Suppose $X \sim \mathbb{P}$. We can embed $\mathbb{P}$ into $\mathcal{F}_k$ via the map:

$$\mathbb{P} \mapsto \mu_{\mathbb{P}} = \mathbb{E}_X k(\cdot, X)$$

Useful because for $f \in \mathcal{F}_k$, $\langle f, \mu_{\mathbb{P}} \rangle_k = \mathbb{E}_X f(X)$
Crucially for us, however:

1. We can exploit the geometric structure of the Hilbert space $\mathcal{F}_k$
2. If $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective, this is a metric
3. Empirical estimate $\hat{\mu}_{\mathbb{P}} = \frac{1}{n} \sum_i k(\cdot, X_i) \approx \mu_{\mathbb{P}}$ converges in RKHS norm as $O_P(n^{-1/2})$

Hilbert-Schmidt Independence Criterion

Problem:

Given *iid* samples $(X_i, Y_i)_{i=1}^n$, are X and Y independent?

Idea:

$\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y \iff \mu_{\mathbb{P}_{XY}} = \mu_{\mathbb{P}_X \mathbb{P}_Y}$

So if $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$, $HSIC = \|\hat{\mu}_{\mathbb{P}_{XY}} - \hat{\mu}_{\mathbb{P}_X \mathbb{P}_Y}\|_k^2$ should be 'small'.

We use $HSIC$ to test:

$$\begin{aligned} \mathcal{H}_0 &: \mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y \\ \mathcal{H}_1 &: \mathbb{P}_{XY} \neq \mathbb{P}_X \mathbb{P}_Y \end{aligned}$$

Can estimate quantiles of $HSIC$ under $\mathcal{H}_0$ by permuting indices of the X_i and recalculating $HSIC$ on many such 'fabricated' datasets.

Note: $HSIC$ can be expressed simply in terms of K and L:

$$\begin{aligned} HSIC &= \frac{1}{n^2}(K \circ L)_{++} - \frac{2}{n^3}(KL)_{++} + \frac{1}{n^4} K_{++} L_{++} \\ &= \frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++} \end{aligned}$$

where $\tilde{\cdot}$ indicates empirical centering.

We can write:

$$k(x, x') = \langle \phi_X(x), \phi_X(x') \rangle$$

Then $\tilde{K}$ is the gram matrix wrt

$$\tilde{k}(x, x') = \langle \phi_X(x) - \hat{\mu}_X, \phi_X(x') - \hat{\mu}_X \rangle$$

This amounts to recentering the feature map ϕ_X with respect to its empirical mean.

Lancaster Interaction

Problem:

Given *iid* samples $(X_i, Y_i, Z_i)_{i=1}^n$, do X, Y and Z exhibit mutual dependence? Specifically, does $\mathbb{P}_{XYZ}$ factorise? ($\mathbb{P}_X\mathbb{P}_{YZ}$, $\mathbb{P}_Y\mathbb{P}_{XZ}$, $\mathbb{P}_{XY}\mathbb{P}_Z$)

Idea:

We consider the Lancaster signed measure:

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

and define:

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

Then $\mathcal{H}_X \cup \mathcal{H}_Y \cup \mathcal{H}_Z \implies \Delta_L P = 0$

We use the norm of the empirical estimate $\|\Delta_L \hat{P}\|_{\mathcal{F}}^2$ to test:

$$\begin{aligned}\mathcal{H}_0 &: \mathcal{H}_X \cup \mathcal{H}_Y \cup \mathcal{H}_Z\\ \mathcal{H}_1 &: \mathbb{P}_{XYZ} \text{ does not factorise}\end{aligned}$$

We test each $\mathcal{H}_\bullet$ separately and reject $\mathcal{H}_0$ iff we reject each $\mathcal{H}_\bullet$

Test eg $\mathcal{H}_X$ by permuting indices of X_i to estimate quantiles under $\mathcal{H}_X$

Note:

(1) $\mathcal{H}_0 \implies \Delta_L P = 0$, but $\impliedby$ is not true. ie there exist $\mathbb{P}_{XYZ}$ that do not factorise yet have $\Delta_L P = 0$. Thus if we fail to reject $\mathcal{H}_0$, we can conclude nothing.

