An Adaptive Test of Independence with Analytic Kernel Embeddings

Wittawat Jitkrittum
Gatsby Unit, University College London
wittawat@gatsby.ucl.ac.uk

Probabilistic Graphical Model Workshop 2017
Institute of Statistical Mathematics, Tokyo

24 Feb 2017

Reference

Coauthors:

Arthur Gretton
Gatsby Unit, UCL

Zoltán Szabó
École Polytechnique

Preprint:

An Adaptive Test of Independence with Analytic Kernel Embeddings
Wittawat Jitkrittum, Zoltán Szabó, Arthur Gretton
https://arxiv.org/abs/1610.04782

- **Python code**: https://github.com/wittawatj/fsic-test

What Is Independence Testing?

- Let $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$ be random vectors following P_{xy}.
- Given a joint sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \sim P_{xy}$ (unknown), test

$$\begin{aligned} H_0 :& P_{xy} = P_x P_y, \\ \text{vs. } H_1 :& P_{xy} \neq P_x P_y. \end{aligned}$$

- $P_{xy} = P_x P_y$ equivalent to $X \perp Y$.
- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n \geq T_\alpha$ (threshold).
- $T_\alpha = (1-\alpha)$-quantile of the null distribution.

What Is Independence Testing?

- Let $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$ be random vectors following P_{xy}.
- Given a joint sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n} \sim P_{xy}$ (unknown), test

$$\begin{aligned} H_0 &: P_{xy} = P_x P_y, \\ \text{vs. } H_1 &: P_{xy} \neq P_x P_y. \end{aligned}$$

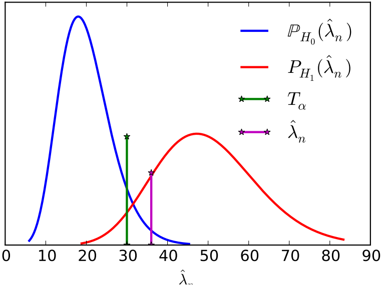
- $P_{xy} = P_x P_y$ equivalent to $X \perp Y$.
- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n \geq T_\alpha$ (threshold).
- $T_\alpha = (1-\alpha)$-quantile of the null distribution.

What Is Independence Testing?

- Let $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$ be random vectors following P_{xy}.
- Given a joint sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \sim P_{xy}$ (unknown), test

$$\begin{aligned} H_0 : & P_{xy} = P_x P_y, \\ \text{vs. } H_1 : & P_{xy} \neq P_x P_y. \end{aligned}$$

- $P_{xy} = P_x P_y$ equivalent to $X \perp Y$.
- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n \geq T_\alpha$ (threshold).
- $T_\alpha = (1-\alpha)$-quantile of the null distribution.

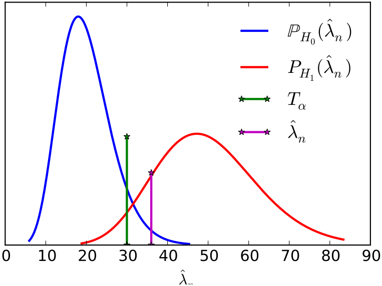


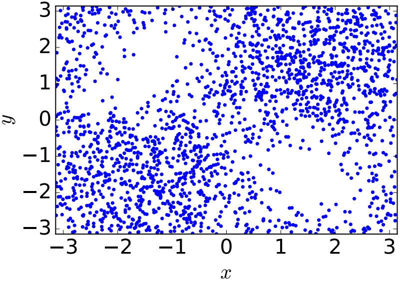
What Is Independence Testing?

- Let $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$ be random vectors following P_{xy}.
- Given a joint sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n} \sim P_{xy}$ (unknown), test

$$\begin{aligned} H_0 :& P_{xy} = P_x P_y, \\ \text{vs. } H_1 :& P_{xy} \neq P_x P_y. \end{aligned}$$

- $P_{xy} = P_x P_y$ equivalent to $X \perp Y$.
- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n \geq T_\alpha$ (threshold).
- $T_\alpha = (1 - \alpha)$-quantile of the null distribution.





Goals

Want a test which is . . .

1. Non-parametric i.e., no parametric assumption on P_{xy}.
2. Linear-time i.e., computational complexity is $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., has a well-defined criterion for parameter tuning.

	Non-parametric	$\mathcal{O}(n)$	Adaptive
Pearson correlation	✗	✓	✓
HSIC [Gretton et al., 2005]	✓	✗	✗
HSIC with RFFs* [Zhang et al., 2016]	✓	✓	✗
FSIC (proposed)	✓	✓	✓

* : RFFs = Random Fourier Features

- Focus on cases where n (sample size) is large.

Goals

Want a test which is ...

1. Non-parametric i.e., no parametric assumption on P_{xy}.
2. Linear-time i.e., computational complexity is $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., has a well-defined criterion for parameter tuning.

	Non-parametric	$\mathcal{O}(n)$	Adaptive
Pearson correlation	✗	✓	✓
HSIC [Gretton et al., 2005]	✓	✗	✗
HSIC with RFFs* [Zhang et al., 2016]	✓	✓	✗
FSIC (proposed)	✓	✓	✓

* : RFFs = Random Fourier Features

- Focus on cases where n (sample size) is large.

Goals

Want a test which is ...

1. Non-parametric i.e., no parametric assumption on P_{xy}.
2. Linear-time i.e., computational complexity is $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., has a well-defined criterion for parameter tuning.

	Non-parametric	$\mathcal{O}(n)$	Adaptive
Pearson correlation	✗	✓	✓
HSIC [Gretton et al., 2005]	✓	✗	✗
HSIC with RFFs* [Zhang et al., 2016]	✓	✓	✗
FSIC (proposed)	✓	✓	✓

* : RFFs = Random Fourier Features

- Focus on cases where n (sample size) is large.

Goals

Want a test which is ...

1. Non-parametric i.e., no parametric assumption on P_{xy}.
2. Linear-time i.e., computational complexity is $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., has a well-defined criterion for parameter tuning.

	Non-parametric	$\mathcal{O}(n)$	Adaptive
Pearson correlation	✗	✓	✓
HSIC [Gretton et al., 2005]	✓	✗	✗
HSIC with RFFs* [Zhang et al., 2016]	✓	✓	✗
FSIC (proposed)	✓	✓	✓

* : RFFs = Random Fourier Features

- Focus on cases where n (sample size) is large.

Goals

Want a test which is ...

1. Non-parametric i.e., no parametric assumption on P_{xy}.
2. Linear-time i.e., computational complexity is $\mathcal{O}(n)$. Fast.
3. Adaptive i.e., has a well-defined criterion for parameter tuning.

	Non-parametric	$\mathcal{O}(n)$	Adaptive
Pearson correlation	✗	✓	✓
HSIC [Gretton et al., 2005]	✓	✗	✗
HSIC with RFFs* [Zhang et al., 2016]	✓	✓	✗
FSIC (proposed)	✓	✓	✓

* : RFFs = Random Fourier Features

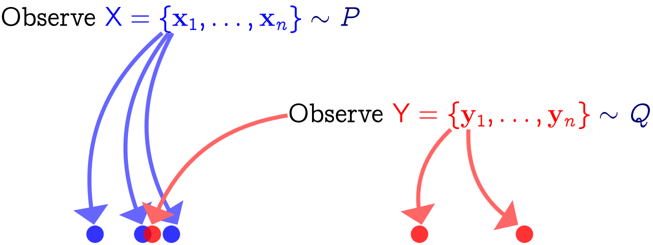
- Focus on cases where n (sample size) is large.

Witness Function [Gretton et al., 2012]

- A function showing the differences of two distributions P and Q.
- Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}\right)$
- Empirical mean embedding of P: $\hat{\mu}_P(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$
- Maximum Mean Discrepancy (MMD): $\|\hat{u}\|_{\mathrm{RKHS}}$.
  - $\mathrm{MMD}(P, Q) = 0$ if and only if $P = Q$.

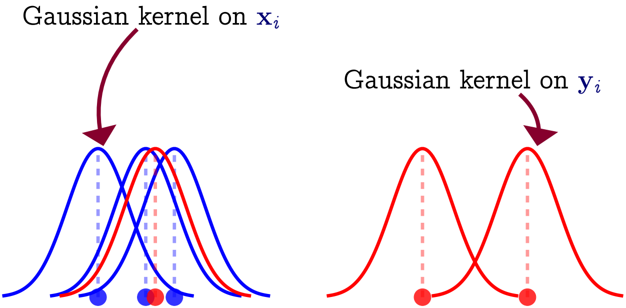
Witness Function [Gretton et al., 2012]

- A function showing the differences of two distributions P and Q.
- Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}\right)$
- Empirical mean embedding of P: $\hat{\mu}_P(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$
- Maximum Mean Discrepancy (MMD): $\|\hat{u}\|_{\mathrm{RKHS}}$.
  - $\mathrm{MMD}(P, Q) = 0$ if and only if $P = Q$.



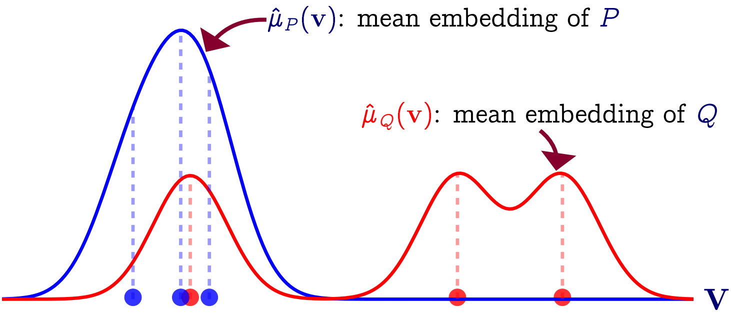
Witness Function [Gretton et al., 2012]

- A function showing the differences of two distributions P and Q.
- Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}\right)$
- Empirical mean embedding of P: $\hat{\mu}_P(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$
- Maximum Mean Discrepancy (MMD): $\|\hat{u}\|_{\text{RKHS}}$.
  - $\text{MMD}(P, Q) = 0$ if and only if $P = Q$.



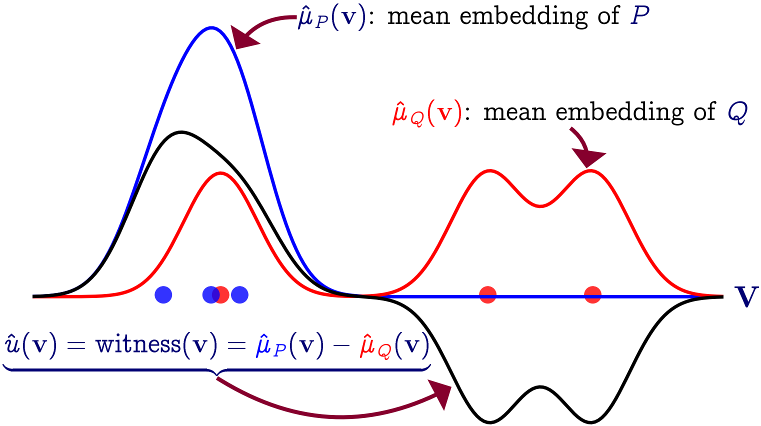
Witness Function [Gretton et al., 2012]

- A function showing the differences of two distributions P and Q.
- Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}\right)$
- Empirical mean embedding of P: $\hat{\mu}_P(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$
- Maximum Mean Discrepancy (MMD): $\|\hat{u}\|_{\text{RKHS}}$.
  - $\text{MMD}(P, Q) = 0$ if and only if $P = Q$.



