Efficiently Learning Significant Fourier Feature Pairs for Statistical Independence Testing

Yixin Ren[1], Yewei Xia[1,4], Hao Zhang[3], Jihong Guan[2], Shuigeng Zhou[1]

[1]Fudan University, [2]Tongji University, [3]SIAT, [4]MBZUAI

Statistical Independence Testing

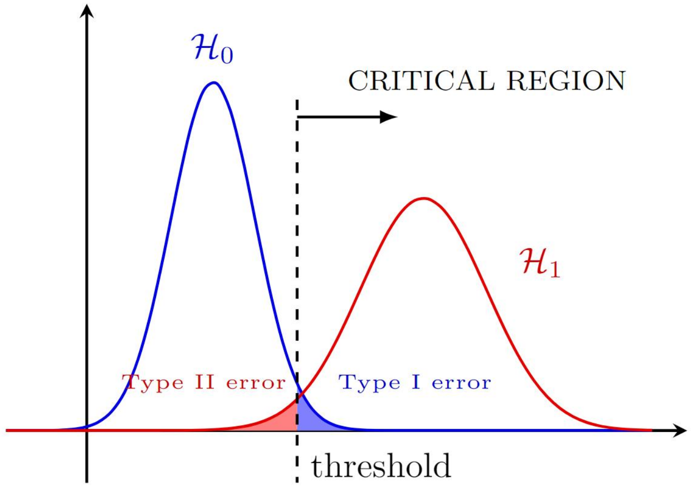
$$\text{whether } \mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y \ ?$$

Statistical Independence Testing

$$\text{whether } \mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y \text{ ?}$$

Hypothesis testing

- Null hypothesis $\mathcal{H}_0$: $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$.
- Alternative hypothesis $\mathcal{H}_1$: $\mathbb{P}_{XY} \neq \mathbb{P}_X \mathbb{P}_Y$.



Given n i.i.d samples $Z := \{(x_i, y_i)\}_{i=1}^n$

Kernel-based Statistical Independence Testing

Definition (Hilbert-Schmidt Independence Criterion)

Let $\mathcal{F}$ be an RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and let $\mathcal{G}$ be a second RKHS on $\mathcal{Y}$ with kernel $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, the HSIC between X and Y, denoted as $\mathrm{HSIC}(X, Y)$ is defined as

$$\mathbf{E}\left[k(X, X')l(Y, Y')\right] + \mathbf{E}\left[k(X, X')\right]\mathbf{E}\left[l(Y, Y')\right] - 2\mathbf{E}_{X'Y'}\left[\mathbf{E}_X k(X, X')\mathbf{E}_Y l(Y, Y')\right],$$

where (X', Y') is a independent copy of (X, Y).

Kernel-based Statistical Independence Testing

Definition (Hilbert-Schmidt Independence Criterion)

Let $\mathcal{F}$ be an RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and let $\mathcal{G}$ be a second RKHS on $\mathcal{Y}$ with kernel $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$, the HSIC between X and Y, denoted as $\mathrm{HSIC}(X, Y)$ is defined as

$$\mathbf{E}\left[k(X, X')l(Y, Y')\right] + \mathbf{E}\left[k(X, X')\right]\mathbf{E}\left[l(Y, Y')\right] - 2\mathbf{E}_{X'Y'}\left[\mathbf{E}_X k(X, X')\mathbf{E}_Y l(Y, Y')\right],$$

where (X', Y') is a independent copy of (X, Y).

- $\mathrm{HSIC}(X, Y) = 0 \Leftrightarrow \mathbb{P}_{XY} = \mathbb{P}_X\mathbb{P}_Y$ with suitable kernels (e.g. Gaussian kernel).

Examples (Gaussian kernel with width σ)

The Gaussian kernel is defined as $k(x, x') := \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$, where σ is the width.