(2) Control of Type I error rate is simple.

Let $A_\bullet = \{\mathcal{H}_\bullet \text{ is rejected}\}$

Then $\mathbb{P}(A_0) = \mathbb{P}(A_X \cap A_Y \cap A_Z) \leq \min\{\mathbb{P}(A_X), \mathbb{P}(A_Y), \mathbb{P}(A_Z)\}$

If $\mathcal{H}_0$ is true, then WLOG $\mathcal{H}_X$ is true. If we use a significance level of α for each $\mathcal{H}_\bullet$ then $\mathbb{P}(A_0) \leq \mathbb{P}(A_X) = \alpha$

(3) $\|\Delta_L \hat{P}\|^2$ can be expressed in terms of K, L and M but is most simply written as $\frac{1}{n^2}(\tilde{K} \circ \tilde{L} \circ \tilde{M})_{++}$

Advantages of Lancaster interaction

- Could instead use HSIC to test each of $\mathcal{H}_X$,$\mathcal{H}_Y$ and $\mathcal{H}_Z$ by testing eg $HSIC((X,Y),Z)$ ('3-way HSIC' in paper)
  - More sensitive to pairwise weak but jointly strong interactions
  - More sensitive in higher dimensions
- Similar when testing for total independence ($\mathbb{P}_{XYZ} = \mathbb{P}_X \mathbb{P}_Y \mathbb{P}_Z$)

See Figures 1 and 2, Sejdinovic et al (2013)

The non-iid case

For both HSIC and Lancaster, *iid* assumption is required for permutation bootstrap.

ie without *iid* data, cannot estimate quantiles of test statistic null distribution so cannot control Type I error. See Figure 2, Rubenstein et al (2016) and Figure 3, Chwialkowski & Gretton (2014)

Essential problem: $(X_{\pi(i)})_{i=1}^n$ has different temporal dependence to $(X_i)_{i=1}^n$

Wild Bootstrap

Previously, we fabricated a new dataset and evaluated our test statistic using it in order to estimate quantiles.

Wild bootstrap directly resamples test statistic under null $\mathcal{H}_0$ subject to conditions on:

(1) Data generating process

- Appropriate mixing conditions

(2) The form of the test statistic

- V-statistic with degenerate core

  ie $\frac{1}{n^2}\sum_{ij} h_(S_i, S_j)$ where $\mathbb{E}_S h(\cdot, S) = 0$

We will consider (2) in some detail but (1) only briefly. But first, we present the statistical test in full.

Lancaster + timeseries

In order to test $\mathcal{H}_0$, we test each of $\mathcal{H}_\bullet$ separately. This is the algorithm to test $\mathcal{H}_Z$

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

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

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

samples = zeros(1,N)

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

Draw random vector W according to $W_t = e^{-1/l_n} W_{t-1} + \sqrt{1 - e^{-2/l_n}}\epsilon_t$

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

end for

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

Reject $\mathcal{H}_Z$

else

Do not reject $\mathcal{H}_Z$

end if

Mixing conditions

Observations need to be drawn from a process that is:

- τ-mixing to use the Wild Bootstrap [Leucht & Neumann (2013)]
- β-mixing to use the Hilbert space CLT [Dehling et al (2013)], needed to show that test statistics satisfy wild bootstrap hypothesis
- Stationary

In practice it is not possible to verify that these conditions hold for a given dataset but do hold for eg Autoregressive processes.

V-statistics

A V-statistic of a 2-argument, symmetric function h given observations $(S_i)_{i=1}^n$ is

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

We call h the core of V and say h is degenerate if $\mathbb{E}_{S_2} h(s_1, S_2) = 0$ for all s_1. In this case we call nV_n a degenerate normalised V-statistic.