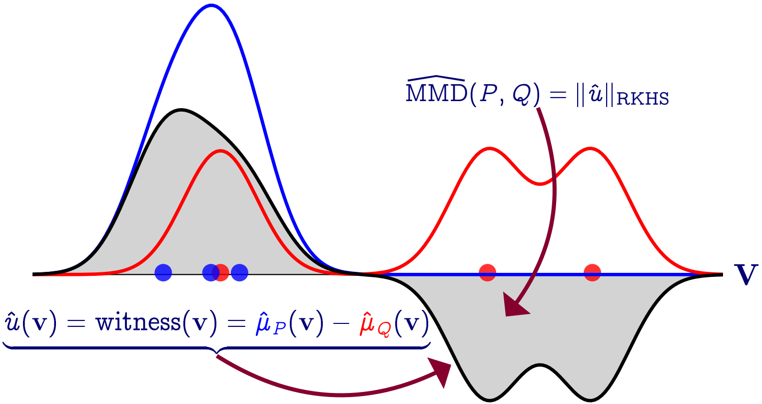
Witness Function [Gretton et al., 2012]

- A function showing the differences of two distributions P and Q.
- Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}\right)$
- Empirical mean embedding of P: $\hat{\mu}_P(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$
- Maximum Mean Discrepancy (MMD): $\|\hat{u}\|_{\mathrm{RKHS}}$.
  - $\mathrm{MMD}(P, Q) = 0$ if and only if $P = Q$.



Witness Function [Gretton et al., 2012]

- A function showing the differences of two distributions P and Q.
- Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma^2}\right)$
- Empirical mean embedding of P: $\hat{\mu}_P(\mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})$
- Maximum Mean Discrepancy (MMD): $\|\hat{u}\|_{\text{RKHS}}$.
  - $\text{MMD}(P, Q) = 0$ if and only if $P = Q$.



Independence Test with HSIC [Gretton et al., 2005]

- Hilbert-Schmidt Independence Criterion.

$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

(need two kernels: k for X, and l for Y).

- Empirical witness:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$$

where $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}) l(\mathbf{y}_i, \mathbf{w})$.

- $\mathrm{HSIC}(X, Y) = 0$ if and only if X and Y are independent.
- Test statistic $= \|\hat{u}\|_{\mathrm{RKHS}}$ ("flatness" of $\hat{u}$). Complexity: $\mathcal{O}(n^2)$.

Key: Can we measure the flatness by other way that costs only $\mathcal{O}(n)$?

Independence Test with HSIC [Gretton et al., 2005]

- Hilbert-Schmidt Independence Criterion.

$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

(need two kernels: k for X, and l for Y).

- Empirical witness:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$$

where $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_i, \mathbf{w})$.

- $\mathrm{HSIC}(X, Y) = 0$ if and only if X and Y are independent.
- Test statistic $= \|\hat{u}\|_{\mathrm{RKHS}}$ ("flatness" of $\hat{u}$). Complexity: $\mathcal{O}(n^2)$.

Key: Can we measure the flatness by other way that costs only $\mathcal{O}(n)$?

Independence Test with HSIC [Gretton et al., 2005]

- Hilbert-Schmidt Independence Criterion.

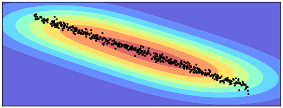
$$\text{HSIC}(X, Y) = \text{MMD}(P_{xy}, P_x P_y) = \|u\|_{\text{RKHS}}$$

(need two kernels: k for X, and l for Y).

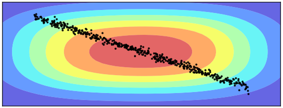
- Empirical witness:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$$

where $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}) l(\mathbf{y}_i, \mathbf{w})$.



$-$



$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$ $\qquad$ $\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$

- $\text{HSIC}(X, Y) = 0$ if and only if X and Y are independent.
- Test statistic $= \|\hat{u}\|_{\text{RKHS}}$ ("flatness" of $\hat{u}$). Complexity: $\mathcal{O}(n^2)$.

Key: Can we measure the flatness by other way that costs only $\mathcal{O}(n)$?

Independence Test with HSIC [Gretton et al., 2005]

- Hilbert-Schmidt Independence Criterion.

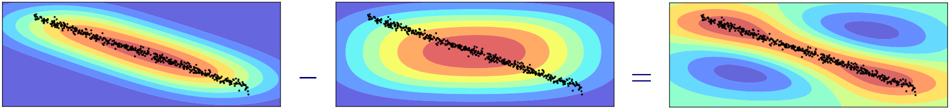
$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

(need two kernels: k for X, and l for Y).

- Empirical witness:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$$

where $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_i, \mathbf{w})$.



$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$ $-$ $\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$ $=$ Witness $\hat{u}(\mathbf{v}, \mathbf{w})$

- $\mathrm{HSIC}(X, Y) = 0$ if and only if X and Y are independent.
- Test statistic $= \|\hat{u}\|_{\mathrm{RKHS}}$ ("flatness" of $\hat{u}$). Complexity: $\mathcal{O}(n^2)$.

Key: Can we measure the flatness by other way that costs only $\mathcal{O}(n)$?

Independence Test with HSIC [Gretton et al., 2005]

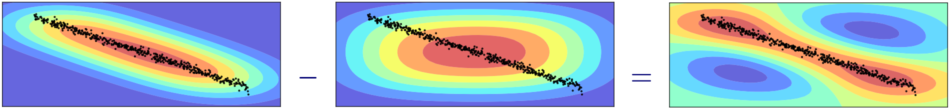
- Hilbert-Schmidt Independence Criterion.

$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

  (need two kernels: k for X, and l for Y).
- Empirical witness:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$$

  where $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}) l(\mathbf{y}_i, \mathbf{w})$.



$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$ $-$ $\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$ $=$ Witness $\hat{u}(\mathbf{v}, \mathbf{w})$

- $\mathrm{HSIC}(X, Y) = 0$ if and only if X and Y are independent.
- Test statistic $= \|\hat{u}\|_{\mathrm{RKHS}}$ ("flatness" of $\hat{u}$). Complexity: $\mathcal{O}(n^2)$.

Key: Can we measure the flatness by other way that costs only $\mathcal{O}(n)$?

Independence Test with HSIC [Gretton et al., 2005]

- Hilbert-Schmidt Independence Criterion.

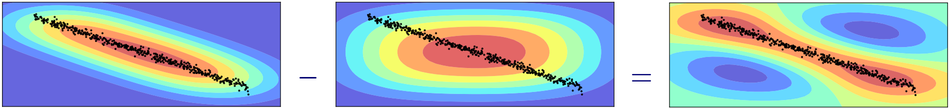
$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

(need two kernels: k for X, and l for Y).

- Empirical witness:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$$

where $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_i, \mathbf{w})$.



$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$ $\qquad$ $\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$ $\qquad$ Witness $\hat{u}(\mathbf{v}, \mathbf{w})$

- $\mathrm{HSIC}(X, Y) = 0$ if and only if X and Y are independent.
- Test statistic $= \|\hat{u}\|_{\mathrm{RKHS}}$ ("flatness" of $\hat{u}$). Complexity: $\mathcal{O}(n^2)$.

Key: Can we measure the flatness by other way that costs only $\mathcal{O}(n)$?

Independence Test with HSIC [Gretton et al., 2005]

- Hilbert-Schmidt Independence Criterion.

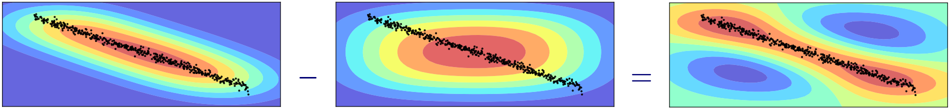
$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

(need two kernels: k for X, and l for Y).

- Empirical witness:

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$$

where $\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{v}) l(\mathbf{y}_i, \mathbf{w})$.



$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$ $\quad$ $\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$ $\quad$ Witness $\hat{u}(\mathbf{v}, \mathbf{w})$

- $\mathrm{HSIC}(X, Y) = 0$ if and only if X and Y are independent.
- Test statistic $= \|\hat{u}\|_{\mathrm{RKHS}}$ ("flatness" of $\hat{u}$). Complexity: $\mathcal{O}(n^2)$.

Key: Can we measure the flatness by other way that costs only $\mathcal{O}(n)$?

Proposal: The Finite Set Independence Criterion (FSIC)

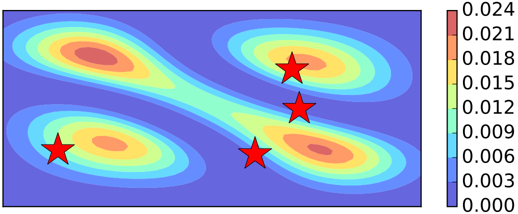
Idea: Evaluate $\hat{u}^2(\mathbf{v}, \mathbf{w})$ at only finitely many test locations.

- A set of random J locations: $\{(\mathbf{v}_1, \mathbf{w}_1), \ldots, (\mathbf{v}_J, \mathbf{w}_J)\}$
- $\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i)$
- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- **But**, what about an unlucky set of locations??
  - Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

Proposal: The Finite Set Independence Criterion (FSIC)

Idea: Evaluate $\hat{u}^2(\mathbf{v}, \mathbf{w})$ at only finitely many test locations.

- A set of random J locations: $\{(\mathbf{v}_1, \mathbf{w}_1), \ldots, (\mathbf{v}_J, \mathbf{w}_J)\}$
- $\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i)$

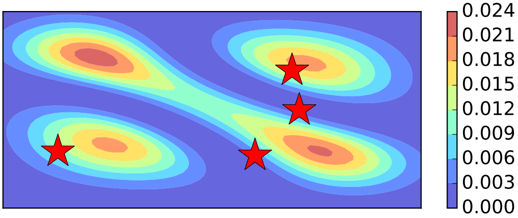


- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- But, what about an unlucky set of locations??
  - Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

Proposal: The Finite Set Independence Criterion (FSIC)

Idea: Evaluate $\hat{u}^2(\mathbf{v}, \mathbf{w})$ at only finitely many test locations.

- A set of random J locations: $\{(\mathbf{v}_1, \mathbf{w}_1), \ldots, (\mathbf{v}_J, \mathbf{w}_J)\}$
- $\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J}\sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i)$

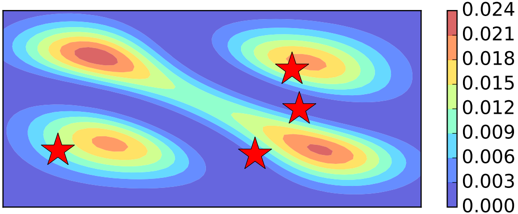


- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- But, what about an unlucky set of locations??
  - Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

Proposal: The Finite Set Independence Criterion (FSIC)

Idea: Evaluate $\hat{u}^2(\mathbf{v}, \mathbf{w})$ at only finitely many test locations.

- A set of random J locations: $\{(\mathbf{v}_1, \mathbf{w}_1), \ldots, (\mathbf{v}_J, \mathbf{w}_J)\}$
- $\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i)$

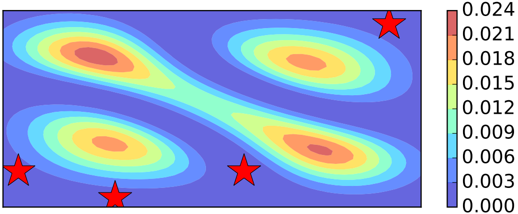


- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- **But**, what about an unlucky set of locations??
  - Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

Proposal: The Finite Set Independence Criterion (FSIC)

Idea: Evaluate $\hat{u}^2(\mathbf{v}, \mathbf{w})$ at only finitely many test locations.

- A set of random J locations: $\{(\mathbf{v}_1, \mathbf{w}_1), \ldots, (\mathbf{v}_J, \mathbf{w}_J)\}$
- $\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i)$

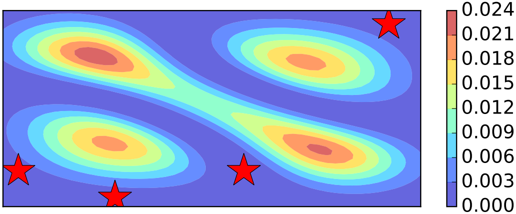


- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- **But**, what about an unlucky set of locations??
  - Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

Proposal: The Finite Set Independence Criterion (FSIC)

Idea: Evaluate $\hat{u}^2(\mathbf{v}, \mathbf{w})$ at only finitely many test locations.

