More Powerful Selective Kernel Tests for Feature Selection

Supplementary

A TRUE POSITIVE RATE (TPR) AND FALSE POSITIVE RATE (FPR)

Let $\mathcal{I}_-$ be the indices of features such that the null holds, i.e., for MMD, we have $\mathcal{I}_- := \{i : \mathrm{MMD}(P^{(i)}, Q^{(i)}) = 0\}$ (and for HSIC, we have $\mathcal{I}_- := \{i : \mathrm{HSIC}(P^{(i)}, Q) = 0\}$). Similarly, let $\mathcal{I}_+$ be the indices of features such that the alternative holds, i.e., for MMD, we have $\mathcal{I}_+ := \{i : \mathrm{MMD}(P^{(i)}, Q^{(i)}) > 0\}$ (and for HSIC, we have $\mathcal{I}_+ := \{i : \mathrm{HSIC}(P^{(i)}, Q) > 0\}$). Then, for a set of selected features $\mathcal{S}_k$ we define FPR and TPR as follows,

$$\mathrm{FPR} = \mathbb{E}\left[\frac{|\mathcal{S}_k \cap \mathcal{I}_- \cap \mathcal{R}|}{|\mathcal{S}_k \cap \mathcal{I}_-|}\right], \qquad \mathrm{TPR} = \mathbb{E}\left[\frac{|\mathcal{S}_k \cap \mathcal{I}_+ \cap \mathcal{R}|}{|\mathcal{S}_k \cap \mathcal{I}_+|}\right],$$

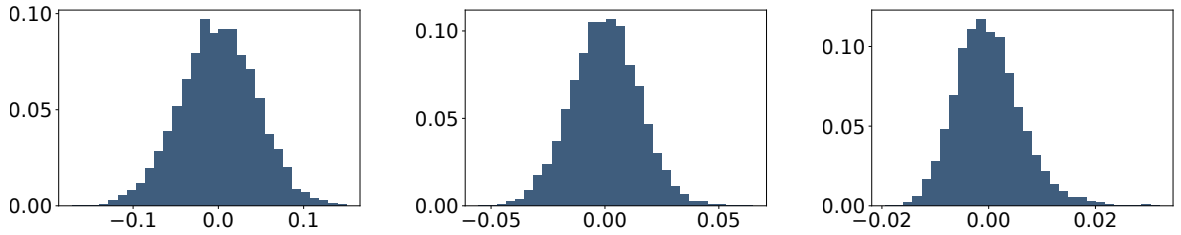
where $\mathcal{R}$ is the set of indices that the algorithm rejections and note that $\mathcal{R} \subseteq \mathcal{S}_k$.

B EMPIRICAL DISTRIBUTIONS OF $\widehat{\mathrm{MMD}}_{\mathbf{Inc}}(X, Y)$ and $\widehat{\mathrm{HSIC}}_{\mathbf{Inc}}(Z)$

In this section, we simulate the empirical distribution of the incomplete estimator for both $\widehat{\mathrm{MMD}}_{\mathrm{Inc}}(X, Y)$ and $\widehat{\mathrm{HSIC}}_{\mathrm{Inc}}(Z)$.

B.1 Empirical distribution of $\widehat{\mathrm{MMD}}_{\mathbf{Inc}}(X, Y)$

Case $P = Q$: For MMD, we let $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$ which means $\widehat{\mathrm{MMD}}_{\mathrm{u}}(X, Y)$ is degenerate whereas we show that $\widehat{\mathrm{MMD}}_{\mathrm{Inc}}(X, Y)$ follows a normal distribution (see Figure 4). When the r is small, the empirical distribution of the incomplete estimators follows a normal distribution butas r gets bigger we expect it to behave like its complete estimator counterpart.



(a) $r = 1$ (b) $r = 10$ (c) $r = 100$

Figure 4: The empirical distribution $\widehat{\mathrm{MMD}}_{\mathrm{Inc}}(X, Y)$ for $r \in \{1, 10, 100\}$. 5000 samples were used.