Kernel-based Statistical Independence Testing

Definition (Estimation of HSIC)

An estimator of HSIC(X, Y) with sample Z is given by

$$\mathrm{HSIC}_b(Z) := \frac{1}{n^2}\sum_{i,j} k_{ij} l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r} k_{ij} l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q} k_{ij} l_{iq} = \frac{1}{n^2}\mathrm{Tr}(\mathbf{KHLH}),$$

where $k_{ij} := k(x_i, x_j)$, $l_{ij} := l(y_i, y_j)$ are the entries of the $n \times n$ kernel matrix $\mathbf{K}$, $\mathbf{L}$, respectively and $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the center matrix and $\mathbf{1}$ is a vector of ones.

$$\mathbf{K}_{n\times n} = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1n} \\ k_{21} & k_{22} & \cdots & k_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n1} & k_{n2} & \cdots & k_{nn} \end{bmatrix}$$

Calculate $\mathrm{HSIC}_b(Z)$ cost $\mathcal{O}(n^2)$ time and $\mathcal{O}(n^2)$ space.
n: sample size.

What's More?

$$\mathrm{HSIC}_b(Z) := \frac{1}{n^2}\sum_{i,j} k_{ij} l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r} k_{ij} l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q} k_{ij} l_{iq} = \frac{1}{n^2}\mathrm{Tr}(\mathbf{KHLH}),$$

- Faster: The time/space computational complexity of current statistic are both quadratic computing time.
- More flexible: The kernel can not be adaptive.

What's More?

$$\mathrm{HSIC}_b(Z) := \frac{1}{n^2}\sum_{i,j} k_{ij} l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r} k_{ij} l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q} k_{ij} l_{iq} = \frac{1}{n^2}\mathrm{Tr}(\mathbf{KHLH}),$$

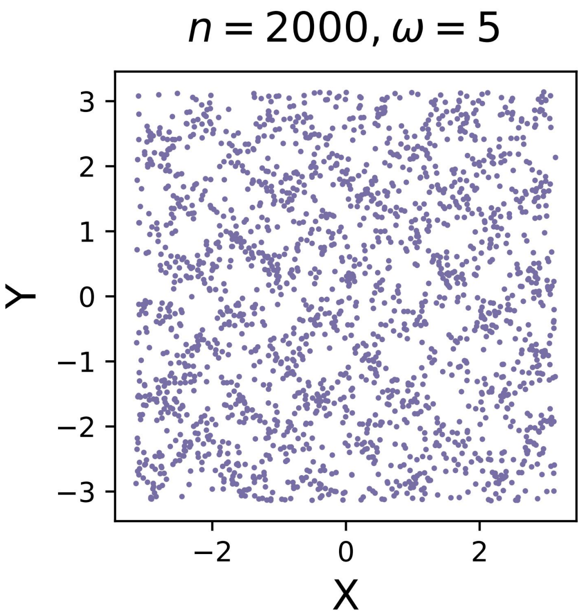
- Faster: The time/space computational complexity of current statistic are both quadratic computing time.

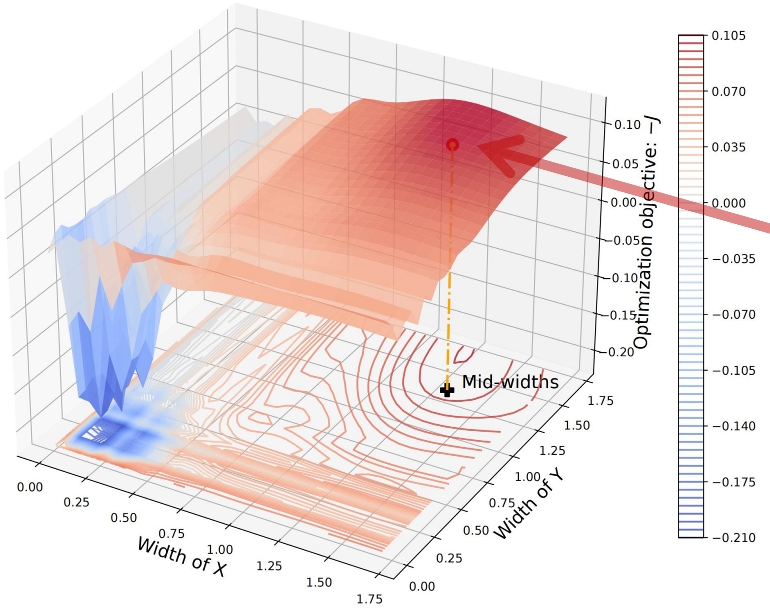
- More flexible: The kernel can not be adaptive.

