Supplementary materials: Post Selection Inference with Kernels

Makoto Yamada[1,2,4], Yuta Umezu[3], Kenji Fukumizu[1,4] , Ichiro Takeuchi[1,3]
[1]RIKEN AIP, [2]JST PRESTO, [3]Nagoya Institute of Technology,
[4] Institute of Statistical Mathematics

Variance of bagging block HSIC

We will derive the variance of bagging block HSIC, and relate it with that of the full U-statistics and (unbagged) single block HSIC.

Let $Z_i = (X_i, Y_i)$ $(i = 1, 2, \ldots, n)$ be i.i.d. samples. HSIC can be expressed in the form of U-statistics of 4th degree:

$$U_n = \frac{1}{\binom{n}{4}} \sum_{S \in \mathfrak{S}_{n,4}} h(Z_S),$$

where $h(z_1, z_2, z_3, z_4)$ is the U-statistic kernel corresponding to HSIC (see [1]), $\mathfrak{S}_{n,k}$ is the set of all k-tuples (in this case $k = 4$) of $\{1, \ldots, n\}$, and Z_S is an abbreviation of $(Z_{i_1}, \ldots, Z_{i_k})$ for $S = (i_1, \ldots, i_k)$.

Consider the block HSIC with block size B. For simplicity, let $M := n/B$ denote the number of blocks for a single block HSIC estimator, and assume that n is taken so that M in an integer. The block HSIC is then defined by

$$W_B = \frac{1}{M} \sum_{b=1}^{M} U_B^{(b)}, \tag{1}$$

where $U_B^{(b)}$ is the U-statistics corresponding to the empirical HSIC computed from only the B samples in the b-the block, namely,

$$U_B^{(b)} = \frac{1}{\binom{B}{4}} \sum_{S} h(Z_S),$$

where the sum is taken for all the quadruplets from b-th block. Note that W_B converges in law to a normal distribution as $n \to \infty$ with B fixed, since $U_B^{(b)}$ $(b = 1, \ldots, M)$ are i.i.d. samples.

Recall that the bagging block HSIC with L random permutations is defined by

$$\xi_{L,B} := \frac{1}{L} \sum_{\ell=1}^{L} W_{\ell,B}, \tag{2}$$

where $W_{\ell,B}$ is defined similarly to W_B in Eq. (1), but with a random permutation of $Z_1, \ldots, Z_n$. We generate L independent uniform random permutations of $\{1, \ldots, n\}$, and make copies of W_B. Note that, by the independence of the random permutations, given $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$, $W_{\ell,B}$ and $W_{\ell',B}$ are independent for $\ell \neq \ell'$, but can be dependent unconditionally. The bagging block HSIC is simply the average over these L copies.

We rewrite $\xi_{L,B}$ with the indicator of index. Let $\mathfrak{I}_{\ell,b}$ be the index set of the b-th block in the ℓ-th permutation. For an arbitrary quadruplet $S = (i_1, i_2, i_3, i_4)$, define $\theta_{\ell,b}(S)$ by

$$\theta_{\ell,b}(S) = \begin{cases} 1 & \text{if } S \in \mathfrak{I}_{\ell,b} \\ 0 & \text{otherwise.} \end{cases}$$

Then, we have

$$\xi_{L,B} = \frac{1}{LM\binom{B}{4}} \sum_{\ell=1}^{L}\sum_{b=1}^{M}\sum_{S\in\mathfrak{S}_{n,4}} \theta_{\ell,b}(S)h(Z_S). \tag{3}$$

Since $E[h(S)] = \mathrm{HSIC}(X,Y)$, it is obvious that

$$E[\xi_{L,B}] = \mathrm{HSIC}(X,Y),$$

and thus $\xi_{L,B}$ is an unbiased estimator of HSIC.

It is not difficult to see that, as estimators of HSIC, the standard U-statistics U_n has less variance than the block HSIC $\xi_{1,B}$, which is regarded as a special type of incomplete U-statistic. The next proposition asserts that the variance of $\xi_{L,B}$, under the assumption of independence $X \perp\!\!\!\perp Y$, interpolates the variances of these two estimators.