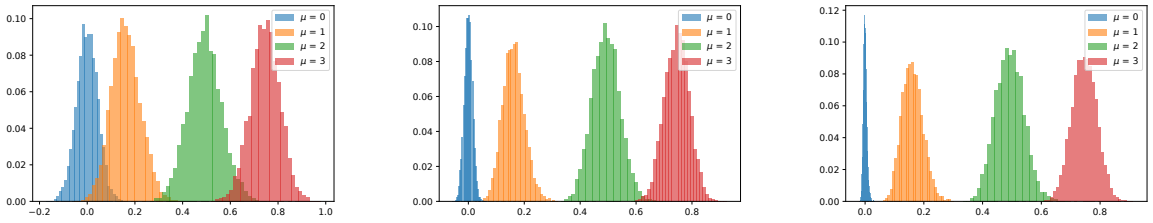
Case $P \neq Q$: We show the empirical distribution of the incomplete estimator for MMD when $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(\mu, 1)$ and $\mu \in \{0, 2, 3\}$. Under the alternative, for our choice in r, the distribution under the alternative is expected to have higher variance than the null distribution.

B.2 Empirical distribution of $\widehat{\mathrm{HSIC}}_{\mathbf{Inc}}(Z)$

For HSIC, let $Z := (X, Y)$ where X and Y is follows a standard normal and is sampled independently of each other. We show that in this case $\widehat{\mathrm{HSIC}}_{\mathrm{Inc}}(Z)$ is also normal (see Figure 6).

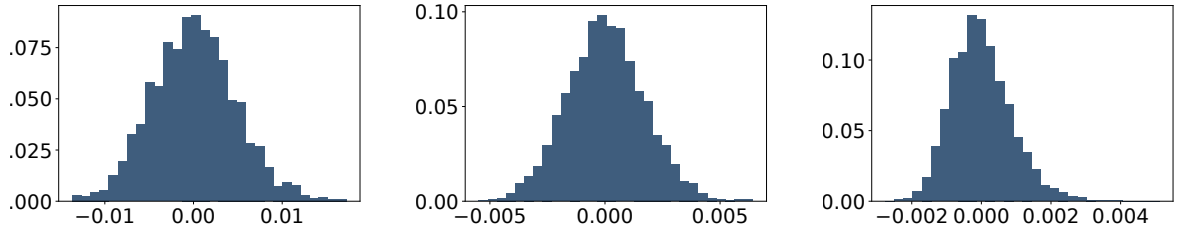
C MULTISCALE BOOSTRAP ALGORITHM FOR HSIC

In this section, we present algorithms for MultiHSIC for incomplete HSIC (Section C.1) and for block HSIC (Section C.2). Algorithm 3 describes the procedure for calculating p-values using multiscale bootstrap.



(a) $r = 1$ (b) $r = 10$ (c) $r = 100$

Figure 5: The empirical distribution $\widehat{\mathrm{MMD}}_{\mathrm{Inc}}(Z)$ for $r \in \{1, 10, 100\}$. 5000 samples were used.



(a) $r = 1$ (b) $r = 10$ (c) $r = 100$

Figure 6: The empirical distribution $\widehat{\mathrm{HSIC}}_{\mathrm{Inc}}(X, Y)$ for $r \in \{1, 10, 100\}$. 5000 samples were used.

C.1 Incomplete HSIC

The parameters $\boldsymbol{T}(\boldsymbol{Z})$ and $\boldsymbol{\Sigma}$ for the incomplete estimator are estimated with the same method as for the incomplete MMD (see Section 3). The algorithm is described in Algorithm 3.

Algorithm 3 MultiHSIC($\boldsymbol{Z}_n, k, \mathcal{M}$): Selective p-values for the null hypothesis $H_{0,i} : \mathrm{HSIC}(P_{\boldsymbol{x}^{(i)}\boldsymbol{y}}) = 0 \mid i \in \mathcal{S}_k$ is selected.

