The Finite-Set Independence Criterion (FSIC)

Wittawat Jitkrittum Zoltán Szabó Arthur Gretton

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

3rd UCL Workshop on the Theory of Big Data

28 June 2017

What Is Independence Testing?

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

$$H_0 : P_{xy} = P_x P_y, \quad \text{vs.} \quad H_1 : P_{xy} \neq P_x P_y.$$

- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n > T_\alpha$ (threshold).
- $T_\alpha = (1 - \alpha)$-quantile of the null distribution.

What Is Independence Testing?

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

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

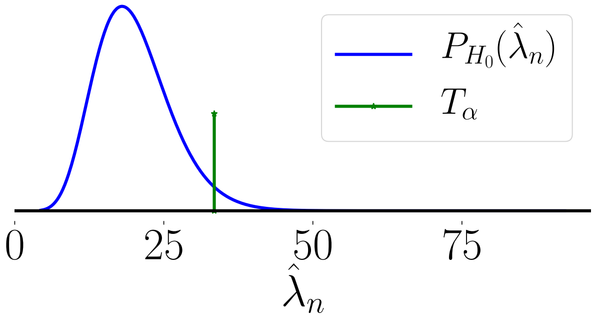
- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n > T_\alpha$ (threshold).
- $T_\alpha = (1-\alpha)$-quantile of the null distribution.

What Is Independence Testing?

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

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

- Compute a test statistic $\hat{\lambda}_n$. Reject H_0 if $\hat{\lambda}_n > T_\alpha$ (threshold).
- $T_\alpha = (1 - \alpha)$-quantile of the null distribution.



Motivations

Modern state-of-the-art test is HSIC [Gretton et al., 2005].

- ✓ Nonparametric i.e., no assumption on P_{xy}. Kernel-based.
- ✗ Slow. Runtime: $\mathcal{O}(n^2)$ where $n =$ sample size.
- ✗ No systematic way to choose kernels.

Propose the Finite-Set Independence Criterion (FSIC).

1. Nonparametric.
2. Linear-time. Runtime complexity: $\mathcal{O}(n)$. Fast.
3. Tunable i.e., well-defined criterion for parameter tuning.

Motivations

Modern state-of-the-art test is HSIC [Gretton et al., 2005].

- ✓ Nonparametric i.e., no assumption on P_{xy}. Kernel-based.
- ✗ Slow. Runtime: $\mathcal{O}(n^2)$ where $n =$ sample size.
- ✗ No systematic way to choose kernels.

Propose the Finite-Set Independence Criterion (FSIC).

1. Nonparametric.
2. Linear-time. Runtime complexity: $\mathcal{O}(n)$. Fast.
3. Tunable i.e., well-defined criterion for parameter tuning.

Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y}) \sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}) \right].$$

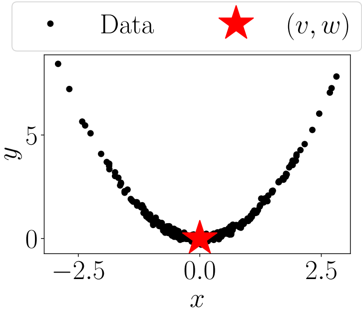
Proposal: The Finite-Set Independence Criterion (FSIC)

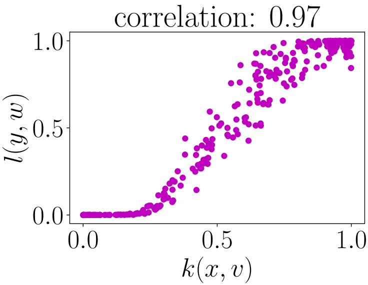
1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$





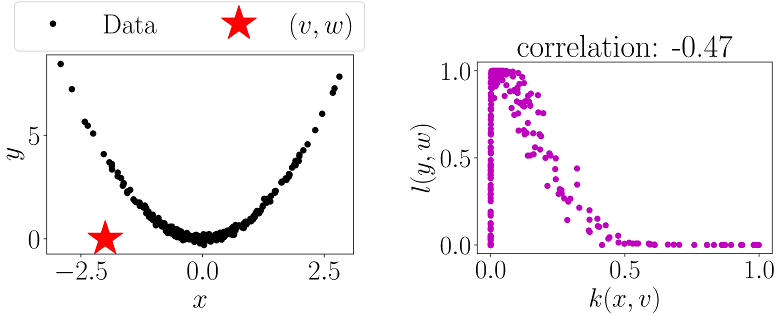
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$