A Frequency Domain Perspective

$$\mathrm{HSIC}(X, Y) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \left| \phi_{\mathbb{P}_{XY}}(\omega) - \phi_{\mathbb{P}_X \mathbb{P}_Y}(\omega) \right|^2 (\mathcal{F}^{-1}\psi)(\omega) d\omega,$$

Example: $(X, Y) \sim p_{xy}(x, y) \propto 1 + \sin(\omega_0 x) \sin(\omega_0 y)$.





Inflexible settings
result in
low test power

We aim to obtain a more flexible $(\mathcal{F}^{-1}\psi)(\omega)$

A Frequency Domain Perspective

Enable more efficient calculation of statistic

- The statistic

$$\mathrm{HSIC}(X, Y) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \left| \phi_{\mathbb{P}_{XY}}(\omega) - \phi_{\mathbb{P}_X \mathbb{P}_Y}(\omega) \right|^2 (\mathcal{F}^{-1}\psi)(\omega) d\omega.$$

Sampling

- Frequency samplings for integral approximation

$$\mathrm{HSIC}_\omega(X, Y) := \frac{1}{D_x D_y} \sum_{i=1}^{D_x} \sum_{j=1}^{D_y} \left| \phi_{\mathbb{P}_{XY}}(\omega_{x;i}, \omega_{y;j}) - \phi_{\mathbb{P}_X \mathbb{P}_Y}(\omega_{x;i}, \omega_{y;j}) \right|^2,$$

where $\{\omega_{x;i}\}_{i=1}^{D_x}, \{\omega_{y;j}\}_{j=1}^{D_y}$ are sampled independently with the measure $\mathcal{F}^{-1}\psi_k, \mathcal{F}^{-1}\psi_l$, respectively. And $\mathcal{F}^{-1}\psi$ is a product measure, i.e., $\mathcal{F}^{-1}\psi = (\mathcal{F}^{-1}\psi_k) \otimes (\mathcal{F}^{-1}\psi_l)$.

This type of approximation also called random Fourier features (RFF).

Obtain Independence Criterion

Design $\mathcal{F}^{-1}\psi$ (take $\mathcal{F}^{-1}\psi_k(\omega)$ for example)

- Designing by kernels with adjustable parameters.

Kernel	$\psi_k(\Delta)$	$\mathcal{F}^{-1}\psi_k(\omega)$	$\mathcal{T}_{\theta_k}(x)$	$p_k(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2\sigma^2}}$	$(2\pi)^{-d_x/2}\sigma e^{-\sigma^2\|\omega\|_2^2/2}$	x/σ	$(2\pi)^{-d_x/2}e^{-\|\omega\|_2^2/2}$
Laplace	$e^{-\frac{\|\Delta\|_1}{\sigma}}$	$\sqrt{\frac{2}{\pi}}\prod_d \frac{\sigma}{\sigma^2+\omega_d^2}$	x/σ	$\sqrt{\frac{2}{\pi}}\prod_d \frac{1}{1+\omega_d^2}$
Mahalanobis	$e^{-\frac{1}{2}\Delta^T\Sigma^{-1}\Delta}$	$(2\pi)^{-d_x/2}\|\Sigma\|^{-1/2}e^{-\omega^T\Sigma^{-1}\omega/2}$	$\Sigma^{1/2}x$	$(2\pi)^{-d_x/2}e^{-\|\omega\|_2^2/2}$

Table: Some popular kernels (parametered σ, Σ) with corresponding density functions.