$\hat{\boldsymbol{T}}(\boldsymbol{Z}), \hat{\boldsymbol{\Sigma}} \leftarrow$EstimateParam($\boldsymbol{Z}_n$)
$\mathcal{S}_k \leftarrow$ the indexes of k largest values of $\{\widehat{\mathrm{HSIC}}(\boldsymbol{Z}_n^{(i)})\}_{i \in \mathcal{I}}$
for $i \in \mathcal{S}_k$ **do**
  for $n' \in \mathcal{M}$ **do**
    $\gamma^2_{n'} \leftarrow \frac{n}{n'}$
    Sample $\{\boldsymbol{y}_i^*\}_{i=1}^B \overset{i.i.d.}{\sim} \mathcal{N}(\hat{\boldsymbol{T}}(\boldsymbol{Z}), \gamma^2_{n'}\hat{\boldsymbol{\Sigma}})$
    $\mathrm{BP}_{\gamma^2_{n'}}(S) \leftarrow \sum_{i=1}^B \mathbb{1}_S^{(i)}(\boldsymbol{y}_i^*)/B$
  end for
  Fit a linear model $\varphi_S(\gamma^2)$ such that $\varphi_S(\gamma^2) = \gamma\bar{\Phi}^{-1}(\mathrm{BP}_{\gamma^2}(S))$.
  $\hat{\beta}_0^{(i)} \leftarrow \hat{\sigma}_i^{-1}\sqrt{l_n}\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}_n^{(i)})$
  $p_i \leftarrow \bar{\Phi}(\hat{\beta}_0^{(i)})/\bar{\Phi}(\hat{\beta}_0^{(i)} + \varphi_S(0))$
end for
return $\{p_i\}_{i=0}^k$ and $\mathcal{S}_k$

The following theorem justifies our use of the multivariate normal model,

Theorem 1. *Assume that* $\lim_{n,l\to\infty} n^{-1}l = \lambda$ *and assume that* $\lim_{n,l\to\infty} n^{-2}l = 0$ *and* $0 < \lambda < \infty$ *then,*

$$l^{\frac{1}{2}}\left(\begin{bmatrix}\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}^{(1)})\\ \vdots \\ \widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}^{(d)})\end{bmatrix} - \begin{bmatrix}\mathrm{HSIC}(P_{\boldsymbol{x}^{(1)}\boldsymbol{y}})\\ \vdots \\ \mathrm{HSIC}(P_{\boldsymbol{x}^{(d)}\boldsymbol{y}})\end{bmatrix}\right) \text{ is asymptotically normal.}$$

The proof can be found in Appendix D.

C.2 Block HSIC

Block estimator as the incomplete estimator: The block estimator $\widehat{\mathrm{HSIC}}_{Blo}$ (Zhang et al., 2018) is an example of an incomplete estimator for HSIC with a fixed design matrix. To see this note that for a given blocksize B, we have a total of $\frac{n}{B}$ blocks. For each block, the complete U-statistic estimator is calculated, i.e., for block t

$$\hat{\eta}(t) = \frac{(B-4)!}{B!} \sum_{(i,j,q,r)\in \mathbf{i}_4^{[(t-1)B+1,tB]}} h(i,j,q,r),$$

where $\mathbf{i}_4^{[u,i]}$ is the set of 4-tuple with each index, between u and i, appearing exactly once. There are a total of $\frac{n}{B}$ blocks that are averaged to produce $\widehat{\mathrm{HSIC}}_{Blo}$, i.e., we have

$$\widehat{\mathrm{HSIC}}_{Blo} = \frac{B}{n} \sum_{t=1}^{\frac{n}{B}} \hat{\eta}(t).$$

Thus, we have shown that $\widehat{\mathrm{HSIC}}_{Blo}$ can be rewritten as $\widehat{\mathrm{HSIC}}_{Inc}$ where we have $\mathcal{D} = \cup_{t=1}^{\frac{n}{B}} \mathbf{i}_4^{[(t-1)B+1,tB]}$. Note that $|\mathcal{D}_{Blo}| = \frac{(B-1)!}{(B-4)!} n$.

Algorithm: The extension to multiscale bootstrap to include the block estimator is simple. It only requires changes in the parameters of the resampling distribution for varying n', as a well as how the signed distance $\hat{\beta}_0^{(i)}$ for feature i is calculated.