Proposition 1 *Assume X_i and Y_i are independent. Then, we have*

$$\mathrm{Var}[\xi_{L,B}] = \Big(1-\frac{1}{L}\Big)\mathrm{Var}[U_n] + \frac{1}{L}\mathrm{Var}[\xi_{1,B}].$$

(Proof) For notational simplicity, we use h_S for $h(Z_S)$. It follows from the expression Eq. (3) that

$$\begin{aligned}
\mathrm{Var}[\xi_{L,B}] &= \frac{1}{(LM\binom{B}{4})^2}\sum_{\ell=1}^{L}\sum_{b=1}^{M}\sum_{\ell'=1}^{L}\sum_{b'=1}^{M}\sum_{S}\sum_{T} E[\theta_{\ell,b}(S)\theta_{\ell',b'}(T)h_S h_T] \\
&= \frac{1}{(LM\binom{B}{4})^2}\sum_{\ell=1}^{L}\sum_{\ell'\neq\ell}\sum_{b=1}^{M}\sum_{b'=1}^{M}\sum_{S}\sum_{T} E[\theta_{\ell,b}(S)\theta_{\ell',b'}(T)h_S h_T] \\
&\quad + \frac{1}{(LM\binom{B}{4})^2}\sum_{\ell=1}^{L}\sum_{b=1}^{M}\sum_{b'=1}^{M}\sum_{S}\sum_{T} E[\theta_{\ell,b}(S)\theta_{\ell,b'}(T)h_S h_T] \\
&=: I + II.
\end{aligned}$$

Let p be the probability that $\theta_{\ell,b}(S)$ takes 1, i.e., $p = P(\theta_{\ell,b}(S) = 1)$. We have

$$p = \frac{\binom{n-4}{B-4}}{\binom{n}{4}} = \frac{\binom{B}{4}}{\binom{n}{4}}.$$

This can be confirmed as follows. p is the probability that S is included in a B-tuple of $\{1,\ldots,n\}$ when we uniformly take it. This is equal to the proportion of B-tuples including S among all the B-tuples, and thus the first equality. The second equality is simple computation.

From the independence of permutations, $\theta_{\ell,b}(S)$ and $\theta_{\ell',b'}(T)$ are independent if $\ell \neq \ell'$. Since h_S is a constant given $\mathbf{Z}_n$, we have

$$\begin{aligned}
I &= \frac{1}{(LM\binom{B}{4})^2} \sum_{(\ell,\ell'):\ell\neq\ell'} \sum_{S,T} \sum_{b,b'} E\big[E[\theta_{\ell,b}(S)\theta_{\ell',b'}(T) \mid \mathbf{Z}_n]E[h_S h_T \mid \mathbf{Z}_n]\big] \\
&= \frac{1}{(LM\binom{B}{4})^2} \sum_{(\ell,\ell'):\ell\neq\ell'} \sum_{S,T} \sum_{b,b'} p^2 E[h_S h_T] \\
&= \frac{1}{(LM\binom{B}{4})^2} L(L-1)M^2 \frac{\binom{B}{4}^2}{\binom{n}{4}^2} \sum_{S,T} E[h_S h_T] \\
&= \frac{L-1}{L} \frac{1}{\binom{n}{4}^2} \sum_{S,T} E[h_S h_T] \\
&= \frac{L-1}{L} E\Big[\Big(\frac{1}{\binom{n}{4}} \sum_{S\in\mathfrak{S}_{n,4}} h_S\Big)^2\Big] \\
&= \frac{L-1}{L} \mathrm{Var}[U_n].
\end{aligned}$$

The second term is given by

$$\begin{aligned}
II &= \frac{1}{(LM\binom{B}{4})^2} \sum_{\ell=1}^{L} \sum_{b,b'} \sum_{S,T} E[\theta_{\ell,b}(S)\theta_{\ell,b'}(T) h_S h_T] \\
&= \frac{1}{L^2} \sum_{\ell=1}^{L} E\Big[\Big(\frac{1}{M} \sum_{b} \frac{1}{\binom{B}{4}^2} \sum_{S} \theta_{\ell,b}(S) h_S\Big)^2\Big].
\end{aligned}$$

In the last line, the value in the squared bracket is exactly the same as a single block HSIC for the ℓ-th sequence. Therefore,

$$II = \frac{1}{L} \mathrm{Var}[\xi_{1,B}].$$

This completes the proof.