Obtain Independence Criterion

Design $\mathcal{F}^{-1}\psi$ (take $\mathcal{F}^{-1}\psi_k(\omega)$ for example)

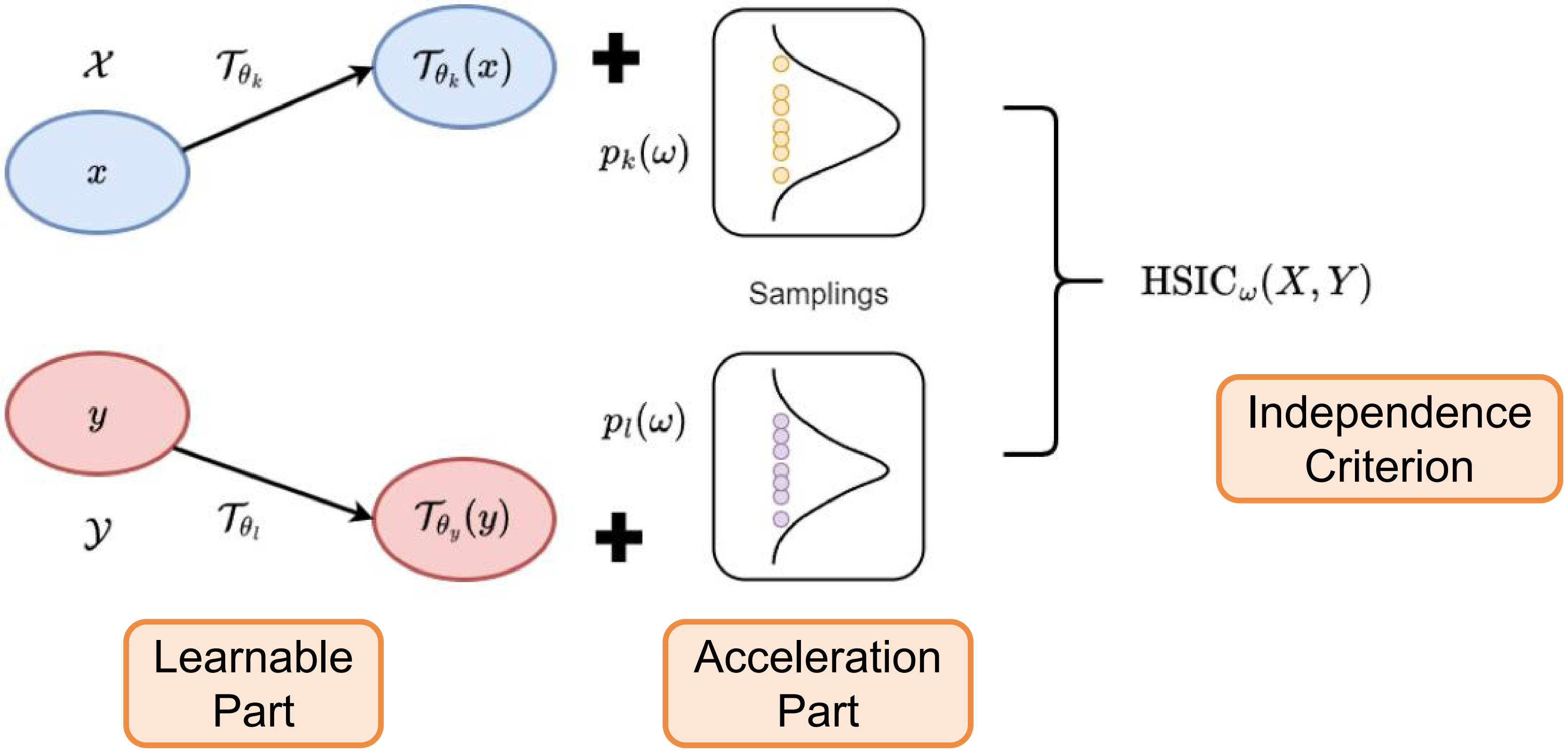
- Designing by kernels with adjustable parameters.