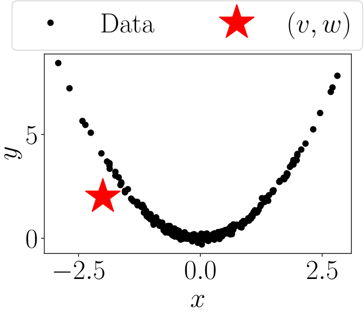
Proposal: The Finite-Set Independence Criterion (FSIC)

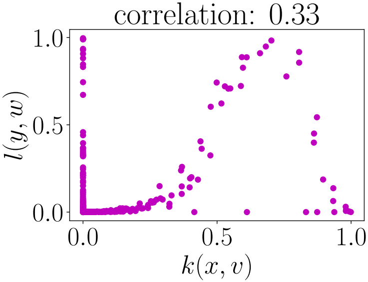
1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance

$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$



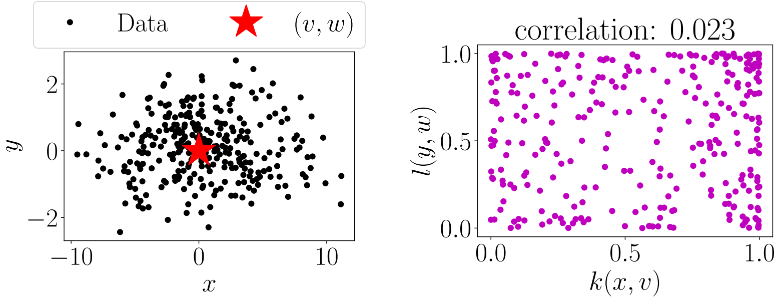


Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance
   $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$

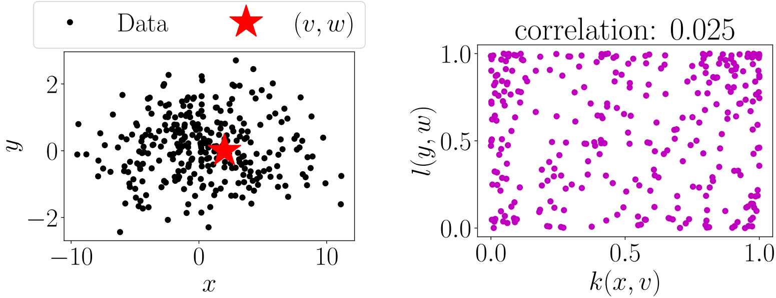


Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x}, \mathbf{y}) \mapsto (k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w}))$ then measure covariance
$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}), l(\mathbf{y}, \mathbf{w})\right].$$



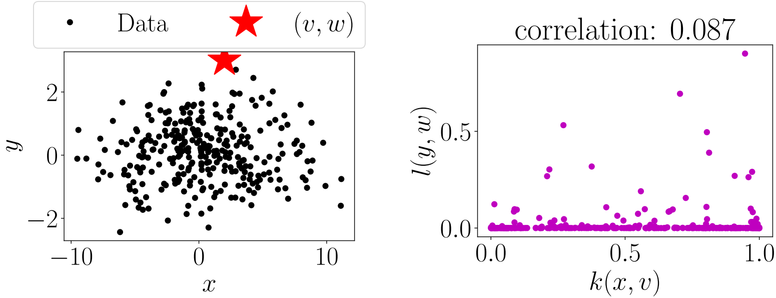
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: k for X, and l for Y.
   - Gaussian kernel: $k(\mathbf{x},\mathbf{v}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{v}\|^2}{2\sigma_x^2}\right)$.
2. Pick some **feature** $(\mathbf{v},\mathbf{w}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(\mathbf{x},\mathbf{y}) \mapsto (k(\mathbf{x},\mathbf{v}), l(\mathbf{y},\mathbf{w}))$ then measure covariance

$$\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R} \times \mathbb{R}$$

$$\mathrm{FSIC}^2(X, Y) = \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x},\mathbf{v}), l(\mathbf{y},\mathbf{w})\right].$$