False positive rate control

To check whether the methods can properly control the desired FPR, we run `hsicInf` using a dataset that has no relationship between input and output. Specifically, we generated the input output pairs as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{I}_{20})$, $\mathbf{0} \in \mathbb{R}$ is the vector whose elements are all zero, $\boldsymbol{I} \in \mathbb{R}^{20\times 20}$ is the identity matrix, and $y \sim N(0,1)$.

Figure 1 shows the FPRs of `hsicInf`, `hsic`, and `split` algorithms. The both the proposed method and `split` successfully control FPR, while `hsic` fails to control FPR. We see that the adjustment of the sampling distribution is crucial for estimating proper p-values. It shows that all FPRs tend to be high when the number of samples are small, and gradually converging to the significance level when the number of samples increases. Since `hsic` cannot control the FPR at the desired level, we do not compare the TPR of `hsic` in the following section.

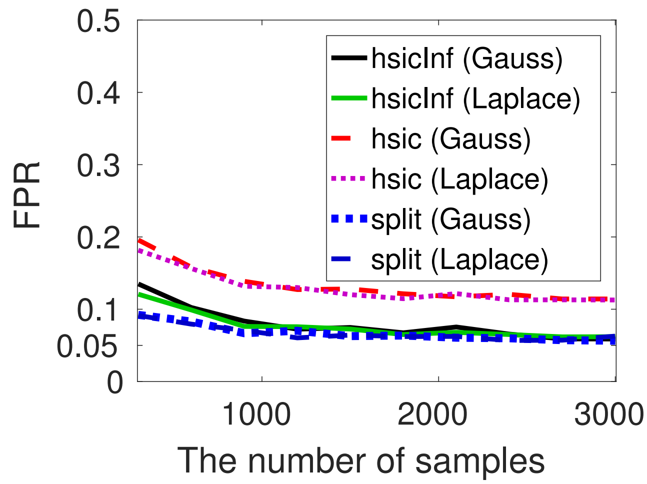


Figure 1: False positive rates at significant level $\alpha = 0.05$ of the proposed methods. Comparison of `hsicInf`, `hsic`, `split`, and `larInf`. We used $B = 10$ and $L = 1$ for the HSIC based approaches. The `hsic` computes p-values without adjusting the sampling distribution by Theorem 1.

Classification data

$$
\begin{aligned}
p(\boldsymbol{x}^{(1,2)}|y=1) &= N\left(\begin{bmatrix} -3 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\
p(\boldsymbol{x}^{(1,2)}|y=2) &= N\left(\begin{bmatrix} 3 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\
p(\boldsymbol{x}^{(1,2)}|y=3) &= 0.5N\left(\begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2.25 \end{bmatrix}\right) + 0.5N\left(\begin{bmatrix} 0 \\ -3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 2.25 \end{bmatrix}\right).
\end{aligned}
$$

Then, we generated the final feature $\boldsymbol{x} = [(\boldsymbol{x}^{(1,2)})^\top \ \widetilde{\boldsymbol{x}}^\top]^\top$ where $\widetilde{\boldsymbol{x}} \in \mathbb{R}^{18}$ and $\widetilde{\boldsymbol{x}} \sim N(\boldsymbol{0}, \boldsymbol{I})$.