Kernel	$\psi_k(\Delta)$	$\mathcal{F}^{-1}\psi_k(\omega)$	$\mathcal{T}_{\theta_k}(x)$	$p_k(\omega)$
Gaussian	$e^{-\frac{\|\Delta\|_2^2}{2\sigma^2}}$	$(2\pi)^{-d_x/2}\sigma e^{-\sigma^2\|\omega\|_2^2/2}$	x/σ	$(2\pi)^{-d_x/2}e^{-\|\omega\|_2^2/2}$
Laplace	$e^{-\frac{\|\Delta\|_1}{\sigma}}$	$\sqrt{\frac{2}{\pi}}\prod_d \frac{\sigma}{\sigma^2+\omega_d^2}$	x/σ	$\sqrt{\frac{2}{\pi}}\prod_d \frac{1}{1+\omega_d^2}$
Mahalanobis	$e^{-\frac{1}{2}\Delta^T\Sigma^{-1}\Delta}$	$(2\pi)^{-d_x/2}\|\Sigma\|^{-1/2}e^{-\omega^T\Sigma^{-1}\omega/2}$	$\Sigma^{1/2}x$	$(2\pi)^{-d_x/2}e^{-\|\omega\|_2^2/2}$

Table: Some popular kernels (parametered σ, Σ) with corresponding density functions.

- Disentangling the sampled objects and the learnable parameters
  - Relocate the learnable component onto X using $\mathcal{T}_{\theta_k}(x)$.
  - Convert the probability measure $\mathcal{F}^{-1}\psi_k$ into a standard distribution $p_k(\omega)$.

Obtain Independence Criterion



Learnable RFF

We obtain the approximation of kernel

- Kernel with learnable mappings

$$\psi_k\left(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x'\right) = \mathcal{F}[\mathcal{F}^{-1}\psi_k(\omega)] = \int e^{-i\omega^T(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x')}p_k(\omega)d\omega.$$

- Use the frequency sampling technique

$$\psi_k^{(\omega)}\left(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x'\right) := \frac{2}{D}\sum_{j=1}^{D/2} e^{-i\omega_{k;j}^T(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x')} = \frac{2}{D}\sum_{j=1}^{D/2}\cos\left(\omega_{k;j}^T(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x')\right),$$

  where $\{\omega_{k;j}\}_{j=1}^{D/2}$ are sampled independently with distribution $p_k(\omega)$.
- The learnable RFF of k

$$\Lambda_k(x) := \sqrt{\frac{2}{D}}\left[\cos(\omega_1^T\mathcal{T}_{\theta_k}x), \sin(\omega_1^T\mathcal{T}_{\theta_k}x), \ldots, \cos(\omega_{D/2}^T\mathcal{T}_{\theta_k}x), \sin(\omega_{D/2}^T\mathcal{T}_{\theta_k}x)\right],$$

- Hence $\psi_k^{(\omega)}\left(\mathcal{T}_{\theta_k}x - \mathcal{T}_{\theta_k}x'\right) = \Lambda_k(x)\Lambda_k(x')^T$.

Independence Criterion

The statistic can be obtained as follows.

- The learnable RFF of samples in matrix form $\mathbf{\Lambda}_X := [\Lambda_k(x_1); \dots; \Lambda_k(x_n)]_{n \times D}$.
- Obtain $\mathbf{\Lambda}_Y$ by analogy. The same number of samplings are used for simplify.
- The statistic with sample Z

$$\mathrm{HSIC}_\omega(Z) := \frac{1}{n^2}\mathrm{Tr}(\mathbf{\Lambda}_X \mathbf{\Lambda}_X^T \mathbf{H} \mathbf{\Lambda}_Y \mathbf{\Lambda}_Y^T \mathbf{H}) = \frac{1}{n^2}\mathrm{Tr}(\mathbf{\Lambda}_X^T \mathbf{H} \mathbf{\Lambda}_Y \mathbf{\Lambda}_Y^T \mathbf{H} \mathbf{\Lambda}_X) = \frac{1}{n^2}\|\mathbf{\Lambda}_{Xc}^T \mathbf{\Lambda}_{Yc}\|_F^2,$$

  where $\mathbf{\Lambda}_{Xc} := \mathbf{H}\mathbf{\Lambda}_X, \mathbf{\Lambda}_{Yc} := \mathbf{H}\mathbf{\Lambda}_Y$.
- The time complexity is $\mathcal{O}\big(nD(d_x + d_y + D)\big)$, i.e. the running time is linear with n.