General Form of FSIC

$$\text{FSIC}^2(X, Y) = \frac{1}{J}\sum_{j=1}^{J} \text{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}_j), l(\mathbf{y}, \mathbf{w}_j)\right],$$

for J **features** $\{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^{J} \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$.

Proposition 1.

Assume

1. *Kernels k and l satisfy some conditions (e.g. Gaussian kernels).*
2. *Features $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$ are drawn from a distribution with a density.*

Then, for any $J \geq 1$,

$$\text{FSIC}(X, Y) = 0 \text{ if and only if } X \text{ and } Y \text{ are independent}$$

Under $H_0 : P_{xy} = P_x P_y$,

$$n\widehat{\text{FSIC}^2} \sim \text{weighted sum of } J \text{ dependent } \chi^2 \text{ variables.}$$

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

General Form of FSIC

$$\mathrm{FSIC}^2(X, Y) = \frac{1}{J}\sum_{j=1}^{J} \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}}\left[k(\mathbf{x}, \mathbf{v}_j), l(\mathbf{y}, \mathbf{w}_j)\right],$$

for J **features** $\{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^{J} \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$.

Proposition 1.

Assume

1. *Kernels k and l satisfy some conditions (e.g. Gaussian kernels).*
2. *Features $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$ are drawn from a distribution with a density.*

Then, for any $J \geq 1$,

$\mathrm{FSIC}(X, Y) = 0$ *if and only if X and Y are independent*

Under $H_0 : P_{xy} = P_x P_y$,

$$n\widehat{\mathrm{FSIC}^2} \sim \text{weighted sum of } J \text{ dependent } \chi^2 \text{ variables.}$$

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

General Form of FSIC

$$\mathrm{FSIC}^2(X, Y) = \frac{1}{J}\sum_{j=1}^{J} \mathrm{cov}^2_{(\mathbf{x},\mathbf{y})\sim P_{xy}} \left[k(\mathbf{x}, \mathbf{v}_j), l(\mathbf{y}, \mathbf{w}_j)\right],$$

for J **features** $\{(\mathbf{v}_j, \mathbf{w}_j)\}_{j=1}^{J} \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$.

Proposition 1.

Assume

1. *Kernels k and l satisfy some conditions (e.g. Gaussian kernels).*
2. *Features $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$ are drawn from a distribution with a density.*

Then, for any $J \geq 1$,

$\mathrm{FSIC}(X, Y) = 0$ *if and only if X and Y are independent*

Under $H_0 : P_{xy} = P_x P_y$,

$$n\widehat{\mathrm{FSIC}^2} \sim \text{weighted sum of } J \text{ dependent } \chi^2 \text{ variables.}$$

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

Normalized FSIC (NFSIC)

- Let $\hat{\mathbf{u}} := \left(\widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_1), l(\mathbf{y}, \mathbf{w}_1)], \ldots, \widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_J), l(\mathbf{y}, \mathbf{w}_J)]\right)^\top \in \mathbb{R}^J$.
- Then, $\widehat{\mathrm{FSIC}^2} = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$.

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

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

- $\hat{\Sigma}_{ij}$ = covariance of $\hat{u}_i$ and $\hat{u}_j$.

Theorem 1 (NFSIC test is consistent).

Assume $\gamma_n \to 0$, and same conditions on k and l as before.

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$. Easy to get threshold T_α.*
2. *Under H_1, $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

- Complexity: $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$. Only need small J.

Normalized FSIC (NFSIC)

- Let $\hat{\mathbf{u}} := \left(\widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_1), l(\mathbf{y}, \mathbf{w}_1)], \ldots, \widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_J), l(\mathbf{y}, \mathbf{w}_J)]\right)^\top \in \mathbb{R}^J$.
- Then, $\widehat{\mathrm{FSIC}^2} = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$.