Let $\hat{\boldsymbol{T}}(\boldsymbol{Z}) := \sqrt{\frac{n}{B}}[\widehat{\mathrm{HSIC}}_{Blo}(\boldsymbol{Z}_n^{(1)}), \ldots, \widehat{\mathrm{HSIC}}_{Blo}(\boldsymbol{Z}_n^{(d)})]^\top$ and $\boldsymbol{T}(\boldsymbol{Z})$ be its population counterpart, namely, $\boldsymbol{T}(\boldsymbol{Z}) = \sqrt{\frac{n}{B}}[\mathrm{HSIC}(P_{\boldsymbol{x}^{(1)}\boldsymbol{y}}), \ldots, \mathrm{HSIC}(P_{\boldsymbol{x}^{(d)}\boldsymbol{y}})]^\top$. Note that $\hat{\boldsymbol{T}}(\boldsymbol{Z})$ can be equivalently written as $\sum_{i=1}^{n/B} \hat{\boldsymbol{\eta}}(i)$ where $\hat{\boldsymbol{\eta}}(i) = [\hat{\eta}^{(1)}(i), \ldots, \hat{\eta}^{(d)}(i)]^\top$, and $\hat{\eta}^{(j)}(i)$ is the complete U-statistic estimator for HSIC applied to the i-th block of $\boldsymbol{Z}^{(j)}$. Then in the limit $n \to \infty$, $B \to \infty$, and $\frac{n}{B} \to \infty$ (Zhang et al., 2018), we have under the null hypothesis

$$\hat{\boldsymbol{T}}(\boldsymbol{Z}) - \boldsymbol{T}(\boldsymbol{Z}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix with its elements as $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}\left(\boldsymbol{\eta}^{(i)}, \boldsymbol{\eta}^{(j)}\right)$. We estimate $\boldsymbol{\Sigma}$ with the sample covariance $\hat{\boldsymbol{\Sigma}}$, i.e., we have $\hat{\boldsymbol{\Sigma}} := \frac{B}{n} \sum_{i=1}^{n/B} [\hat{\boldsymbol{\eta}}(i) - \overline{\boldsymbol{\eta}}][\hat{\boldsymbol{\eta}}(i) - \overline{\boldsymbol{\eta}}]^\top$. Then for varying n', instead of resampling n' samples from $\boldsymbol{Z}$, we produce samples directly from $\mathcal{N}(\hat{\boldsymbol{T}}(\boldsymbol{Z}), \frac{n}{n'}\hat{\boldsymbol{\Sigma}})$ as before. The sign distance $\hat{\beta}_0^{(i)}$ is $\hat{\sigma}_i^{-1} \sqrt{\frac{n}{B}} \widehat{\mathrm{HSIC}}_{Blo}(\boldsymbol{Z}_n^{(i)})$ where $\hat{\sigma}_i^{-1}$ is the i-th diagonal element of $\hat{\boldsymbol{\Sigma}}$.

Empirical Results: In this experiment, we use the same setup as Figure 3 for the Logit problem and the results are shown in Figure 7. Our aim is to investigate the behaviour of our test when B the block size increases. In Zaremba et al. (2013, Section 5), they investigated the behaviour of the block estimator under finite samples and found that there can have severe bias under the null hypothesis.

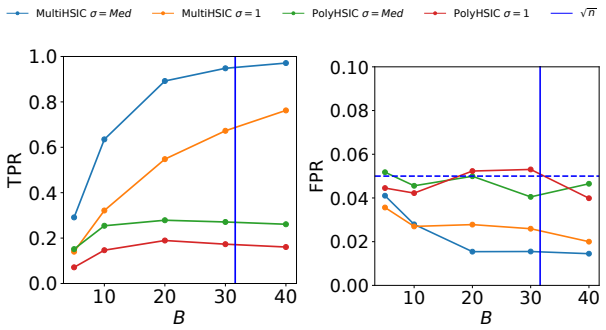


Figure 7: Logistic experiment. B increases for $\widehat{\mathrm{HSIC}}_B$. We use a Gaussian kernel with its bandwidth either set to be 1 or chosen with the median heuristic. We use $n = 1000$.

In our results, we observed that there was a large deviation for the nominal size α and an increase in the TPR. We speculate that this is due to the positive bias in finite samples of the skewness of the block estimator. These experiments show that the effect is more pronounced for MultiHSIC (than PolyHSIC) which may be because of our choice in parameterising the bootstrap samples as a normal distribution. We note that the effect of FPR going below the nominal α is not just for very large values of B but even for the recommended heuristic $B = \sqrt{n}$. It would be interesting to investigate this problem and correct for it in future works.