Asymptotic Behavior

Proposition (Asymptotics)

Let $h_{ijqr}^{(\omega)} := \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,q,r)} k_{tu}^{(\omega)} l_{tu}^{(\omega)} + k_{tu}^{(\omega)} l_{vw}^{(\omega)} - 2k_{uv}^{(\omega)} l_{tv}^{(\omega)}$. *Then, Under the null hypothesis* $\mathcal{H}_0$, $HSIC_\omega(Z)$ *coverages in distribution to*

$$nHSIC_\omega(Z) \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l \chi_{1l}^2, \quad \lambda_l g_l(z_j) = \int_{z_i,z_q,z_r} h_{ijqr}^{(\omega)} g_l(z_i) dF_{z_i,z_q,z_r},$$

where $\chi_{11}^2, \chi_{12}^2, \dots$ *are independent* χ_1^2 *variates and* λ_l *is the solution to eigenvalue problem. Also, under* $\mathcal{H}_1$, $HSIC_\omega(Z)$ *converges in distribution as*

$$n^{\frac{1}{2}}\Big(HSIC_\omega(Z) - \mathbf{E}_Z HSIC_\omega(Z)\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_\omega^2), \ \sigma_\omega^2 := 16\Big[\mathbf{E}_i(\mathbf{E}_{j,q,r} h_{ijqr}^{(\omega)})^2 - (\mathbf{E}_Z h_{ijqr}^{(\omega)})^2\Big]$$

with the simplified notation $\mathbf{E}_{j,q,r} := \mathbf{E}_{z_j,z_q,z_r}$ *and* $\mathbf{E}_Z := \mathbf{E}_{z_i,z_j,z_q,z_r}$.

Construct Optimization Objective

- According to the asymptotics, the power of the test

$$\mathbb{P}_{\mathcal{H}_1}\left(n\mathrm{HSIC}_\omega(Z) > r_\omega\right) \to \Phi\left(\frac{n\mathbf{E}_Z\mathrm{HSIC}_\omega(Z) - r_\omega}{\sqrt{n}\sigma_\omega}\right),$$

  where Φ is the standard normal CDF.

- The optimization objective

$$J := \frac{(n\mathrm{HSIC}_\omega(Z) - \widehat{c_\alpha})}{(\sqrt{n}\widehat{\sigma}_\omega)},$$

  where $\mathrm{HSIC}_\omega(Z)$ is the criterion, $\widehat{c_\alpha}$ is a estimate of threshold and the estimate of variance $\widehat{\sigma}_\omega^2 := 16\left[\frac{1}{n}\sum_i\left(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}^{(\omega)}\right)^2 - \mathrm{HSIC}_\omega^2(Z)\right]$.

Obtain $\widehat{c_\alpha}$ with Gamma approximation.

- Determined completely by the first two moments of distribution under $\mathcal{H}_0$.

$$n\mathrm{HSIC}_\omega(Z) \sim \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma \Gamma(\gamma)}, \text{ where } \gamma = \frac{(\mathbf{E}[\mathrm{HSIC}_\omega(Z)])^2}{\mathrm{var}[\mathrm{HSIC}_\omega(Z)]}, \beta = \frac{n\mathbf{Var}[\mathrm{HSIC}_\omega(Z)]}{\mathbf{E}[\mathrm{HSIC}_\omega(Z)]}$$

- The $(1-\alpha)$-quantile of Gamma distribution

$$\int_0^{\widehat{c_\alpha}} \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma \Gamma(\gamma)} dx = 1 - \alpha.$$

Obtain $\widehat{c_\alpha}$ with Gamma approximation.