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

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

- $\hat{\Sigma}_{ij}$ = covariance of $\hat{u}_i$ and $\hat{u}_j$.

Theorem 1 (NFSIC test is consistent).

Assume $\gamma_n \to 0$, and same conditions on k and l as before.

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$. Easy to get threshold T_α.*
2. *Under H_1, $\mathbb{P}(reject\ H_0) \to 1$ as $n \to \infty$.*

- Complexity: $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$. Only need small J.

Normalized FSIC (NFSIC)

- Let $\hat{\mathbf{u}} := \left(\widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_1), l(\mathbf{y}, \mathbf{w}_1)], \ldots, \widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_J), l(\mathbf{y}, \mathbf{w}_J)]\right)^\top \in \mathbb{R}^J$.
- Then, $\widehat{\mathrm{FSIC}^2} = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$.

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

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

- $\hat{\Sigma}_{ij}$ = covariance of $\hat{u}_i$ and $\hat{u}_j$.

Theorem 1 (NFSIC test is consistent).

Assume $\gamma_n \to 0$, and same conditions on k and l as before.

1. *Under H_0, $\hat{\lambda}_n \stackrel{d}{\to} \chi^2(J)$ as $n \to \infty$. Easy to get threshold T_α.*
2. *Under H_1, $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

- Complexity: $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$. Only need small J.

Normalized FSIC (NFSIC)

- Let $\hat{\mathbf{u}} := \left(\widehat{\text{cov}}[k(\mathbf{x}, \mathbf{v}_1), l(\mathbf{y}, \mathbf{w}_1)], \ldots, \widehat{\text{cov}}[k(\mathbf{x}, \mathbf{v}_J), l(\mathbf{y}, \mathbf{w}_J)]\right)^\top \in \mathbb{R}^J$.
- Then, $\widehat{\text{FSIC}^2} = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$.

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

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

- $\hat{\Sigma}_{ij}$ = covariance of $\hat{u}_i$ and $\hat{u}_j$.

Theorem 1 (NFSIC test is consistent).

Assume $\gamma_n \to 0$, and same conditions on k and l as before.

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$. Easy to get threshold T_α.*
2. *Under H_1, $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

- Complexity: $\mathcal{O}(J^3 + J^2 n + (d_x + d_y)Jn)$. Only need small J.

Normalized FSIC (NFSIC)

- Let $\hat{\mathbf{u}} := \left(\widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_1), l(\mathbf{y}, \mathbf{w}_1)], \ldots, \widehat{\mathrm{cov}}[k(\mathbf{x}, \mathbf{v}_J), l(\mathbf{y}, \mathbf{w}_J)]\right)^\top \in \mathbb{R}^J$.
- Then, $\widehat{\mathrm{FSIC}^2} = \frac{1}{J}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}$.

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

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

- $\hat{\Sigma}_{ij}$ = covariance of $\hat{u}_i$ and $\hat{u}_j$.

Theorem 1 (NFSIC test is consistent).

Assume $\gamma_n \to 0$, and same conditions on k and l as before.

1. *Under H_0, $\hat{\lambda}_n \xrightarrow{d} \chi^2(J)$ as $n \to \infty$. Easy to get threshold T_α.*
2. *Under H_1, $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.*

- Complexity: $\mathcal{O}(J^3 + J^2 n + (d_x + d_y) J n)$. Only need small J.

Tuning Features and Kernels

- Split the data into training (tr) and test (te) sets.

Procedure:

1. Choose $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$ and Gaussian widths by maximizing $\hat{\lambda}_n^{(\mathrm{tr})}$ (i.e., computed on the training set). Gradient ascent.
2. Reject H_0 if $\hat{\lambda}_n^{(\mathrm{te})} > (1-\alpha)$-quantile of $\chi^2(J)$.

- Splitting avoids overfitting.

Theorem 2.