- A set of random J locations: $\{(\mathbf{v}_1, \mathbf{w}_1), \ldots, (\mathbf{v}_J, \mathbf{w}_J)\}$
- $\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i)$



- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- **But**, what about an unlucky set of locations??
  - Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??
- No. Population $\mathrm{FSIC}(X, Y) = 0$ iff $X \perp Y$, almost surely.

Requirements on the Kernels

Definition 1 (Analytic kernels).

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be analytic if for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{v} \to k(\mathbf{x}, \mathbf{v})$ is a real analytic function on $\mathcal{X}$.

- Analytic: Taylor series about $\mathbf{x}_0$ converges for all $\mathbf{x}_0 \in \mathcal{X}$.
- $\implies$ k is infinitely differentiable.

Definition 2 (Characteristic kernels).

- Let P, Q be two distributions, and g be a kernel.
- Let $\mu_P(\mathbf{v}) := \mathbb{E}_{\mathbf{z}\sim P}[g(\mathbf{z}, \mathbf{v})]$ and $\mu_Q(\mathbf{v}) := \mathbb{E}_{\mathbf{z}\sim Q}[g(\mathbf{z}, \mathbf{v})]$.

g is said to be characteristic if $P \neq Q$ implies $\mu_P \neq \mu_Q$.

Requirements on the Kernels

Definition 1 (Analytic kernels).

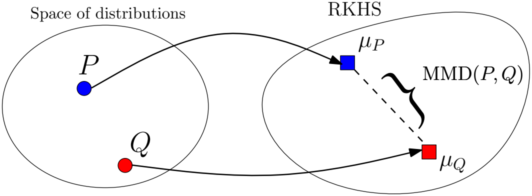
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be analytic if for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{v} \to k(\mathbf{x}, \mathbf{v})$ is a real analytic function on $\mathcal{X}$.

- Analytic: Taylor series about $\mathbf{x}_0$ converges for all $\mathbf{x}_0 \in \mathcal{X}$.
- $\implies$ k is infinitely differentiable.

Definition 2 (Characteristic kernels).

- Let P, Q be two distributions, and g be a kernel.
- Let $\mu_P(\mathbf{v}) := \mathbb{E}_{\mathbf{z}\sim P}[g(\mathbf{z}, \mathbf{v})]$ and $\mu_Q(\mathbf{v}) := \mathbb{E}_{\mathbf{z}\sim Q}[g(\mathbf{z}, \mathbf{v})]$.

g is said to be characteristic if $P \neq Q$ implies $\mu_P \neq \mu_Q$.



FSIC Is a Dependence Measure

Proposition 1.

Assume

1. *The product kernel* $g((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := k(\mathbf{x}, \mathbf{x}')l(\mathbf{y}, \mathbf{y}')$ *is characteristic and analytic (i.e.,* k, l *are Gaussian kernels).*
2. *Test locations* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$ *where* η *has a density.*

Then, η*-almost surely,* $\mathrm{FSIC}(X, Y) = 0$ *iff* X *and* Y *are independent.*

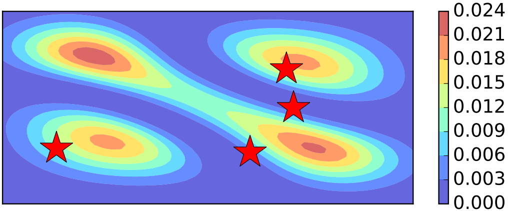
FSIC Is a Dependence Measure

Proposition 1.

Assume

1. *The product kernel* $g((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := k(\mathbf{x}, \mathbf{x}')l(\mathbf{y}, \mathbf{y}')$ *is* *characteristic and analytic (i.e.,* k, l *are Gaussian kernels).*
2. *Test locations* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J} \sim \eta$ *where* η *has a density.*

Then, η*-almost surely,* $\mathrm{FSIC}(X, Y) = 0$ *iff* X *and* Y *are independent.*

Let's plot  as

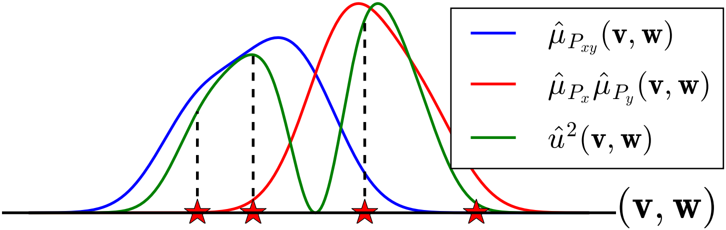
FSIC Is a Dependence Measure

Proposition 1.

Assume

1. *The product kernel* $g((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := k(\mathbf{x}, \mathbf{x}')l(\mathbf{y}, \mathbf{y}')$ *is* *characteristic* *and* *analytic* *(i.e.,* k, l *are Gaussian kernels).*
2. *Test locations* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J} \sim \eta$ *where* η *has a density.*

Then, η*-almost surely,* $\mathrm{FSIC}(X, Y) = 0$ *iff* X *and* Y *are independent.*



FSIC Is a Dependence Measure

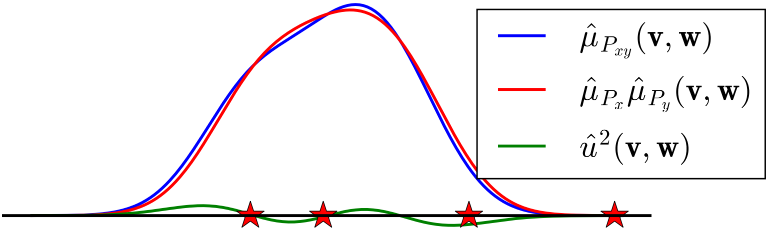
Proposition 1.

Assume

1. *The product kernel* $g((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := k(\mathbf{x}, \mathbf{x}')l(\mathbf{y}, \mathbf{y}')$ *is* *characteristic* *and* *analytic* *(i.e.,* k, l *are Gaussian kernels).*
2. *Test locations* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J} \sim \eta$ *where* η *has a density.*

Then, η*-almost surely,* $\mathrm{FSIC}(X, Y) = 0$ *iff* X *and* Y *are independent.*

Under H_0,



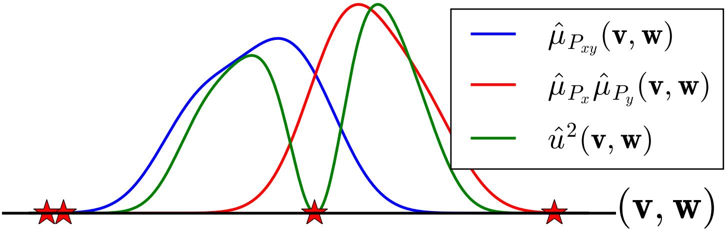
FSIC Is a Dependence Measure

Proposition 1.

Assume

1. *The product kernel* $g((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := k(\mathbf{x}, \mathbf{x}')l(\mathbf{y}, \mathbf{y}')$ *is* *characteristic* *and* *analytic* *(i.e.,* k, l *are Gaussian kernels).*
2. *Test locations* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$ *where* η *has a density.*

Then, η*-almost surely,* $\mathrm{FSIC}(X, Y) = 0$ *iff* X *and* Y *are independent.*



- Under H_1, u is not a zero function ($P \mapsto \mathbb{E}_{\mathbf{z}\sim P}[g(\mathbf{z}, \cdot)]$ is injective).
- u is analytic. So, $R_u = \{(\mathbf{v}, \mathbf{w}) \mid u(\mathbf{v}, \mathbf{w}) = 0\}$ has 0 Lebesgue measure.
- So, $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$ will not be in R_u (with probability 1).

Alternative View of the Witness $u(\mathbf{v}, \mathbf{w})$

The witness $u(\mathbf{v}, \mathbf{w})$ can be rewritten as

$$
\begin{aligned}
u(\mathbf{v}, \mathbf{w}) &:= \mu_{xy}(\mathbf{v}, \mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}) \\
&= \mathbb{E}_{\mathbf{xy}}[k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})]\mathbb{E}_{\mathbf{y}}[l(\mathbf{y}, \mathbf{w})], \\
&= \mathrm{cov}_{\mathbf{xy}}[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})].
\end{aligned}
$$

1. Transforming $\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v})$ and $\mathbf{y} \mapsto l(\mathbf{y}, \mathbf{w})$ (from $\mathbb{R}^{d_y}$ to $\mathbb{R}$).
2. Then, take the covariance.

> The kernel transformations turn the linear covariance into a dependence measure.

Alternative View of the Witness $u(\mathbf{v}, \mathbf{w})$

The witness $u(\mathbf{v}, \mathbf{w})$ can be rewritten as

$$\begin{aligned} u(\mathbf{v}, \mathbf{w}) &:= \mu_{xy}(\mathbf{v}, \mathbf{w}) - \mu_x(\mathbf{v})\mu_y(\mathbf{w}) \\ &= \mathbb{E}_{\mathbf{xy}}[k(\mathbf{x}, \mathbf{v})l(\mathbf{y}, \mathbf{w})] - \mathbb{E}_{\mathbf{x}}[k(\mathbf{x}, \mathbf{v})]\mathbb{E}_{\mathbf{y}}[l(\mathbf{y}, \mathbf{w})], \\ &= \mathrm{cov}_{\mathbf{xy}}[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})]. \end{aligned}$$

1. Transforming $\mathbf{x} \mapsto k(\mathbf{x}, \mathbf{v})$ and $\mathbf{y} \mapsto l(\mathbf{y}, \mathbf{w})$ (from $\mathbb{R}^{d_y}$ to $\mathbb{R}$).
2. Then, take the covariance.

The kernel transformations turn the linear covariance into a dependence measure.

Alternative Form of $\hat{u}(\mathbf{v}, \mathbf{w})$

- Recall $\widehat{\mathrm{FSIC}^2} = \frac{1}{J}\sum_{i=1}^{J} \hat{u}(\mathbf{v}_i, \mathbf{w}_i)^2$
- Let $\widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w})$ be an unbiased estimator of $\mu_x(\mathbf{v})\mu_y(\mathbf{w})$.
- $\widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_j, \mathbf{w})$.
- An unbiased estimator of $u(\mathbf{v}, \mathbf{w})$ is

$$\begin{aligned}\hat{u}(\mathbf{v}, \mathbf{w}) &= \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) \\ &= \frac{2}{n(n-1)}\sum_{i<j} h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)),\end{aligned}$$

  where

$$h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := \frac{1}{2}(k(\mathbf{x}, \mathbf{v}) - k(\mathbf{x}', \mathbf{v}))(l(\mathbf{y}, \mathbf{w}) - l(\mathbf{y}', \mathbf{w})).$$

- For a fixed $(\mathbf{v}, \mathbf{w})$, $\hat{u}(\mathbf{v}, \mathbf{w})$ is a one-sample 2^{nd}-order U-statistic.

Alternative Form of $\hat{u}(\mathbf{v}, \mathbf{w})$

- Recall $\widehat{\mathrm{FSIC}^2} = \frac{1}{J}\sum_{i=1}^{J} \hat{u}(\mathbf{v}_i, \mathbf{w}_i)^2$
- Let $\widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w})$ be an unbiased estimator of $\mu_x(\mathbf{v})\mu_y(\mathbf{w})$.
- $\widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) := \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(\mathbf{x}_i, \mathbf{v})l(\mathbf{y}_j, \mathbf{w})$.
- An unbiased estimator of $u(\mathbf{v}, \mathbf{w})$ is

$$\begin{aligned}\hat{u}(\mathbf{v}, \mathbf{w}) &= \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \widehat{\mu_x\mu_y}(\mathbf{v}, \mathbf{w}) \\ &= \frac{2}{n(n-1)}\sum_{i<j} h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)),\end{aligned}$$

where

$$h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := \frac{1}{2}(k(\mathbf{x}, \mathbf{v}) - k(\mathbf{x}', \mathbf{v}))(l(\mathbf{y}, \mathbf{w}) - l(\mathbf{y}', \mathbf{w})).$$

- For a fixed $(\mathbf{v}, \mathbf{w})$, $\hat{u}(\mathbf{v}, \mathbf{w})$ is a one-sample 2^{nd}-order U-statistic.