Examples:
Sample mean: $V = \frac{1}{n} \sum_i x_i$ (1 arg V-stat)

(Biased) sample variance: $V = \frac{1}{n^2} \sum_{ij} (x_i - x_j)^2$

$\frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++}$, $\frac{1}{n^2}(\tilde{K} \circ \tilde{L} \circ \tilde{M})_{++}$ can be expressed as n-argument V-statistics, but not as 2-argument.

The Wild Bootstrap

Theorem: (Leucht & Neumann 2013, informal)

Suppose that $nV = \frac{1}{n}\sum_{ij} h(S_i, S_j)$ is a normalised degenerate V-statistic, h is a kernel that is Lipschitz continuous, and that the $(S_i)_{i=1}^n$ is drawn from an appropriately mixing process.

Then an auxiliary process $(W_i)_{i=1}^n$ can be (randomly) drawn such that

$$nV^{(b)} = \frac{1}{n}\sum_{ij} W_i h(S_i, S_j) W_j$$

converges in distribution to nV as $n \longrightarrow \infty$

Note: nV converges in distribution to an infinite weighted sum of χ^2 random variables.

To use the wild bootstrap in our tests we need to show that under the null hypothesis, the conditions hold.

The Task

(1) $\frac{1}{n^2}(\tilde{K} \circ \tilde{L})_{++}$ and $\frac{1}{n^2}(\tilde{K} \circ \tilde{L} \circ \tilde{M})_{++}$ are n-arg V-statistics but can be expressed asymptotically as 2-arg V-statistics.

- Previous approach used the Hoeffding decomposition, expressing V_n as a sum of V_{n-1}, V_{n-2},... V_2. Then show that all other terms decay faster than V_2 term.

(2) Having expressed them as asymptotically 2-arg V-statistics, show that they are degenerate.

Our proof (very sketched outline)

The proof is rather simple in concept, but there is a large notational overhead.

For simplicity, we will stick to the derivation for HSIC, but the proof is essentially the same for Lancaster.

We define $\bar{k}$ to be the population centred kernel with feature map $\phi_X(x) - \mu_X$, and write $\bar{K}$ to be its Gram matrix.

By observing that

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

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

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

and similarly for $\tilde{L}$. We can thus write

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

Each of these terms can be interpreted as a term involving empirical estimates of the covariance operator and mean embeddings *with respect to* $\bar{k}$

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

$\bar{\mu}_X$ and $\bar{\mu}_Y$ are estimators of a quantity that is zero because we performed population centering of the feature maps.

$\bar{C}_{XY}$ estimates zero because we assume $\mathcal{H}_0$ to be true.

The rate of convergence for each is $O_p(n^{-1/2})$ by the Hilbert space CLT for random processes. Thus $\|\bar{\mu}_X \otimes \bar{\mu}_Y\|^2 = O_p(n^{-1})$. It follows that

$$nHSIC \longrightarrow n\|\bar{C}_{XY}\|^2 = \frac{1}{n}(\bar{K} \circ \bar{L})_{++}$$

Observe that $\frac{1}{n}(\bar{K} \circ \bar{L})_{++} = \frac{1}{n}\sum_{ij} \bar{k} \otimes \bar{l}((X_i, Y_i), (X_j, Y_j))$ and this is degenerate.

Summary: we recentre the feature maps of the kernels with respect to their population means, and expand to express them as inner products of covariance operators / mean embeddings, all but one of which decay to 0. The remaining term is a normalised degenerate V-statistic and so satisfies the hypothesis of the Wild Bootstrap.

The proof for the Lancaster statistic $\frac{1}{n}(\tilde{K} \circ \tilde{L} \circ \tilde{M})_{++}$ is very similar, but is more algebraically involved. See paper for details.

Results and Open Questions

See Experiments section from Rubenstein et al (2016).