D PROOFS

In this section, we provide proofs for our statements in Section 4. Before we begin, recall that

$$h(i,j,q,r) = \frac{1}{4!}\sum_{(s,t,u,v)}^{(i,j,q,r)} \boldsymbol{K}_{st}[\boldsymbol{L}_{st} + \boldsymbol{L}_{uv} - 2\boldsymbol{L}_{su}]$$

is the order-4 U-statistic kernel for HSIC. We define the conditional expectation of the U-statistic kernel

$$\begin{aligned}
h_4 &= h(i,j,q,r),\\
h_3 &= \mathbb{E}[h(i,j,q,r)\,|\,i,\,j,\,q\,],\\
h_2 &= \mathbb{E}[h(i,j,q,r)\,|\,i,\,j\,]\\
h_1 &= \mathbb{E}[h(i,j,q,r)\,|\,i\,].
\end{aligned}$$

Let c be the smallest integer such that $h_c \neq \mathrm{HSIC}$. When $P \perp\!\!\!\perp Q$, we have $h_1 = 0$ and $\mathrm{HSIC} = h_1$ so $c > 1$. However when $P \not\perp\!\!\!\perp Q$, $h_1 \neq \mathrm{HSIC}$ so $c = 1$. Similarly, we show that $\widehat{\mathrm{HSIC}}_{Inc}$ is asymptotically normal under mild assumptions.

Theorem 2 (Asymptotic Distribution of $\widehat{\mathrm{HSIC}}_{Inc}$)**.** *Let c be the smallest integer such that $h_c \neq \mathrm{HSIC}$ (h_c defined in Appendix D) and let $\lim_{n,l\to\infty} n^{-c}l = \lambda$ $(0 \leq \lambda \leq \infty)$ and let $\mathcal{D}$ be constructed by selecting l subsets with replacement from $\boldsymbol{i}_4^n$ then,*

1. *If $\lambda = 0$ then, $l^{\frac{1}{2}}(\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{z}) - \mathrm{HSIC}(P_{xy}) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$,*
2. *If $0 < \lambda < \infty$ then, $l^{\frac{1}{2}}(\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{z}) - \mathrm{HSIC}(P_{xy})) \xrightarrow{d} \lambda^{\frac{1}{2}}V + T$,*
3. *If $\lambda = \infty$ then, $n^{\frac{c}{2}}(\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{z}) - \mathrm{HSIC}(P_{xy})) \xrightarrow{d} V$,*

where V is a random variable with the limit distribution of $n^{c/2}(\widehat{\mathrm{HSIC}}_u(\boldsymbol{z}) - \mathrm{HSIC})$ and $T \sim \mathcal{N}(0,\sigma^2)$ where $\sigma^2 = \mathrm{Var}[h(i,j,q,r)]$.

Proof. See Janson (1984, Corollary 1) and Lee (2019, Theorem 1, Section 4.3.3) □

Corollary 1 (Asymptotic Distribution of $\widehat{\mathrm{HSIC}}_{\mathrm{Inc}}$)**.** *Assume that $\lim_{n,l\to\infty} n^{-2}l = 0$ and $0 < \lim_{n,l\to\infty} n^{-1}l = \lambda < \infty$,*

- *If $X \perp\!\!\!\perp Y$, then $l^{\frac{1}{2}}\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{z}) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$,*
- *If $X \not\perp\!\!\!\perp Y$, then $l^{\frac{1}{2}}(\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{z}) - \mathrm{HSIC}(P_{xy})) \xrightarrow{d} \mathcal{N}(0,\lambda\sigma_u^2 + \sigma^2)$,*

where $\sigma^2 = \mathrm{Var}[h(i,j,q,r)]$ and σ_u^2 is the variance of the complete U-statistic counterpart, see Song et al. (2012, Theorem 5).

Proof. When $P \perp\!\!\!\perp Q$, then $c \geq 2$ then the result immediately follows from Theorem 2 for the case $\lambda = 0$.