Alternative Form of $\hat{u}(\mathbf{v}, \mathbf{w})$

- Recall $\widehat{\mathrm{FSIC}^2} = \frac{1}{J}\sum_{i=1}^{J} \hat{u}(\mathbf{v}_i, \mathbf{w}_i)^2$
- Let $\widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w})$ be an unbiased estimator of $\mu_x(\mathbf{v})\mu_y(\mathbf{w})$.
- $\widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w}) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(\mathbf{x}_i, \mathbf{v}) l(\mathbf{y}_j, \mathbf{w})$.
- An unbiased estimator of $u(\mathbf{v}, \mathbf{w})$ is

$$\begin{aligned} \hat{u}(\mathbf{v}, \mathbf{w}) &= \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \widehat{\mu_x \mu_y}(\mathbf{v}, \mathbf{w}) \\ &= \frac{2}{n(n-1)} \sum_{i<j} h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)), \end{aligned}$$

where

$$h_{(\mathbf{v},\mathbf{w})}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) := \frac{1}{2}(k(\mathbf{x}, \mathbf{v}) - k(\mathbf{x}', \mathbf{v}))(l(\mathbf{y}, \mathbf{w}) - l(\mathbf{y}', \mathbf{w})).$$

- For a fixed $(\mathbf{v}, \mathbf{w})$, $\hat{u}(\mathbf{v}, \mathbf{w})$ is a one-sample 2^{nd}-order U-statistic.

Asymptotic Distribution of $\hat{\mathbf{u}}$

$$\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J}\sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i) = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}},$$

where $\hat{\mathbf{u}} = \left(\hat{u}(\mathbf{v}_1, \mathbf{w}_1), \ldots, \hat{u}(\mathbf{v}_J, \mathbf{w}_J)\right)^\top$.

Proposition 2 (Asymptotic distribution of $\hat{\mathbf{u}}$).

For any fixed locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$, we have $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(0, \Sigma)$.

- $\Sigma_{ij} = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)] - u(\mathbf{v}_i, \mathbf{w}_i)u(\mathbf{v}_j, \mathbf{w}_j)$,
- $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'}k(\mathbf{x}', \mathbf{v})$,
- $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'}l(\mathbf{y}', \mathbf{w})$.

Under H_0,

$$n\widehat{\mathrm{FSIC}^2} = \frac{n}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}} \sim \text{weighted sum of dependent } \chi^2 \text{ variables.}$$

- Difficult to get $(1-\alpha)$-quantile for the threshold.

Asymptotic Distribution of $\hat{\mathbf{u}}$

$$\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J}\sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i) = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}},$$

where $\hat{\mathbf{u}} = (\hat{u}(\mathbf{v}_1, \mathbf{w}_1), \ldots, \hat{u}(\mathbf{v}_J, \mathbf{w}_J))^\top$.

Proposition 2 (Asymptotic distribution of $\hat{\mathbf{u}}$).

For any fixed locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$, *we have* $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$.

- $\Sigma_{ij} = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)] - u(\mathbf{v}_i, \mathbf{w}_i)u(\mathbf{v}_j, \mathbf{w}_j)$,
- $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'}k(\mathbf{x}', \mathbf{v})$,
- $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'}l(\mathbf{y}', \mathbf{w})$.

Under H_0,

$$n\widehat{\mathrm{FSIC}^2} = \frac{n}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}} \sim \text{weighted sum of dependent } \chi^2 \text{ variables.}$$

- Difficult to get $(1-\alpha)$-quantile for the threshold.

Asymptotic Distribution of $\hat{\mathbf{u}}$

$$\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J}\sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i) = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}},$$

where $\hat{\mathbf{u}} = (\hat{u}(\mathbf{v}_1, \mathbf{w}_1), \ldots, \hat{u}(\mathbf{v}_J, \mathbf{w}_J))^\top$.

Proposition 2 (Asymptotic distribution of $\hat{\mathbf{u}}$).

For any fixed locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$, *we have* $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$.

- $\Sigma_{ij} = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)] - u(\mathbf{v}_i, \mathbf{w}_i)u(\mathbf{v}_j, \mathbf{w}_j)$,
- $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'}k(\mathbf{x}', \mathbf{v})$,
- $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'}l(\mathbf{y}', \mathbf{w})$.

Under H_0,

$$n\widehat{\mathrm{FSIC}^2} = \frac{n}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}} \sim \text{weighted sum of dependent } \chi^2 \text{ variables.}$$

- Difficult to get $(1-\alpha)$-quantile for the threshold.

Asymptotic Distribution of $\hat{\mathbf{u}}$

$$\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J}\sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i) = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}},$$

where $\hat{\mathbf{u}} = (\hat{u}(\mathbf{v}_1, \mathbf{w}_1), \ldots, \hat{u}(\mathbf{v}_J, \mathbf{w}_J))^\top$.

Proposition 2 (Asymptotic distribution of $\hat{\mathbf{u}}$).

For any fixed locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$, *we have* $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$.

- $\Sigma_{ij} = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)] - u(\mathbf{v}_i, \mathbf{w}_i)u(\mathbf{v}_j, \mathbf{w}_j)$,
- $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'}k(\mathbf{x}', \mathbf{v})$,
- $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'}l(\mathbf{y}', \mathbf{w})$.

Under H_0,

$$n\widehat{\mathrm{FSIC}^2} = \frac{n}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}} \sim \text{weighted sum of dependent } \chi^2 \text{ variables.}$$

- Difficult to get $(1-\alpha)$-quantile for the threshold.

Asymptotic Distribution of $\hat{\mathbf{u}}$

$$\widehat{\mathrm{FSIC}^2}(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i) = \frac{1}{J} \hat{\mathbf{u}}^\top \hat{\mathbf{u}},$$

where $\hat{\mathbf{u}} = (\hat{u}(\mathbf{v}_1, \mathbf{w}_1), \ldots, \hat{u}(\mathbf{v}_J, \mathbf{w}_J))^\top$.

Proposition 2 (Asymptotic distribution of $\hat{\mathbf{u}}$).

For any fixed locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$, *we have* $\sqrt{n}(\hat{\mathbf{u}} - \mathbf{u}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$.

- $\Sigma_{ij} = \mathbb{E}_{\mathbf{xy}}[\tilde{k}(\mathbf{x}, \mathbf{v}_i)\tilde{l}(\mathbf{y}, \mathbf{w}_i)\tilde{k}(\mathbf{x}, \mathbf{v}_j)\tilde{l}(\mathbf{y}, \mathbf{w}_j)] - u(\mathbf{v}_i, \mathbf{w}_i)u(\mathbf{v}_j, \mathbf{w}_j)$,
- $\tilde{k}(\mathbf{x}, \mathbf{v}) := k(\mathbf{x}, \mathbf{v}) - \mathbb{E}_{\mathbf{x}'} k(\mathbf{x}', \mathbf{v})$,
- $\tilde{l}(\mathbf{y}, \mathbf{w}) := l(\mathbf{y}, \mathbf{w}) - \mathbb{E}_{\mathbf{y}'} l(\mathbf{y}', \mathbf{w})$.

Under H_0,

$$n\widehat{\mathrm{FSIC}^2} = \frac{n}{J} \hat{\mathbf{u}}^\top \hat{\mathbf{u}} \sim \text{weighted sum of dependent } \chi^2 \text{ variables.}$$

- Difficult to get $(1 - \alpha)$-quantile for the threshold.

Normalized FSIC (NFSIC)

$$\widehat{\mathrm{NFSIC}^2}(X, Y) = \hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

with a regularization parameter $\gamma_n \geq 0$.

- **Key**: NFSIC = FSIC normalized by the covariance.

Theorem 1 (NFSIC test is consistent).

Assume

1. *The product kernel is characteristic and analytic.*
2. $\lim_{n\to\infty} \gamma_n = 0$.

Then, for any k, l and $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$,

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$.*
2. *Under H_1, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) = 1$, η-almost surely.*

Normalized FSIC (NFSIC)

$$\widehat{\mathrm{NFSIC}}^2(X, Y) = \hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

with a regularization parameter $\gamma_n \geq 0$.

- **Key**: NFSIC = FSIC normalized by the covariance.

Theorem 1 (NFSIC test is consistent).

Assume

1. *The product kernel is characteristic and analytic.*
2. $\lim_{n\to\infty} \gamma_n = 0$.

Then, for any k, l and $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$,

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$.*
2. *Under H_1, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) = 1$, η-almost surely.*

Normalized FSIC (NFSIC)

$$\widehat{\mathrm{NFSIC}}^2(X, Y) = \hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

with a regularization parameter $\gamma_n \geq 0$.

- **Key**: NFSIC = FSIC normalized by the covariance.

Theorem 1 (NFSIC test is consistent).

Assume

1. *The product kernel is characteristic and analytic.*
2. $\lim_{n\to\infty} \gamma_n = 0$.

Then, for any k, l and $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$,

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$.*
2. *Under H_1, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) = 1$, η-almost surely.*

Normalized FSIC (NFSIC)

$$\widehat{\mathrm{NFSIC}^2}(X, Y) = \hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\mathbf{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

with a regularization parameter $\gamma_n \geq 0$.

- **Key**: NFSIC = FSIC normalized by the covariance.

Theorem 1 (NFSIC test is consistent).

Assume

1. *The product kernel is characteristic and analytic.*
2. $\lim_{n\to\infty} \gamma_n = 0$.

Then, for any k, l *and* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$,

1. *Under* H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ *as* $n \to \infty$.
2. *Under* H_1, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) = 1$, η*-almost surely.*

Normalized FSIC (NFSIC)

$$\widehat{\mathrm{NFSIC}^2}(X, Y) = \hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

with a regularization parameter $\gamma_n \geq 0$.

- **Key**: NFSIC = FSIC normalized by the covariance.

Theorem 1 (NFSIC test is consistent).

Assume

1. *The product kernel is characteristic and analytic.*
2. $\lim_{n\to\infty} \gamma_n = 0$.

Then, for any k, l *and* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$,

1. *Under* H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ *as* $n \to \infty$.
2. *Under* H_1, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) = 1$, η*-almost surely.*

Normalized FSIC (NFSIC)

$$\widehat{\mathrm{NFSIC}}^2(X, Y) = \hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

with a regularization parameter $\gamma_n \geq 0$.

- **Key**: NFSIC = FSIC normalized by the covariance.

Theorem 1 (NFSIC test is consistent).

Assume

1. *The product kernel is characteristic and analytic.*
2. $\lim_{n\to\infty} \gamma_n = 0$.

Then, for any k, l *and* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$,

1. *Under* H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ *as* $n \to \infty$.
2. *Under* H_1, $\lim_{n\to\infty} \mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) = 1$, η*-almost surely.*

Asymptotically, false positive rate is at α under H_0, and always reject under H_1.

An Estimator of $\widehat{\text{NFSIC}^2}$

$$\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

- Test locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$.
- $\mathbf{K} = [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J\times n}$
- $\mathbf{L} = [l(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J\times n}$. (No $n \times n$ Gram matrix.)

Estimators

1. $\hat{\mathbf{u}} = \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.
2. $\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top}{n}$ where $\boldsymbol{\Gamma} := (\mathbf{K} - n^{-1}\mathbf{K}\mathbf{1}_n\mathbf{1}_n^\top) \circ (\mathbf{L} - n^{-1}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^\top) - \hat{\mathbf{u}}\mathbf{1}_n^\top$.

- $\hat{\lambda}_n$ can be computed in $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$ time.

Main Point: Linear in n. Cubic in J (small).

An Estimator of $\widehat{\text{NFSIC}^2}$

$$\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

- Test locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$.
- $\mathbf{K} = [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J\times n}$
- $\mathbf{L} = [l(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J\times n}$. (No $n \times n$ Gram matrix.)

Estimators

1. $\hat{\mathbf{u}} = \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.
2. $\hat{\Sigma} = \frac{\Gamma\Gamma^\top}{n}$ where $\Gamma := (\mathbf{K} - n^{-1}\mathbf{K}\mathbf{1}_n\mathbf{1}_n^\top) \circ (\mathbf{L} - n^{-1}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^\top) - \hat{\mathbf{u}}\mathbf{1}_n^\top$.