- *This procedure increases a lower bound on $\mathbb{P}$(reject $H_0 \mid H_1$ true) (test power).*
- *Asymptotically, false rejection rate is α.*

Tuning Features and Kernels

- Split the data into training (tr) and test (te) sets.

Procedure:

1. Choose $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^J$ and Gaussian widths by maximizing $\hat{\lambda}_n^{(\mathrm{tr})}$ (i.e., computed on the training set). Gradient ascent.
2. Reject H_0 if $\hat{\lambda}_n^{(\mathrm{te})} > (1-\alpha)$-quantile of $\chi^2(J)$.

- Splitting avoids overfitting.

Theorem 2.

- *This procedure increases a lower bound on $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$ (test power).*
- *Asymptotically, false rejection rate is α.*

Tuning Features and Kernels

- Split the data into training (tr) and test (te) sets.

Procedure:

1. Choose $\{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{J}$ and Gaussian widths by maximizing $\hat{\lambda}_n^{(\mathrm{tr})}$ (i.e., computed on the training set). Gradient ascent.
2. Reject H_0 if $\hat{\lambda}_n^{(\mathrm{te})} > (1-\alpha)$-quantile of $\chi^2(J)$.

- Splitting avoids overfitting.

Theorem 2.

- *This procedure increases a lower bound on $\mathbb{P}(\text{reject } H_0 \mid H_1 \text{ true})$ (test power).*
- *Asymptotically, false rejection rate is α.*

Simulation Settings

- Gaussian kernels $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma_x^2}\right)$ for both X and Y.

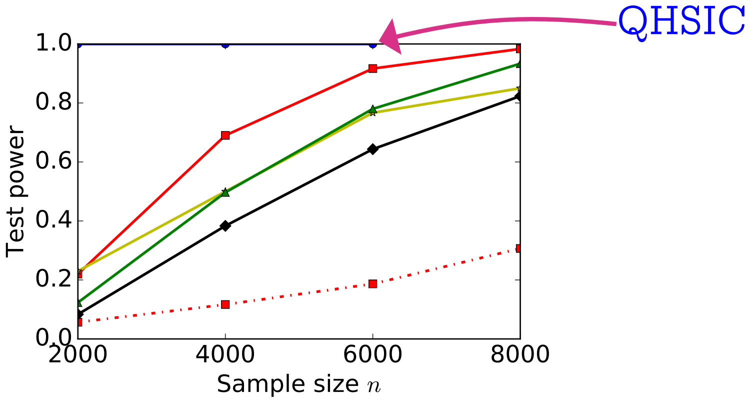
	Method	Description
1	**NFSIC-opt**	**NFSIC with optimization.** $\mathcal{O}(n)$.
2	**QHSIC** **[Gretton et al., 2005]**	**State-of-the-art HSIC.** $\mathcal{O}(n^2)$.
3	NFSIC-med	NFSIC with random features.
4	NyHSIC	Linear-time HSIC with Nystrom approx.
5	FHSIC	Linear-time HSIC with random Fourier features
6	RDC [Lopez-Paz et al., 2013]	Canonical Correlation Analysis with cosine basis.

NFSIC-opt NFSIC-med QHSIC NyHSIC FHSIC RDC

- $J = 10$ in NFSIC.

Youtube Video (X) vs. Caption (Y).

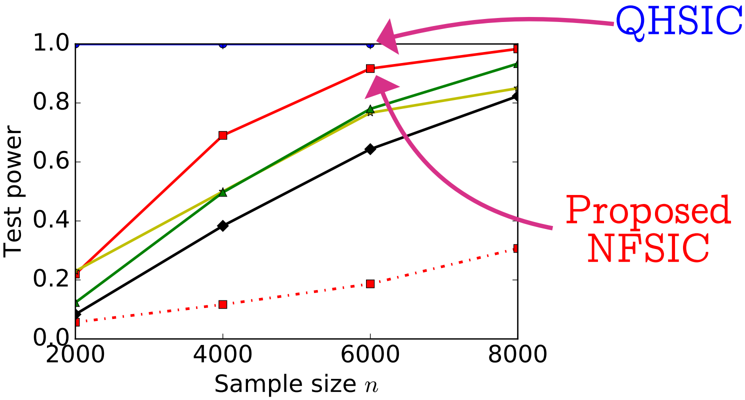
- $X \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms descriptors [Wang and Schmid, 2013].
- $Y \in \mathbb{R}^{1878}$: Bag of words. Term frequency.
- $\alpha = 0.01$.