- Determined completely by the first two moments of distribution under $\mathcal{H}_0$.

$$n\mathrm{HSIC}_\omega(Z) \sim \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma \Gamma(\gamma)}, \text{ where } \gamma = \frac{(\mathbf{E}[\mathrm{HSIC}_\omega(Z)])^2}{\mathrm{var}[\mathrm{HSIC}_\omega(Z)]}, \beta = \frac{n\mathbf{Var}[\mathrm{HSIC}_\omega(Z)]}{\mathbf{E}[\mathrm{HSIC}_\omega(Z)]}$$

- The $(1-\alpha)$-quantile of Gamma distribution

$$\int_0^{\widehat{c_\alpha}} \frac{x^{\gamma-1}e^{-x/\beta}}{\beta^\gamma \Gamma(\gamma)} dx = 1-\alpha.$$

The linear-time estimations of two moments of distribution under $\mathcal{H}_0$.

Theorem (Linear-Time Estimations)

Under $\mathcal{H}_0$, the estimation of mean and variance with bias of $\mathcal{O}(n^{-1})$ to $\mathbf{E}_Z[nHSIC_\omega(Z)]$ and $\mathbf{Var}_Z[nHSIC_\omega(Z)]$, denote as $\mathcal{E}_0$ and $\mathcal{V}_0$, respectively, are given by

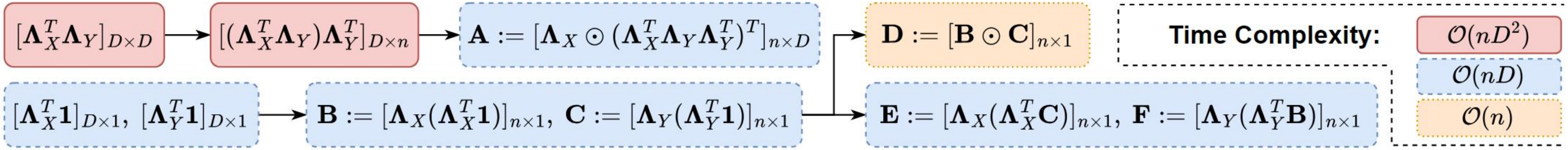
$$\mathcal{E}_0 := \frac{[\mathbf{1}^T \mathbf{\Lambda}_{Xc}^{\cdot 2}\mathbf{1}][\mathbf{1}^T \mathbf{\Lambda}_{Yc}^{\cdot 2}\mathbf{1}]}{(n-1)^2}, \mathcal{V}_0 := \frac{2n(n-4)(n-5)}{(n-1)(n-2)(n-3)} \frac{[\mathbf{1}^T(\mathbf{\Lambda}_{Xc}^T\mathbf{\Lambda}_{Xc})^{\cdot 2}\mathbf{1}][\mathbf{1}^T(\mathbf{\Lambda}_{Yc}^T\mathbf{\Lambda}_{Yc})^{\cdot 2}\mathbf{1}]}{n^4},$$

where $()^{\cdot 2}$ is the entrywise matrix power. Both $\mathcal{E}_0$ and $\mathcal{V}_0$ can be calculated in $\mathcal{O}(nD^2)$ time.

The estimate of variance $\widehat{\sigma}_\omega^2$ also can be calculated in linear-time

- By the definition $\widehat{\sigma}_\omega^2 := 16\left[\frac{1}{n}\sum_i(\frac{1}{n^3}\sum_{j,q,r} h_{ijqr}^{(\omega)})^2 - \mathrm{HSIC}_\omega^2(Z)\right]$.
- Calculate $\sum_{j,q,r} h_{ijqr}^{(\omega)}$ in linear time