- $\hat{\lambda}_n$ can be computed in $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$ time.

Main Point: Linear in n. Cubic in J (small).

An Estimator of $\widehat{\text{NFSIC}^2}$

$$\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

- Test locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$.
- $\mathbf{K} = [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J\times n}$
- $\mathbf{L} = [l(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J\times n}$. (No $n \times n$ Gram matrix.)

Estimators

1. $\hat{\mathbf{u}} = \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.
2. $\hat{\Sigma} = \frac{\Gamma\Gamma^\top}{n}$ where $\Gamma := (\mathbf{K} - n^{-1}\mathbf{K}\mathbf{1}_n\mathbf{1}_n^\top) \circ (\mathbf{L} - n^{-1}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^\top) - \hat{\mathbf{u}}\mathbf{1}_n^\top$.

- $\hat{\lambda}_n$ can be computed in $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$ time.

Main Point: Linear in n. Cubic in J (small).

An Estimator of $\widehat{\text{NFSIC}^2}$

$$\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

- Test locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$.
- $\mathbf{K} = [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J \times n}$
- $\mathbf{L} = [l(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J \times n}$. (No $n \times n$ Gram matrix.)

Estimators

1. $\hat{\mathbf{u}} = \frac{(\mathbf{K} \circ \mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n) \circ (\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.
2. $\hat{\Sigma} = \frac{\Gamma\Gamma^\top}{n}$ where $\Gamma := (\mathbf{K} - n^{-1}\mathbf{K}\mathbf{1}_n\mathbf{1}_n^\top) \circ (\mathbf{L} - n^{-1}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^\top) - \hat{\mathbf{u}}\mathbf{1}_n^\top$.

- $\hat{\lambda}_n$ can be computed in $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$ time.

Main Point: Linear in n. Cubic in J (small).

An Estimator of $\widehat{\mathrm{NFSIC}^2}$

$$\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}},$$

- Test locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J \sim \eta$.
- $\mathbf{K} = [k(\mathbf{v}_i, \mathbf{x}_j)] \in \mathbb{R}^{J\times n}$
- $\mathbf{L} = [l(\mathbf{w}_i, \mathbf{y}_j)] \in \mathbb{R}^{J\times n}$. (No $n \times n$ Gram matrix.)

Estimators

1. $\hat{\mathbf{u}} = \frac{(\mathbf{K}\circ\mathbf{L})\mathbf{1}_n}{n-1} - \frac{(\mathbf{K}\mathbf{1}_n)\circ(\mathbf{L}\mathbf{1}_n)}{n(n-1)}$.
2. $\hat{\Sigma} = \frac{\Gamma\Gamma^\top}{n}$ where $\Gamma := (\mathbf{K} - n^{-1}\mathbf{K}\mathbf{1}_n\mathbf{1}_n^\top) \circ (\mathbf{L} - n^{-1}\mathbf{L}\mathbf{1}_n\mathbf{1}_n^\top) - \hat{\mathbf{u}}\mathbf{1}_n^\top$.

- $\hat{\lambda}_n$ can be computed in $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$ time.

Main Point: Linear in n. Cubic in J (small).

Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\mathrm{NFSIC}^2}$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

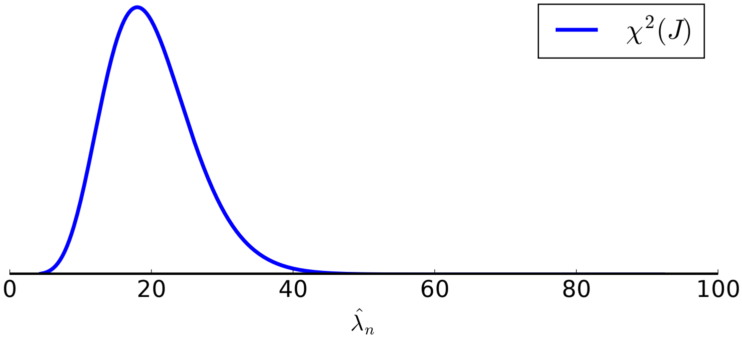
Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\mathrm{NFSIC}^2}$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\mathrm{NFSIC}^2}$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

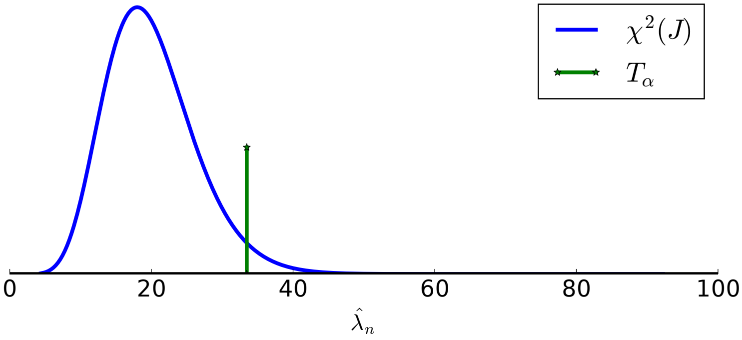
Under $H_0 : X \perp Y$, we have $\hat{\lambda}_n \sim \chi^2(J)$ as $n \to \infty$.



Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\mathrm{NFSIC}^2}$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

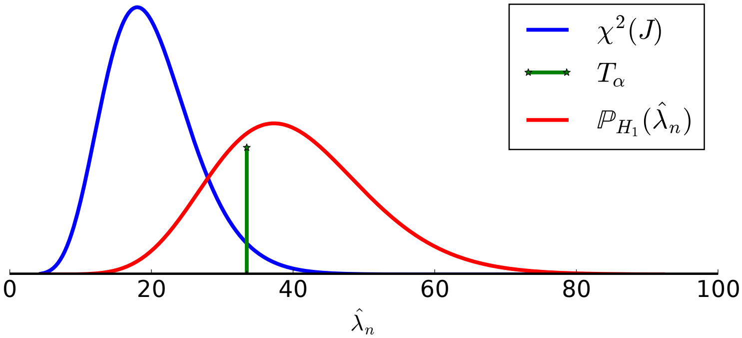
Under $H_0 : X \perp Y$, we have $\hat{\lambda}_n \sim \chi^2(J)$ as $n \to \infty$.



Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\mathrm{NFSIC}}^2$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

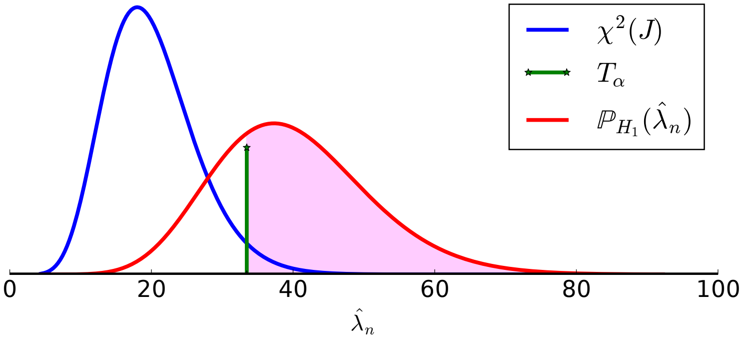
Under H_1, $\hat{\lambda}_n$ will be large. Follows some distribution $\mathbb{P}_{H_1}(\hat{\lambda}_n)$



Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\text{NFSIC}^2}$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

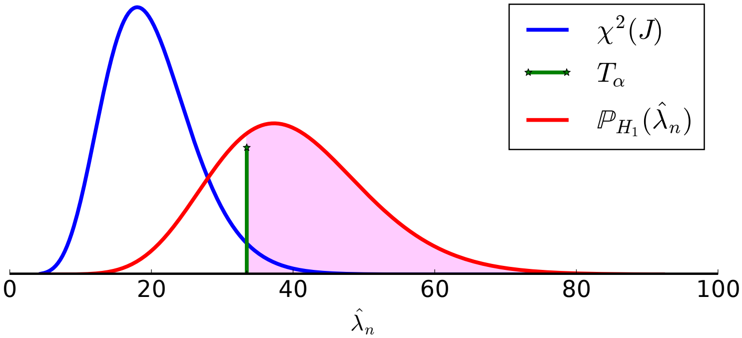
Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\hat{\lambda}_n \geq T_\alpha)$



Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\mathrm{NFSIC}^2}$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\hat{\lambda}_n \geq T_\alpha)$

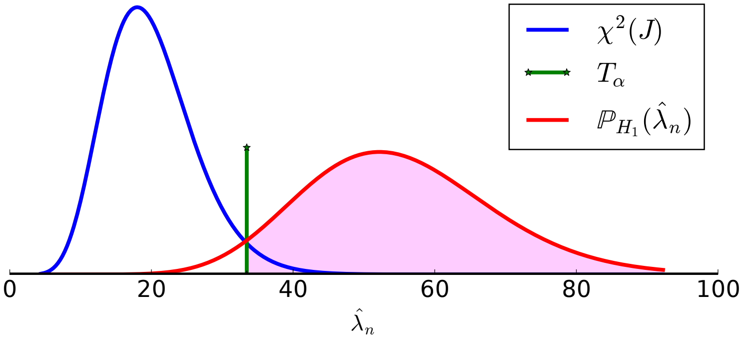


Idea: Pick locations and Gaussian widths to maximize (lower bound of) test power.

Optimizing Test Locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$

- Test $\widehat{\mathrm{NFSIC}^2}$ is consistent for any random locations $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$.
- In practice, tuning them will increase the test power.

Test power $= \mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true}) = \mathbb{P}(\hat{\lambda}_n \geq T_\alpha)$



Idea: Pick locations and Gaussian widths to maximize (lower bound of) test power.

Optimization Objective = Power Lower Bound

- Recall $\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$.

Theorem 2 (A lower bound on the test power).

- *Let* $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \Sigma^{-1}\mathbf{u}$.

With some conditions, for any k, l, *and* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$, *the test power satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *where*

$$L(\lambda_n) = 1 - 62e^{-\xi_1 \gamma_n^2 (\lambda_n - T_\alpha)^2/n} - 2e^{-\lfloor 0.5n \rfloor (\lambda_n - T_\alpha)^2/\left[\xi_2 n^2\right]}$$
$$- 2e^{-\left[(\lambda_n - T_\alpha)\gamma_n(n-1)/3 - \xi_3 n - c_3\gamma_n^2 n(n-1)\right]^2/\left[\xi_4 n^2 (n-1)\right]},$$

where $\xi_1, \ldots, \xi_4, c_3 > 0$ *are constants. For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Do: Locations and Gaussian widths = $\arg\max L(\lambda_n) = \arg\max \lambda_n$

Optimization Objective = Power Lower Bound

- Recall $\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\boldsymbol{\Sigma}} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$.

Theorem 2 (A lower bound on the test power).

- *Let* $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \boldsymbol{\Sigma}^{-1}\mathbf{u}$.

With some conditions, for any k, l, *and* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$, *the test power satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *where*

$$L(\lambda_n) = 1 - 62e^{-\xi_1\gamma_n^2(\lambda_n - T_\alpha)^2/n} - 2e^{-\lfloor 0.5n \rfloor(\lambda_n - T_\alpha)^2/\left[\xi_2 n^2\right]} - 2e^{-\left[(\lambda_n - T_\alpha)\gamma_n(n-1)/3 - \xi_3 n - c_3\gamma_n^2 n(n-1)\right]^2/\left[\xi_4 n^2(n-1)\right]},$$

where $\xi_1, \ldots, \xi_4, c_3 > 0$ *are constants. For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Do: Locations and Gaussian widths $= \arg\max L(\lambda_n) = \arg\max \lambda_n$

Optimization Objective = Power Lower Bound

- Recall $\hat{\lambda}_n := n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$.

Theorem 2 (A lower bound on the test power).

- *Let* $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \Sigma^{-1}\mathbf{u}$.