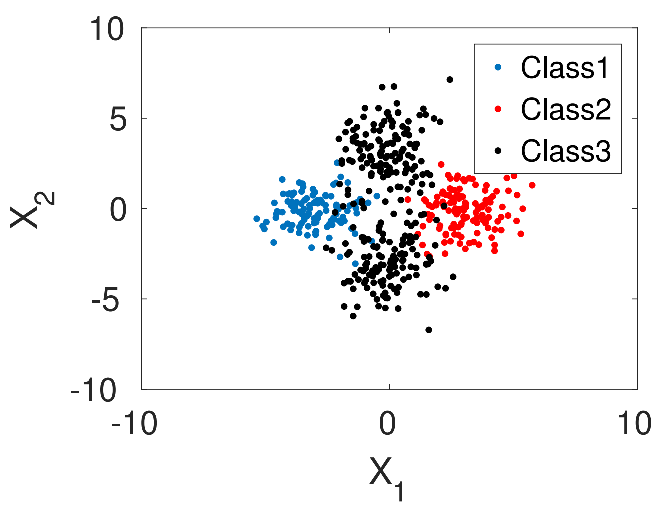


Figure 2: The multi-class classification dataset.

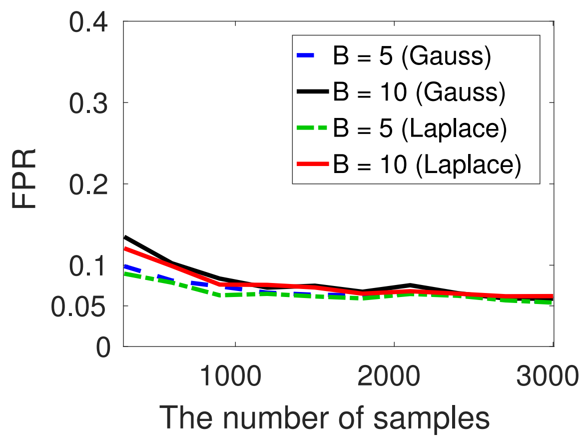
Block parameter comparison $L = 1$

For this experiment, we first generated the input matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ where $\boldsymbol{x} \sim N(\mathbf{0}, \bar{\boldsymbol{\Sigma}})$, $[\bar{\boldsymbol{\Sigma}}]_{ij} = 0.95\delta_{ij} + 0.05, i, j \in \{1, 2, 3, 4, 5\}$, $[\bar{\boldsymbol{\Sigma}}]_{ii} = \delta_{ij}, i, j \in \{6, \ldots, d\}$, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise, and $d = \{20, 500\}$ and $n = \{300, 600, \ldots, 3000\}$.

Then, we generated the corresponding output variable as

- **Linear:** $Y = \sum_{i=1}^{5} X_i + 0.1E$,
- **Additive Non-linear:** $Y = \sum_{i=1}^{5} X_i^2 + 0.1E$,
- **Non-additive Non-linear:**
  $Y = X_1 \exp(X_2) X_3 \exp(X_4) X_5 + 0.1E$,

where $E \sim N(0, 1)$ is an independent random noise.