- For large n, NFSIC is comparable to HSIC.

Youtube Video (X) vs. Caption (Y).

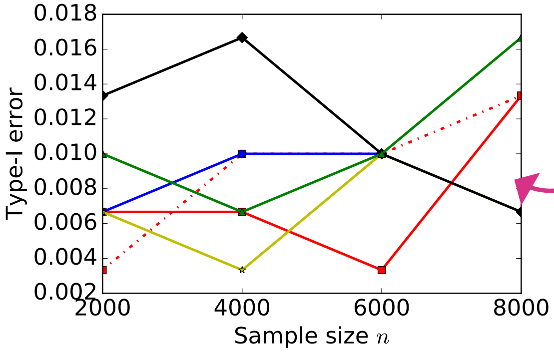
- $X \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms descriptors [Wang and Schmid, 2013].
- $Y \in \mathbb{R}^{1878}$: Bag of words. Term frequency.
- $\alpha = 0.01$.



- For large n, NFSIC is comparable to HSIC.

Youtube Video (X) vs. Caption (Y).

- $X \in \mathbb{R}^{2000}$: Fisher vector encoding of motion boundary histograms descriptors [Wang and Schmid, 2013].
- $Y \in \mathbb{R}^{1878}$: Bag of words. Term frequency.
- $\alpha = 0.01$.



Exchange (X, Y) pairs. H_0 true.

- For large n, NFSIC is comparable to HSIC.

Conclusions

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

An Adaptive Test of Independence with Analytic Kernel Embeddings
Wittawat Jitkrittum, Zoltán Szabó, Arthur Gretton
https://arxiv.org/abs/1610.04782
(to appear in ICML 2017)

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

Questions?

Thank you

Reference

Coauthors:

Zoltán Szabó
École Polytechnique

Arthur Gretton
Gatsby Unit, UCL

An Adaptive Test of Independence with Analytic Kernel Embeddings
Wittawat Jitkrittum, Zoltán Szabó, Arthur Gretton
https://arxiv.org/abs/1610.04782
(to appear in ICML 2017)

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

Requirements on the Kernels

Definition 1 (Analytic kernels).

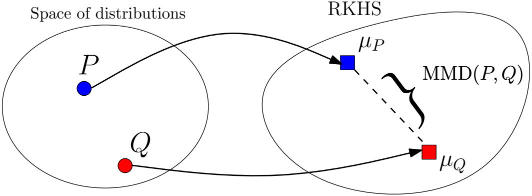
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be analytic if for all $\mathbf{x} \in \mathcal{X}$, $\mathbf{v} \to k(\mathbf{x}, \mathbf{v})$ is a real analytic function on $\mathcal{X}$.

- 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 $\mu_P(\mathbf{v}) := \mathbb{E}_{\mathbf{z}\sim P}[k(\mathbf{z}, \mathbf{v})]$.

k is said to be characteristic if μ_P is unique for distinct P. Equivalently, $P \mapsto \mu_P$ is injective.



Optimization Objective = Power Lower Bound

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

Theorem 3 (A lower bound on the test power).

1. *With some conditions, the test power* $\mathbb{P}_{H_1}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *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.*

2. *For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Optimization Objective = Power Lower Bound

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

Theorem 3 (A lower bound on the test power).

1. *With some conditions, the test power* $\mathbb{P}_{H_1}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *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.*

2. *For large* n*,* $L(\lambda_n)$ *is increasing in* λ_n*.*

Optimization Objective = Power Lower Bound

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

Theorem 3 (A lower bound on the test power).

1. *With some conditions, the test power* $\mathbb{P}_{H_1}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *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.*

2. *For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Optimization Objective = Power Lower Bound

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

Theorem 3 (A lower bound on the test power).