With some conditions, for any k, l, *and* $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$, *the test power satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *where*

$$L(\lambda_n) = 1 - 62e^{-\xi_1 \gamma_n^2 (\lambda_n - T_\alpha)^2/n} - 2e^{-\lfloor 0.5n \rfloor (\lambda_n - T_\alpha)^2 / \left[\xi_2 n^2\right]}$$
$$- 2e^{-\left[(\lambda_n - T_\alpha)\gamma_n (n-1)/3 - \xi_3 n - c_3 \gamma_n^2 n(n-1)\right]^2 / \left[\xi_4 n^2 (n-1)\right]},$$

where $\xi_1, \ldots, \xi_4, c_3 > 0$ *are constants. For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Do: Locations and Gaussian widths = $\arg\max L(\lambda_n) = \arg\max \lambda_n$

Optimization Objective = Power Lower Bound

- Recall $\hat{\lambda}_n := n\hat{\mathbf{u}}^\top\left(\hat{\Sigma}+\gamma_n\mathbf{I}\right)^{-1}\hat{\mathbf{u}}$.

Theorem 2 (A lower bound on the test power).

- *Let* $\mathrm{NFSIC}^2(X,Y) := \lambda_n := n\mathbf{u}^\top\Sigma^{-1}\mathbf{u}$.

With some conditions, for any k, l, *and* $\{(\mathbf{v}_i,\mathbf{w}_i)\}_{i=1}^J$, *the test power satisfies* $\mathbb{P}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *where*

$$L(\lambda_n) = 1 - 62e^{-\xi_1\gamma_n^2(\lambda_n - T_\alpha)^2/n} - 2e^{-\lfloor 0.5n\rfloor(\lambda_n-T_\alpha)^2/\left[\xi_2 n^2\right]} - 2e^{-\left[(\lambda_n-T_\alpha)\gamma_n(n-1)/3-\xi_3 n - c_3\gamma_n^2 n(n-1)\right]^2/\left[\xi_4 n^2(n-1)\right]},$$

where $\xi_1,\ldots,\xi_4, c_3 > 0$ *are constants. For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Do: Locations and Gaussian widths $= \arg\max L(\lambda_n) = \arg\max \lambda_n$

Optimization Procedure

- $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \Sigma^{-1}\mathbf{u}$ is unknown.
- Split the data into 2 disjoint sets: training (tr) and test (te) sets.

Procedure:

1. Estimate λ_n with $\hat{\lambda}_n^{(\mathrm{tr})}$ (i.e., computed on the training set).
2. Optimize all $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ and Gaussian widths with gradient ascent.
3. Independence test with $\hat{\lambda}_n^{(\mathrm{te})}$. Reject H_0 if $\hat{\lambda}_n^{(\mathrm{te})} \geq T_\alpha$.

- Splitting avoids overfitting.

Optimization Procedure

- $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top\Sigma^{-1}\mathbf{u}$ is unknown.
- Split the data into 2 disjoint sets: training (tr) and test (te) sets.

Procedure:

1. Estimate λ_n with $\hat{\lambda}_n^{(\mathrm{tr})}$ (i.e., computed on the training set).
2. Optimize all $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ and Gaussian widths with gradient ascent.
3. Independence test with $\hat{\lambda}_n^{(\mathrm{te})}$. Reject H_0 if $\hat{\lambda}_n^{(\mathrm{te})} \geq T_\alpha$.

- Splitting avoids overfitting.

Optimization Procedure

- $\text{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \Sigma^{-1}\mathbf{u}$ is unknown.
- Split the data into 2 disjoint sets: training (tr) and test (te) sets.

Procedure:

1. Estimate λ_n with $\hat{\lambda}_n^{(\mathrm{tr})}$ (i.e., computed on the training set).
2. Optimize all $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ and Gaussian widths with gradient ascent.
3. Independence test with $\hat{\lambda}_n^{(\mathrm{te})}$. Reject H_0 if $\hat{\lambda}_n^{(\mathrm{te})} \geq T_\alpha$.

- Splitting avoids overfitting.

Optimization Procedure

- $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \Sigma^{-1}\mathbf{u}$ is unknown.
- Split the data into 2 disjoint sets: training (tr) and test (te) sets.

Procedure:

1. Estimate λ_n with $\hat{\lambda}_n^{(\mathsf{tr})}$ (i.e., computed on the training set).
2. Optimize all $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ and Gaussian widths with gradient ascent.
3. Independence test with $\hat{\lambda}_n^{(\mathsf{te})}$. Reject H_0 if $\hat{\lambda}_n^{(\mathsf{te})} \geq T_\alpha$.

- Splitting avoids overfitting.

But, what does this do to $\mathbb{P}(\hat{\lambda}_n \geq T_\alpha)$ when H_0 holds?

Optimization Procedure

- $\mathrm{NFSIC}^2(X, Y) := \lambda_n := n\mathbf{u}^\top \Sigma^{-1}\mathbf{u}$ is unknown.
- Split the data into 2 disjoint sets: training (tr) and test (te) sets.

Procedure:

1. Estimate λ_n with $\hat{\lambda}_n^{(\mathrm{tr})}$ (i.e., computed on the training set).
2. Optimize all $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ and Gaussian widths with gradient ascent.
3. Independence test with $\hat{\lambda}_n^{(\mathrm{te})}$. Reject H_0 if $\hat{\lambda}_n^{(\mathrm{te})} \geq T_\alpha$.

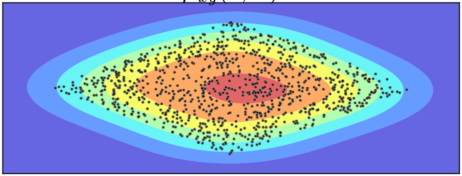
- Splitting avoids overfitting.

But, what does this do to $\mathbb{P}(\hat{\lambda}_n \geq T_\alpha)$ when H_0 holds?

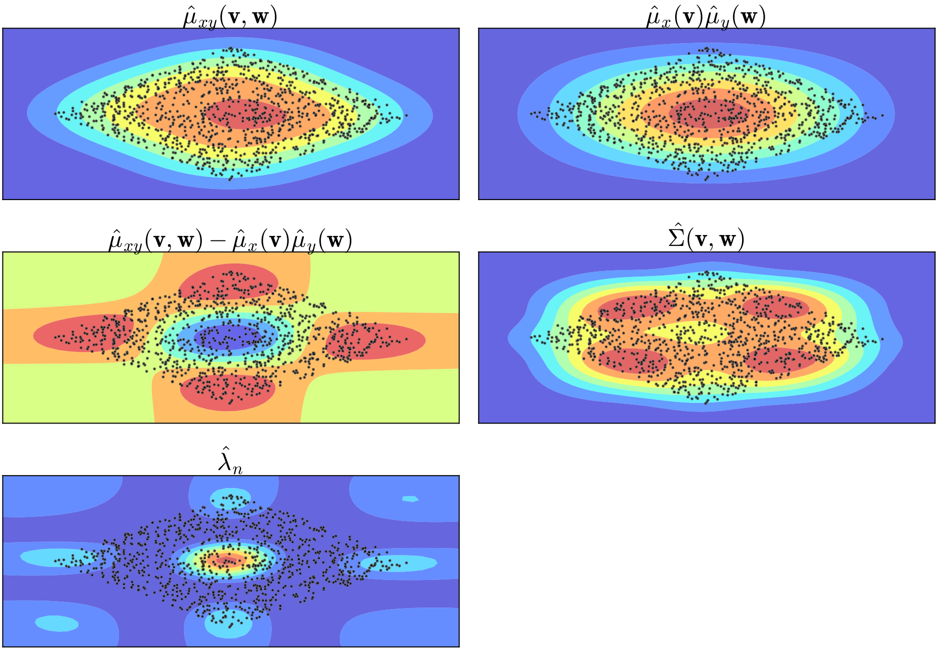
- Still asymptotically at α.
- $\lambda_n = 0$ iff X, Y independent.
- So, under H_0, we do $\arg\max 0 =$ arbitrary locations.
- Asymptotic null distribution is $\chi^2(J)$ for any locations.

Demo: 2D Rotation

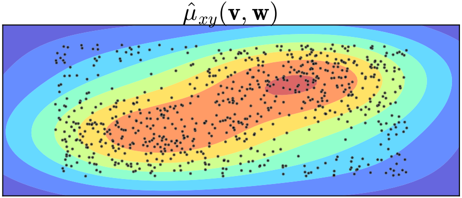
$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$



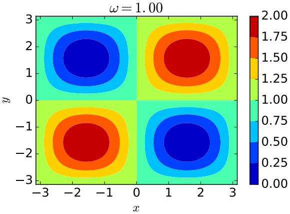
Demo: 2D Rotation



Demo: Sin Problem ($\omega = 1$)

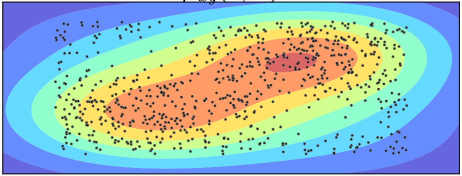


$p(\mathbf{x}, \mathbf{y}) =$

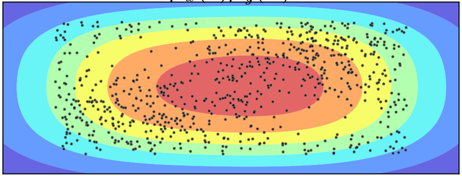


Demo: Sin Problem ($\omega = 1$)

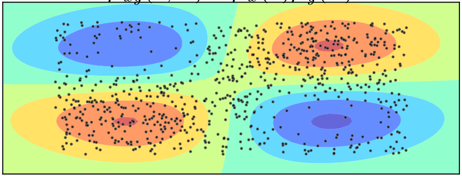
$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w})$



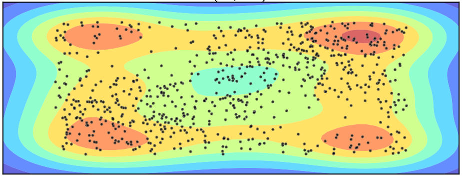
$\hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$



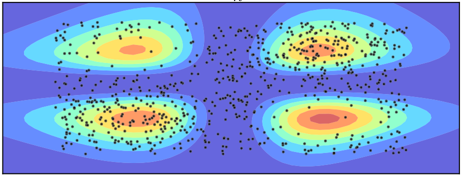
$\hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w})$



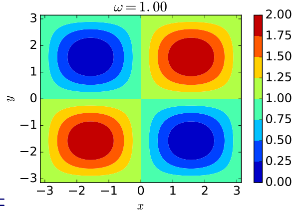
$\hat{\Sigma}(\mathbf{v}, \mathbf{w})$



$\hat{\lambda}_n$



$p(\mathbf{x}, \mathbf{y}) =$



Simulation Settings

- n = full sample size
- All methods use Gaussian kernels for both X and Y.

Compare 6 methods

Method	Description	Tuning	Test size	Complex.
NFSIC-opt	Proposed	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning.	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heu.	n	$\mathcal{O}(n^2)$
NyHSIC	NyStrom HSIC	Median heu.	n	$\mathcal{O}(n)$
FHSIC	HSIC + RFFs*	Median heu.	n	$\mathcal{O}(n)$
RDC	RFFs + CCA	Median heu.	n	$\mathcal{O}(n \log n)$

* : Random Fourier features

- Given a problem, report rejection rate of H_0.
- 10 features for all (except QHSIC). $J = 10$ in NFSIC.

Simulation Settings

- n = full sample size
- All methods use Gaussian kernels for both X and Y.

Compare 6 methods

Method	Description	Tuning	Test size	Complex.
NFSIC-opt	Proposed	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning.	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heu.	n	$\mathcal{O}(n^2)$
NyHSIC	NyStrom HSIC	Median heu.	n	$\mathcal{O}(n)$
FHSIC	HSIC + RFFs*	Median heu.	n	$\mathcal{O}(n)$
RDC	RFFs + CCA	Median heu.	n	$\mathcal{O}(n \log n)$

* : Random Fourier features

- Given a problem, report rejection rate of H_0.
- 10 features for all (except QHSIC). $J = 10$ in NFSIC.

Simulation Settings

- $n =$ full sample size
- All methods use Gaussian kernels for both X and Y.