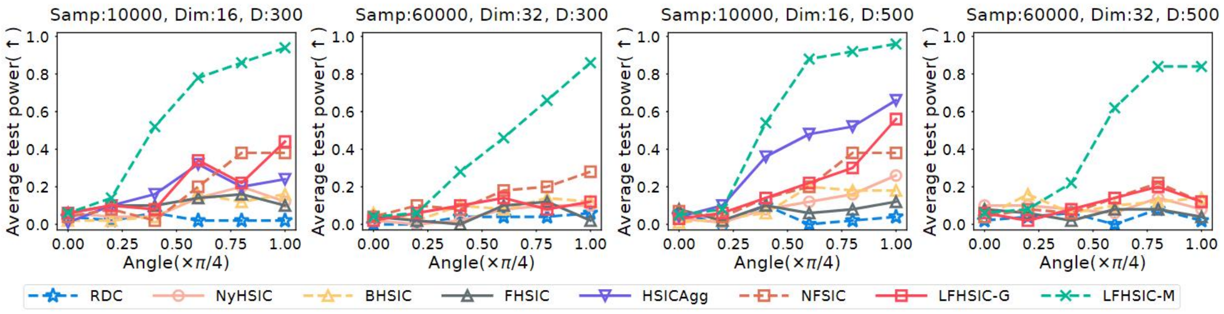
$$\sum_{j,q,r} h_{ijqr}^{(\omega)} = \frac{1}{2}\left[n\mathbf{1}^T\mathbf{A}\mathbf{1} + n^2(\mathbf{A}\mathbf{1})_i + (\mathbf{1}^T\mathbf{C})\mathbf{B}_i + (\mathbf{1}^T\mathbf{B})\mathbf{C}_i - n\mathbf{E}_i - n\mathbf{F}_i - n\mathbf{D}_i - \mathbf{1}^T\mathbf{D}\right],$$



As a result, the optimization objective can be computed in linear time.

Experiments

Results on ISA Datasets (Large Scale)



Results on 3DShape (High-dimensional) and Real Datasets

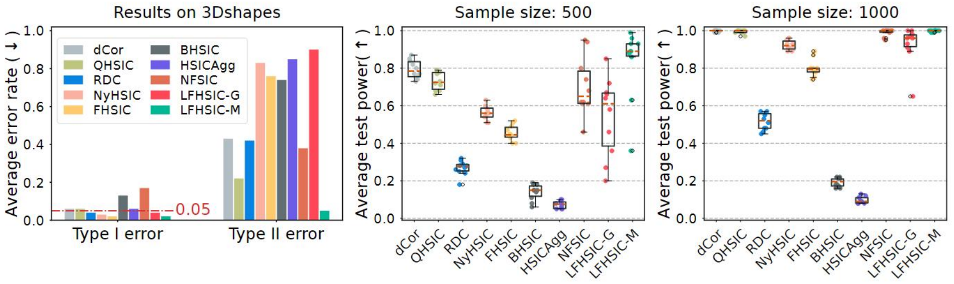
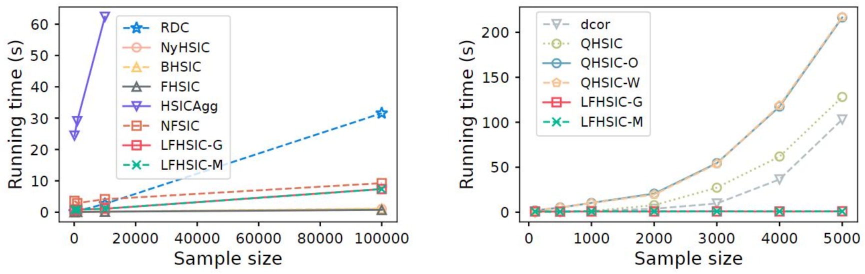


Figure: The results on two real data. Left: 3DShapes. Right two: MSD Dataset.

Running time comparing



Our test can handle 100,000 samples within 10 seconds.

- Flexible: handle large scale/high-dimensional settings well.
- Fast: linear-time time/space computation complexity.

Conclusions

The method achieves flexible independence testing in linear time (w.r.t sample size).

1. Flexible: The kernel can be *adaptive*.
2. Fast: *linear-time time/space* complexity.

Thank you for you attention!