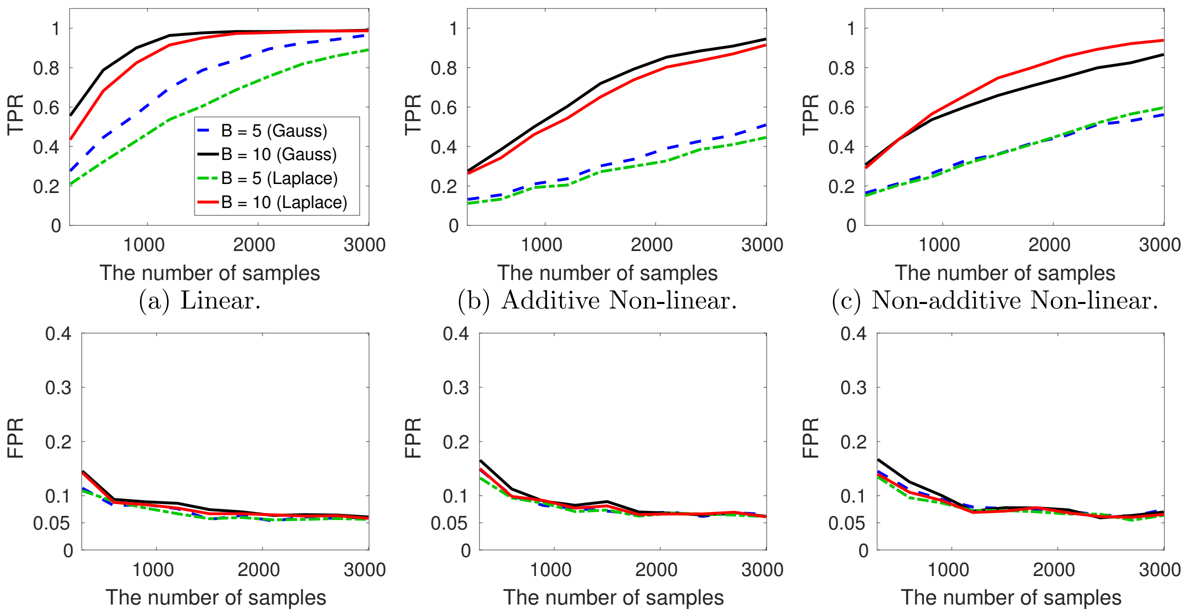


(a)

Figure 3: False positive rates at significant level $\alpha = 0.05$ of the proposed methods. FPRs for `hsicInf` with different block parameter B.

References

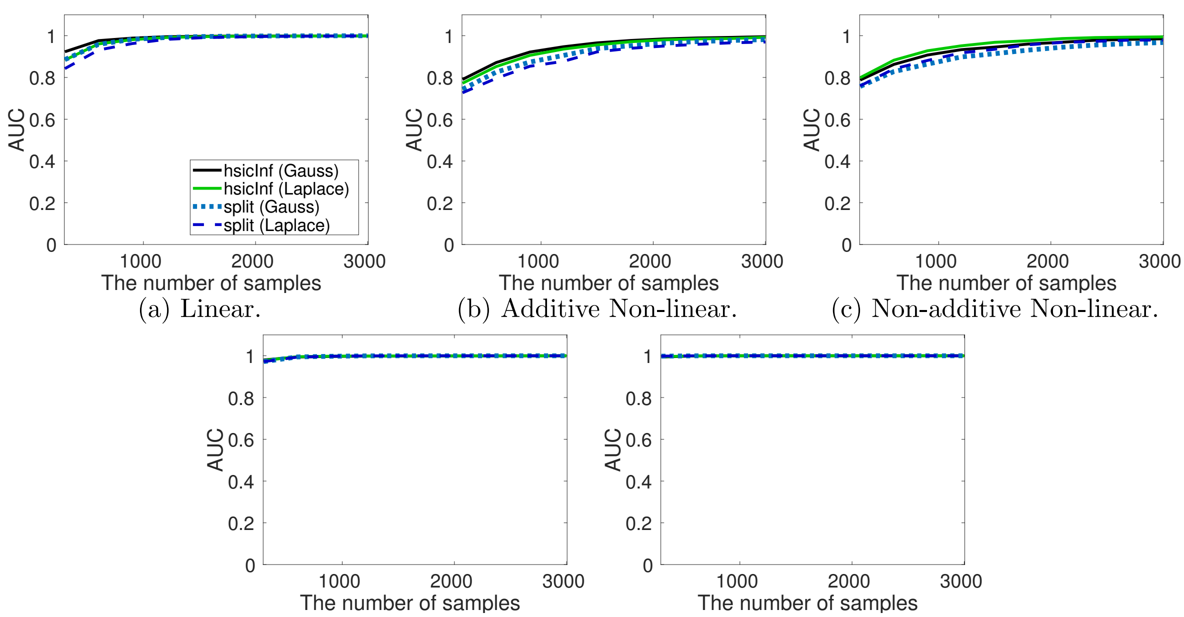
[1] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.



(a) Linear. (b) Additive Non-linear. (c) Non-additive Non-linear.

(d) Linear. (e) Additive Non-linear. (f) Non-additive Non-linear.