1. *With some conditions, the test power* $\mathbb{P}_{H_1}\left(\hat{\lambda}_n \geq T_\alpha\right) \geq L(\lambda_n)$ *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.*

2. *For large* n, $L(\lambda_n)$ *is increasing in* λ_n.

Set test locations and Gaussian widths = $\arg\max L(\lambda_n) = \arg\max \lambda_n$

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

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

- J 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 $\widehat{\text{NFSIC}^2}$

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

- J 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 $\widehat{\text{NFSIC}^2}$

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

- J 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).

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

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

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

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

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

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

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.

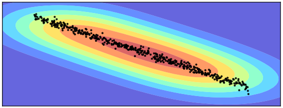
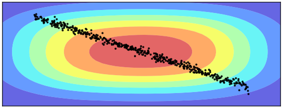
$$\text{HSIC}(X, Y) = \text{MMD}(P_{xy}, P_x P_y) = \|u\|_{\text{RKHS}}$$

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

 $-$ 

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

- $\text{HSIC}(X, Y) = 0$ 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]

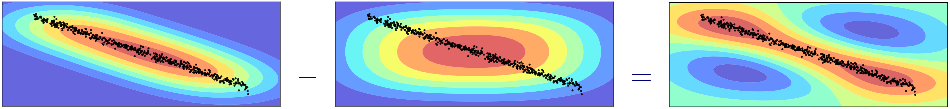
- Hilbert-Schmidt Independence Criterion.

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

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



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

- $\text{HSIC}(X, Y) = 0$ 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.

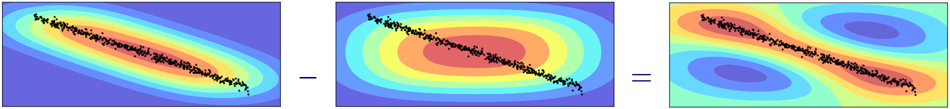
$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

(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.

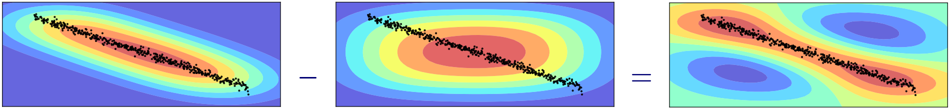
$$\mathrm{HSIC}(X, Y) = \mathrm{MMD}(P_{xy}, P_x P_y) = \|u\|_{\mathrm{RKHS}}$$

(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.

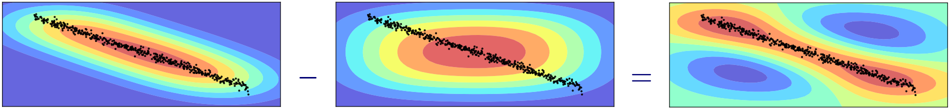
$$\text{HSIC}(X, Y) = \text{MMD}(P_{xy}, P_x P_y) = \|u\|_{\text{RKHS}}$$

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



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

- $\text{HSIC}(X, Y) = 0$ 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 $\mathcal{O}(n)$?

Proposal: The Finite Set Independence Criterion (FSIC)

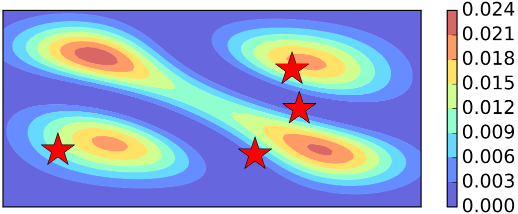
Idea: Evaluate $\hat{u}^2(\mathbf{v}, \mathbf{w})$ at only finitely many test locations.