Compare 6 methods

Method	Description	Tuning	Test size	Complex.
NFSIC-opt	Proposed	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning.	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heu.	n	$\mathcal{O}(n^2)$
NyHSIC	NyStrom HSIC	Median heu.	n	$\mathcal{O}(n)$
FHSIC	HSIC + RFFs*	Median heu.	n	$\mathcal{O}(n)$
RDC	RFFs + CCA	Median heu.	n	$\mathcal{O}(n \log n)$

* : Random Fourier features

- Given a problem, report rejection rate of H_0.
- 10 features for all (except QHSIC). $J = 10$ in NFSIC.

Simulation Settings

- $n =$ full sample size
- All methods use Gaussian kernels for both X and Y.

Compare 6 methods

Method	Description	Tuning	Test size	Complex.
NFSIC-opt	Proposed	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning.	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heu.	n	$\mathcal{O}(n^2)$
NyHSIC	NyStrom HSIC	Median heu.	n	$\mathcal{O}(n)$
FHSIC	HSIC + RFFs*	Median heu.	n	$\mathcal{O}(n)$
RDC	RFFs + CCA	Median heu.	n	$\mathcal{O}(n \log n)$

* : Random Fourier features

- Given a problem, report rejection rate of H_0.
- 10 features for all (except QHSIC). $J = 10$ in NFSIC.

Simulation Settings

- n = full sample size
- All methods use Gaussian kernels for both X and Y.

Compare 6 methods

Method	Description	Tuning	Test size	Complex.
NFSIC-opt	Proposed	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning.	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heu.	n	$\mathcal{O}(n^2)$
NyHSIC	NyStrom HSIC	Median heu.	n	$\mathcal{O}(n)$
FHSIC	HSIC + RFFs*	Median heu.	n	$\mathcal{O}(n)$
RDC	RFFs + CCA	Median heu.	n	$\mathcal{O}(n \log n)$

* : Random Fourier features

- Given a problem, report rejection rate of H_0.
- 10 features for all (except QHSIC). $J = 10$ in NFSIC.

Simulation Settings

- n = full sample size
- All methods use Gaussian kernels for both X and Y.

Compare 6 methods

Method	Description	Tuning	Test size	Complex.
NFSIC-opt	Proposed	Gradient descent	$n/2$	$\mathcal{O}(n)$
NFSIC-med	No tuning.	Random locations	n	$\mathcal{O}(n)$
QHSIC	Full HSIC	Median heu.	n	$\mathcal{O}(n^2)$
NyHSIC	NyStrom HSIC	Median heu.	n	$\mathcal{O}(n)$
FHSIC	HSIC + RFFs*	Median heu.	n	$\mathcal{O}(n)$
RDC	RFFs + CCA	Median heu.	n	$\mathcal{O}(n \log n)$

* : Random Fourier features

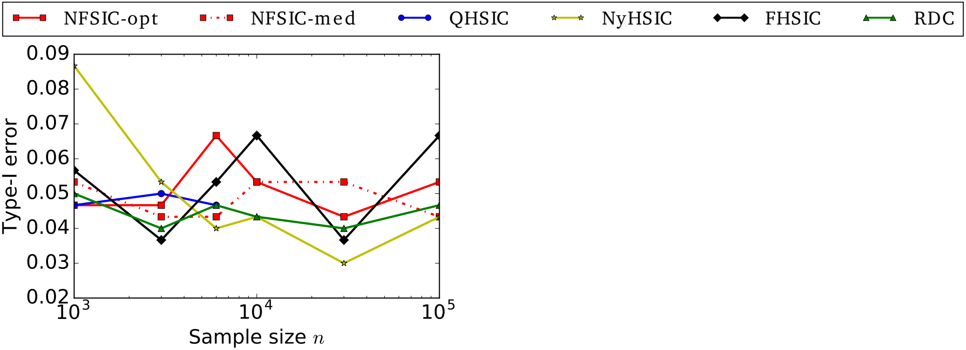
- Given a problem, report rejection rate of H_0.
- 10 features for all (except QHSIC). $J = 10$ in NFSIC.

Toy Problem 1: Independent Gaussians

- $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ and $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{dy})$.
- Independent X, Y. So, H_0 holds.
- Set $\alpha := 0.05$, $d_x = d_y = 250$.

Toy Problem 1: Independent Gaussians

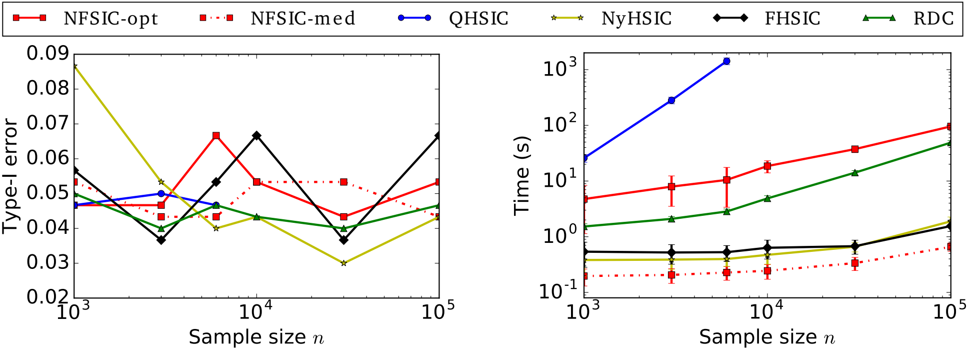
- $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ and $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{dy})$.
- Independent X, Y. So, H_0 holds.
- Set $\alpha := 0.05$, $d_x = d_y = 250$.



- Correct type-I errors (false positive rate).

Toy Problem 1: Independent Gaussians

- $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ and $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{dy})$.
- Independent X, Y. So, H_0 holds.
- Set $\alpha := 0.05$, $d_x = d_y = 250$.



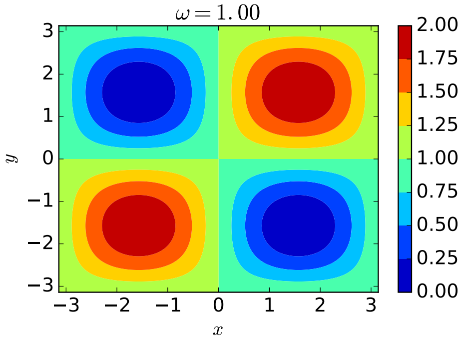
- Correct type-I errors (false positive rate).

Toy Problem 2: Sinusoid

- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.

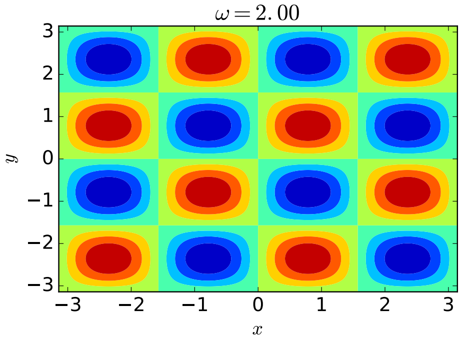
Toy Problem 2: Sinusoid

- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.



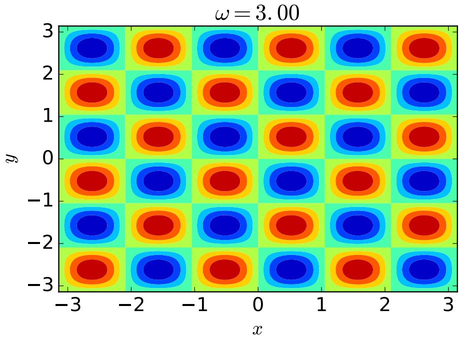
Toy Problem 2: Sinusoid

- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.



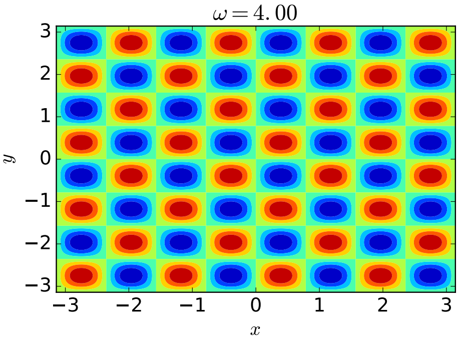
Toy Problem 2: Sinusoid

- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.



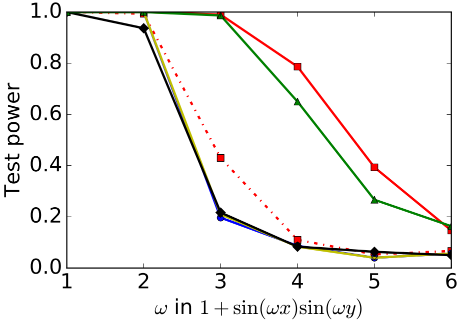
Toy Problem 2: Sinusoid

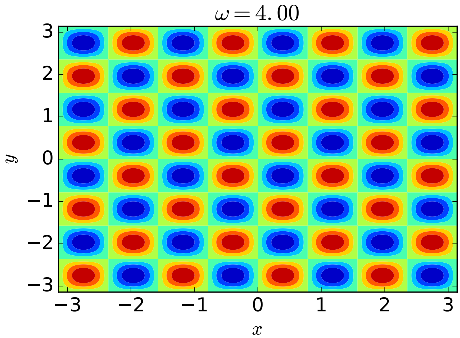
- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.



Toy Problem 2: Sinusoid

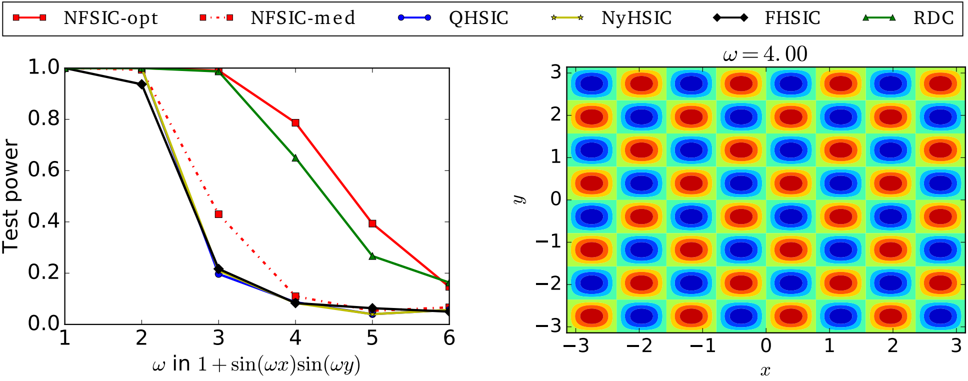
- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.





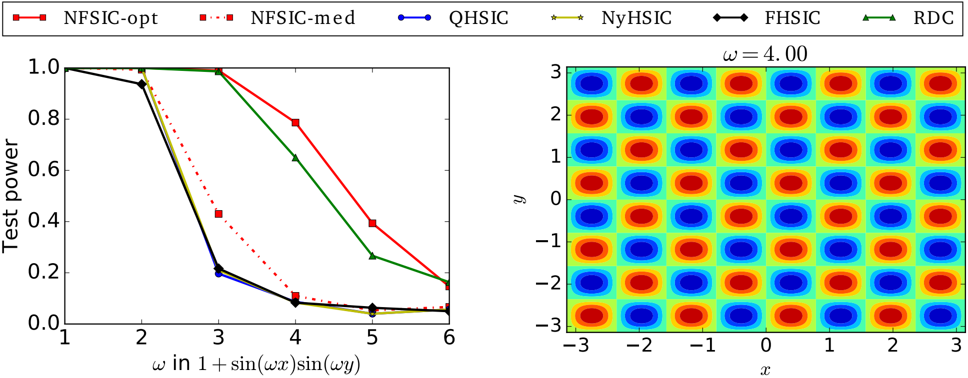
Toy Problem 2: Sinusoid

- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.