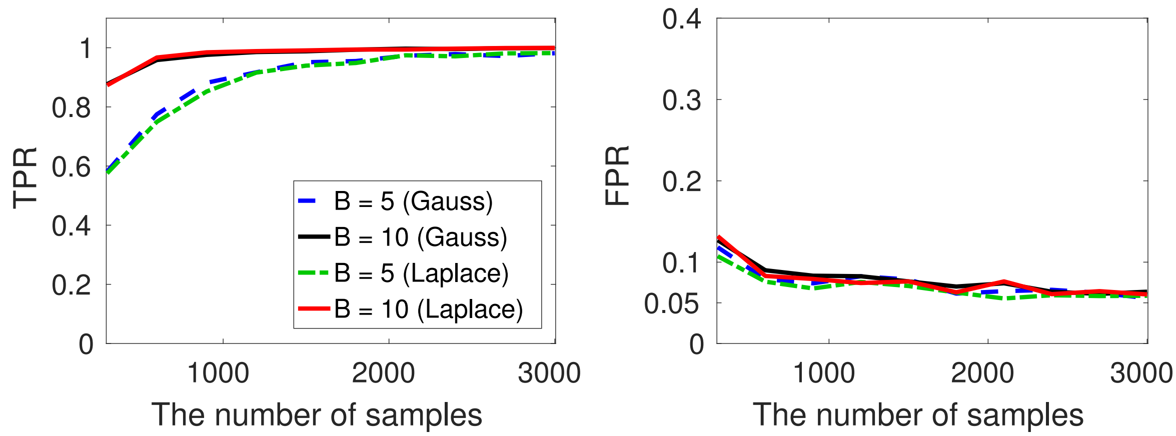
Figure 4: The results for `hsicInf` in uni-variate setups with different block parameter B. (a)-(c): TPR for the three datasets. (d)-(f): FPR for the three datasets.



(a) Linear. (b) Additive Non-linear. (c) Non-additive Non-linear.

(d) Multi-variate. (e) Multi-class.

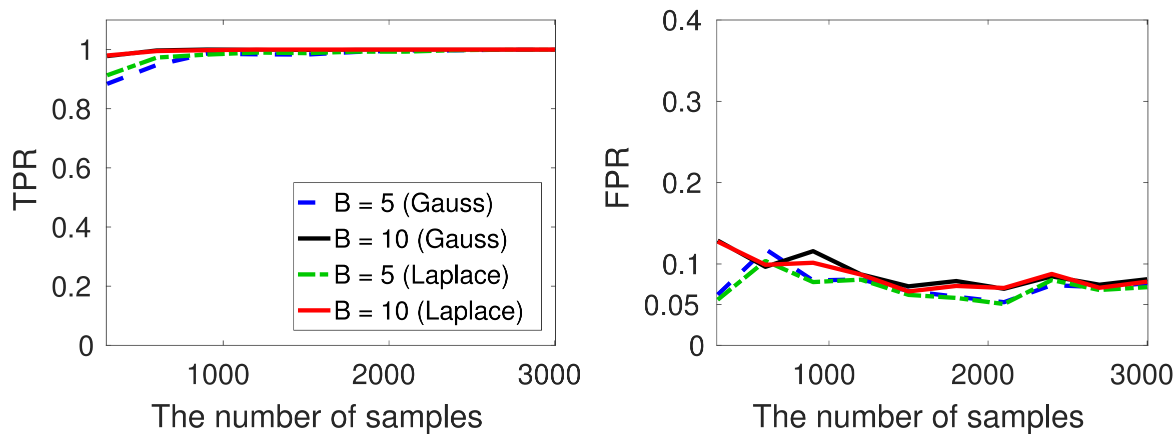
Figure 5: The average AUC scores.



(a) `hsicInf` (TPR).

(b) `hsicInf` (FPR).

Figure 6: The results for the multi-variate regression dataset. TPRs and FPRs of `hsicInf` with different block size B.



(a) `hsicInf` (TPR).

(b) `hsicInf` (FPR).

Figure 7: The results for the multi-class classification dataset. TPRs and FPRs of `hsicInf` with different block size B.