A Kernel Test for Three-Variable Interactions with Random Processes

Abstract

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

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

1. Introduction

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

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

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

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

- The main contributions of this paper are the folllowing:

  – To show that the Lancaster test statistic is such a statistic

  – This is done using a new style of technique which in particular gives a significantly simpler proof that HSIC is also such a statistic (and thus simplifies the proofs used in [HSIC+time series])

  – To show that the multiple testing corrections used in [Lancaster] are too conservative, and therefore that we can improve test power by using a more relaxed correction.

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

2. Background

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

2.1. Kernel Mean Embedding (and HSIC?)

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

2.2. Lancaster

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

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

$\mathbb{P}_{XZ}\mathbb{P}_Y - \mathbb{P}_X\mathbb{P}_{YZ} + 2\mathbb{P}_X\mathbb{P}_Y\mathbb{P}_Z$. It can be shown that if any variable is independent of the other two, then $\Delta_L P = 0$. That is,

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

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

Given kernels k, l and m on $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$ respectively, the kernel $k \otimes l \otimes m$ defines a kernel on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Letting K, L and M denote the gram matrices, where for example $K_{ij} = k(X_i, X_j)$, we can write

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

2.3. Time series

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

2.3.1. τ-MIXING

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

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

where

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

2.3.2. β-MIXING

Definition 2. *A process $(X_t)_t$ is* β-mixing *(also known as* absolutely regular*) if $\beta(m) \longrightarrow 0$ as $m \longrightarrow \infty$, where*

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

where the second supremum is taken over all finite partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of the sample space such that $A_i \in \mathcal{A}_1^n$ 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 1. *Suppose that the process $(X_t, Y_t, Z_t)_t$ is β-mixing. Then any 'sub-process' is also β-mixing (for example $(X_t, Y_t)_t$ or $(X_t)_t$)*

2.4. V-statistics

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

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

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

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

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

2.5. Wild Bootstrap

2.6. Hilbert spaced random variable CLT

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

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

3. Lancaster Interaction for Random Processes

- Statement of Wild Bootstrap theorem (maybe in background though?)
- Proof that Lancaster satisfies WB theorem hypothesis
- ...
- Multiple testing correction (maybe in next section though?)

4. p-values for Lancaster test

- In [Lancaster], they use the Holm-Bonferroni correction. Show here that this isn't actually necessary - that the 'naive' correction works and is therefore more powerful as we use $[\alpha, \alpha, \alpha]$ as the thresholds rather than $[\alpha/3, \alpha/2, \alpha]$ or whatever.

5. Experiments

5.1. Artificial data

5.2. Real data

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

6. Proofs

Proof of Lemma 1:

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

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

Thus

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

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

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

Acknowledgments

cheers!