Toy Problem 2: Sinusoid

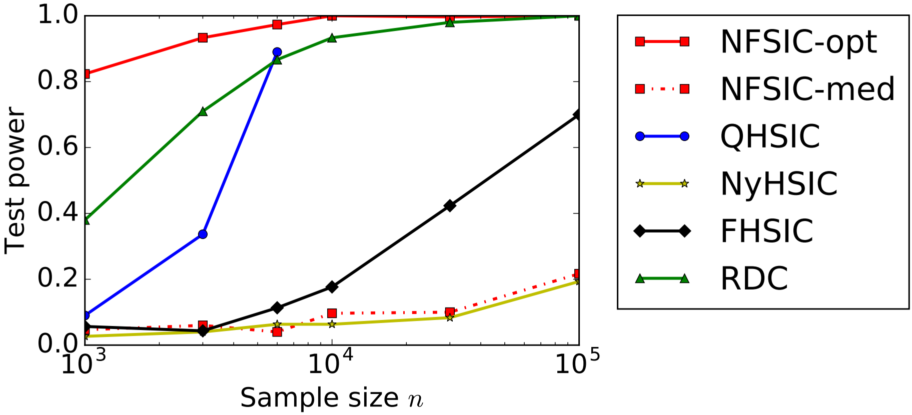
- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.



Main Point: NFSIC can handle well the local changes in the joint space.

Toy Problem 3: Gaussian Sign

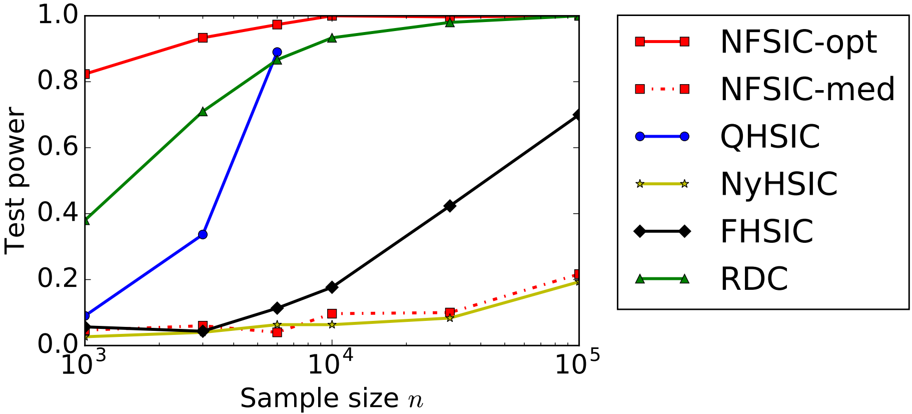
- $y = |Z| \prod_{i=1}^{d_x} \text{sign}(x_i)$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ and $Z \sim \mathcal{N}(0,1)$ (noise).
- Full interaction among $x_1, \ldots, x_{d_x}$.
- Need to consider all $x_1, \ldots, x_d$ to detect the dependency.



Main Point: NFSIC can handle feature interaction.

Toy Problem 3: Gaussian Sign

- $y = |Z| \prod_{i=1}^{d_x} \mathrm{sign}(x_i)$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ and $Z \sim \mathcal{N}(0, 1)$ (noise).
- Full interaction among $x_1, \ldots, x_{d_x}$.
- Need to consider all $x_1, \ldots, x_d$ to detect the dependency.



Main Point: NFSIC can handle feature interaction.

HSIC vs. FSIC

Recall the witness

$$\hat{u}(\mathbf{v}, \mathbf{w}) = \hat{\mu}_{xy}(\mathbf{v}, \mathbf{w}) - \hat{\mu}_x(\mathbf{v})\hat{\mu}_y(\mathbf{w}).$$

HSIC [Gretton et al., 2005]

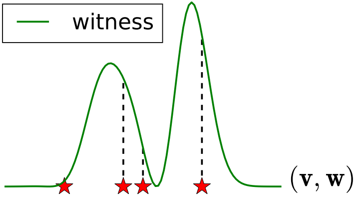
$= \|\hat{u}\|_{\mathrm{RKHS}}$

Good when difference between p_{xy} and $p_x p_y$ is spatially diffuse.

- $\hat{u}$ is almost flat.

FSIC [proposed]

$= \frac{1}{J}\sum_{i=1}^{J} \hat{u}^2(\mathbf{v}_i, \mathbf{w}_i)$



Good when difference between p_{xy} and $p_x p_y$ is local.

- $\hat{u}$ is mostly zero, has many peaks (feature interaction).

Real Problem 1: Million Song Data

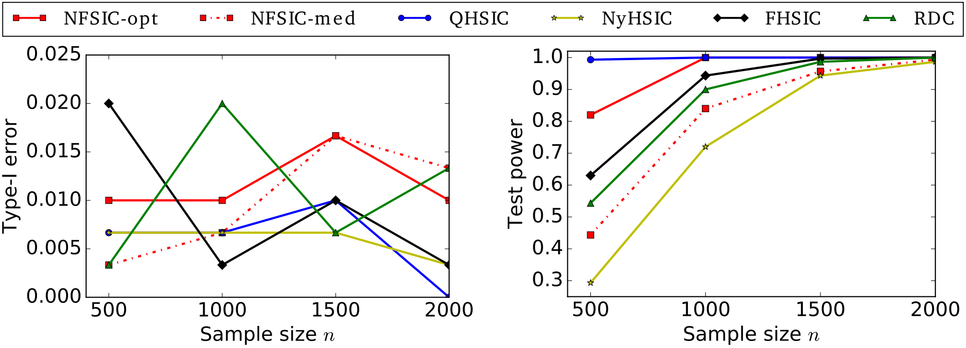
Song (X) vs. year of release (Y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $X \in \mathbb{R}^{90}$ contains audio features.
- $Y \in \mathbb{R}$ is the year of release.

Real Problem 1: Million Song Data

Song (X) vs. year of release (Y).

- Western commercial tracks from 1922 to 2011 [Bertin-Mahieux et al., 2011].
- $X \in \mathbb{R}^{90}$ contains audio features.
- $Y \in \mathbb{R}$ is the year of release.



- Break (X, Y) pairs to simulate H_0.
- H_1 is true.

Real Problem 2: Videos and Captions

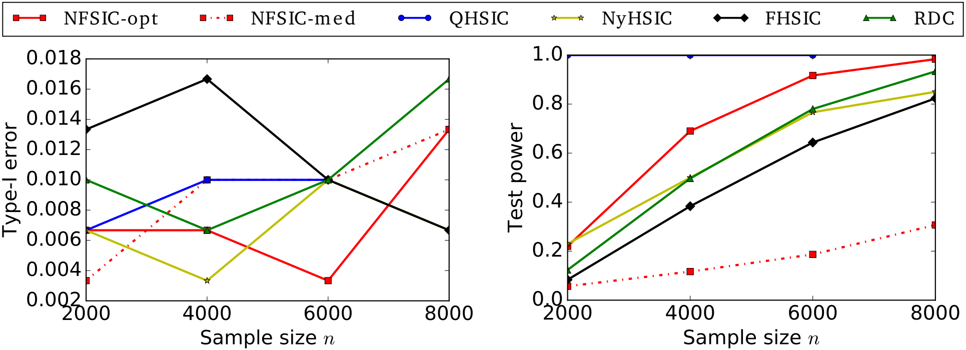
Youtube video (X) vs. caption (Y).

- VideoStory46K [Habibian et al., 2014]
- $X \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms descriptors [Wang and Schmid, 2013].
- $Y \in \mathbb{R}^{1878}$: bag of words. TF.

Real Problem 2: Videos and Captions

Youtube video (X) vs. caption (Y).

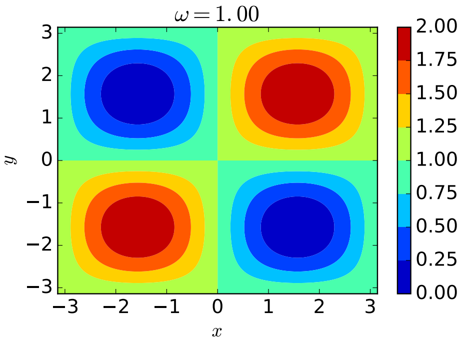
- VideoStory46K [Habibian et al., 2014]
- $X \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms descriptors [Wang and Schmid, 2013].
- $Y \in \mathbb{R}^{1878}$: bag of words. TF.

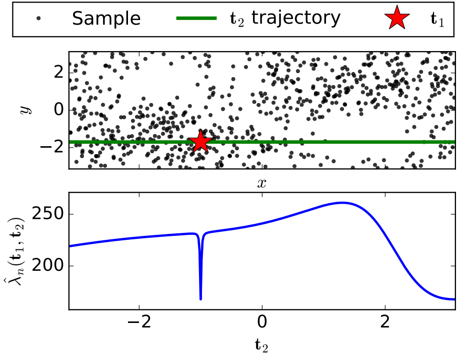


- Break (X, Y) pairs to simulate H_0.
- H_1 is true.

Penalize Redundant Test Locations

- Consider the Sin problem. Use $J = 2$ locations.
- Optimization objective: $\hat{\lambda}_n$.
- Write $\mathbf{t} = (\mathbf{v}, \mathbf{w})$. Fix $\mathbf{t}_1$ at ★. Plot $\mathbf{t}_2 \to \hat{\lambda}_n(\mathbf{t}_1, \mathbf{t}_2)$.

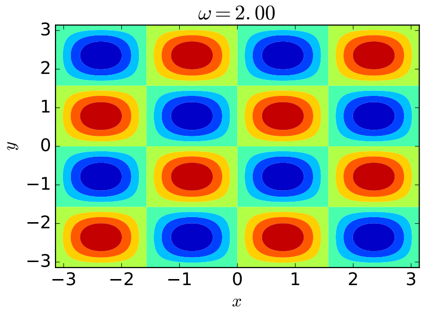


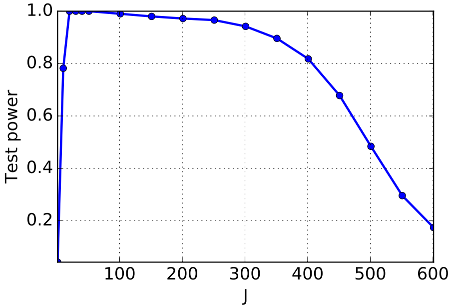


- The optimized $\mathbf{t}_1, \mathbf{t}_2$ will not be in the same neighbourhood.

Test Power vs. J

- Test power *does not* always increase with J (number of test locations).
- $n = 800$.





- Accurate estimation of $\hat{\Sigma} \in \mathbb{R}^{J\times J}$ in $\hat{\lambda}_n = n\hat{\mathbf{u}}^\top \left(\hat{\Sigma} + \gamma_n \mathbf{I}\right)^{-1} \hat{\mathbf{u}}$ becomes more difficult.
- Large J defeats the purpose of a linear-time test.

Conclusions

- Proposed The Finite Set Independence Criterion (**FSIC**).
- Independece test based on FSIC is
  1. non-parametric,
  2. linear-time,
  3. adaptive (parameteris automatically tuned).

Future works

- Any way to interpret the learned $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$?
- Relative efficiency of FSIC vs. block HSIC, RFF-HSIC.

https://github.com/wittawatj/fsic-test

Conclusions

- Proposed The Finite Set Independence Criterion (**FSIC**).
- Independece test based on FSIC is
  1. non-parametric,
  2. linear-time,
  3. adaptive (parameteris automatically tuned).

Future works

- Any way to interpret the learned $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$?
- Relative efficiency of FSIC vs. block HSIC, RFF-HSIC.

`https://github.com/wittawatj/fsic-test`

Questions?

Thank you

References I

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011).
The million song dataset.
In *International Conference on Music Information Retrieval (ISMIR)*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012).
A Kernel Two-Sample Test.
Journal of Machine Learning Research, 13:723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005).
Measuring Statistical Dependence with Hilbert-Schmidt Norms.
In *Algorithmic Learning Theory (ALT)*, pages 63–77.

References II

Habibian, A., Mensink, T., and Snoek, C. G. (2014).
Videostory: A new multimedia embedding for few-example recognition and translation of events.
In *ACM International Conference on Multimedia*, pages 17–26.

Wang, H. and Schmid, C. (2013).
Action recognition with improved trajectories.
In *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2016).
Large-Scale Kernel Methods for Independence Testing.