For $P \not\perp\!\!\!\perp Q$, then $c = 1$ thus, under our assumptions, we obtain our result from Theorem 2. □

Theorem 1. *Assume that $\lim_{n,l\to\infty} n^{-1}l = \lambda$ and assume that $\lim_{n,l\to\infty} n^{-2}l = 0$ and $0 < \lambda < \infty$ then,* $l^{\frac{1}{2}}\left(\begin{bmatrix}\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}^{(1)})\\ \vdots \\ \widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}^{(d)})\end{bmatrix} - \begin{bmatrix}\mathrm{HSIC}(P_{\boldsymbol{x}^{(1)}\boldsymbol{y}})\\ \vdots \\ \mathrm{HSIC}(P_{\boldsymbol{x}^{(d)}\boldsymbol{y}})\end{bmatrix}\right)$ *is asymptotically normal.*

Proof. This proof is identical to the proof of Yamada et al. (2019, Theorem 5). From Cramér-Wold theorem, it is sufficient to prove that for every $\boldsymbol{\eta} \in \mathbb{R}^d$,

$$\boldsymbol{\eta}^\top \begin{bmatrix} \widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}^{(1)}) \\ \vdots \\ \widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}^{(d)}) \end{bmatrix} \xrightarrow{d} \boldsymbol{\eta}^\top \boldsymbol{V}$$

where $\boldsymbol{V}$ is some normal distribution. Under our assumptions, for all i $\widehat{\mathrm{HSIC}}_{Inc}(\boldsymbol{Z}^{(i)})$ follows a normal distribution. Following from the continuous mapping theorem, for all $\boldsymbol{\eta} \in \mathbb{R}^d$ we have as desired. □

E Additional Experiments

In this section, we provide additional experiments with HSIC. The first is a benchmarking experiment similar to the one performed in Section 5.2. The second uses the Divorce dataset (Yöntem et al., 2019) where people were given a questionnaire about their marriage and asked to rate each statement about their marriage from 0 to 4 depending on the truthfulness.

E.1 Benchmark

The goal is to rediscover the original features with statistical significance. As seen in the Table 3, the results indicate that MultiSel achieves higher power (as with the MMD).

	MultiSel-HSIC		PolySel-HSIC	
Dataset	TPR	FPR	TPR	FPR
Pulsar ($n = 100$)	0.705	0.023	0.625	0.025
Heart ($n = 138$)	0.469	0.029	0.410	0.030
Wine ($n = 200$)	0.800	0.042	0.730	0.058

Table 3: The TPR and FPR for the benchmarking experiment using $\widehat{\mathrm{HSIC}}_{Inc}$. The results are averaged over 100 trials, with $\alpha = 0.05$.

E.2 Divorce Dataset

We report the calculated p-values of each statistical test of dependency between a selected statement and the outcome of divorce. In the experiment, we chose $r = 15$, $k = 15$ (out of 54) and $n = 150$ with the results summaries in Table 4. We found that MultiSel declared 6 more statements as significantly (than PolySel) with a significance level at $\alpha = 0.05$, including statements such as "I feel aggressive when I argue with my wife." and "My wife and most of our goals are common.". We do not know the ground truth but the 6 statements seem plausible. The results suggest that MultiSel has higher detection rate.

	p-values	
	MultiSel-HSIC	PolySel-HSIC
My argument with my wife is not calm.	<0.01	<0.01
Fights often occur suddenly.	<0.01	0.41
I can insult my spouse during our discussions.	<0.01	0.09
When fighting with my spouse, I usually use expressions such as you always or you never.	<0.01	0.17
We're compatible with my wife about what love should be.	<0.01	0.43
My wife and most of our goals are common.	<0.01	0.25
I feel aggressive when I argue with my wife.	<0.01	0.22
We're starting a fight before I know what's going on.	<0.01	<0.01
I can use negative statements about my wife's personality during our discussions.	<0.01	0.05
I hate my wife's way of bringing it up.	<0.01	<0.01
I enjoy our holidays with my wife.	0.12	0.27
When we fight, I remind her of my wife's inadequate issues.	0.13	0.04
When I argue with my wife, it will eventually work for me to contact him.	0.16	0.14
I know my wife's hopes and wishes.	0.77	0.56
I can use offensive expressions during our discussions.	0.94	0.89

Table 4: The resultant p-values from one trial of the divorce dataset using HSIC_{Inc}.