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

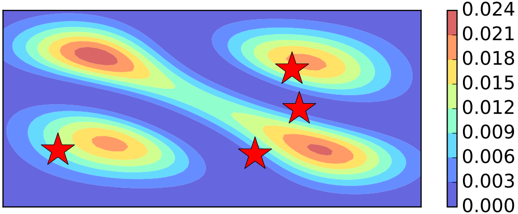


- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

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

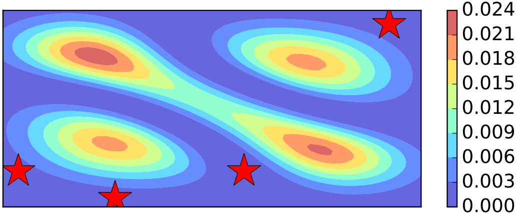


- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

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

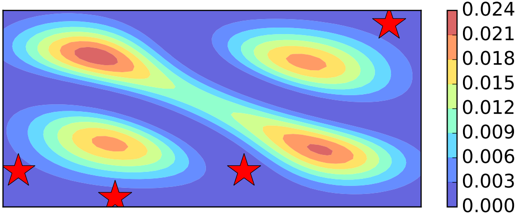


- Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.
- Can $\mathrm{FSIC}^2(X, Y) = 0$ even if X and Y are dependent??

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.
- 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.

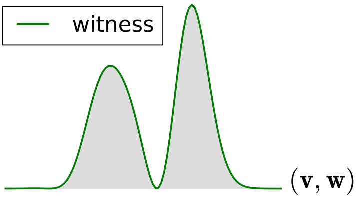
HSIC vs. FSIC

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]

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

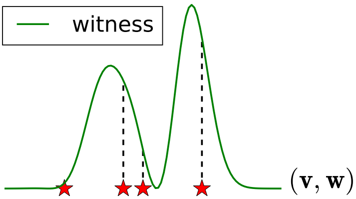


Good when difference between p_{xy} and $p_x p_y$ is spatially diffuse.

- $\hat{u}$ 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.

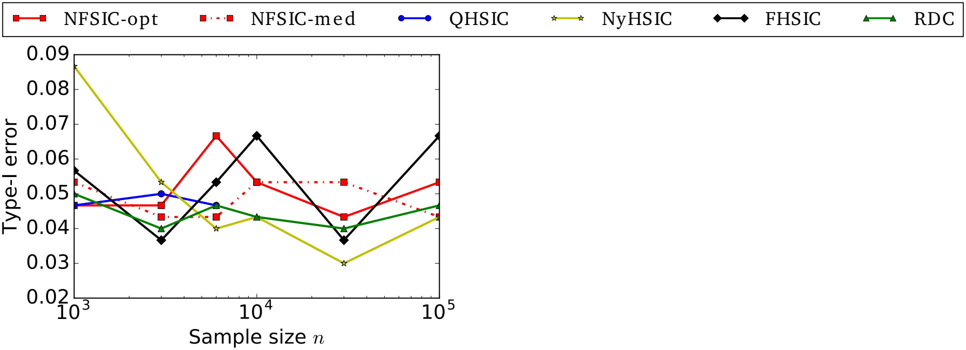
- $\hat{u}$ is mostly zero, has many peaks (feature interaction).

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

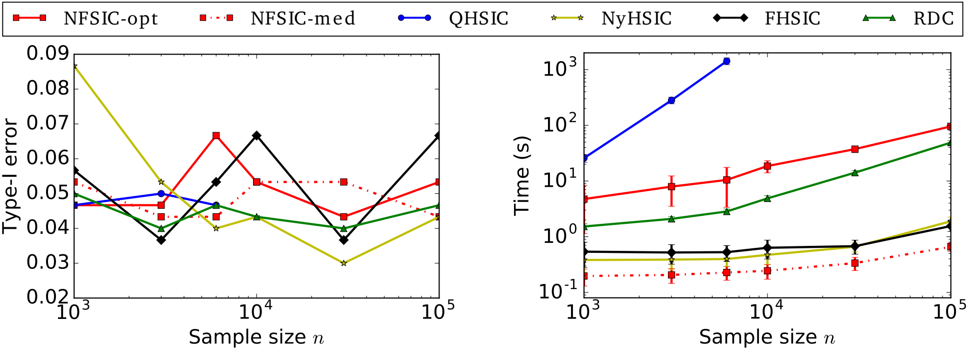
- $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_x})$ and $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{dy})$.
- Independent X, Y. So, H_0 holds.
- Set $\alpha := 0.05$, $d_x = d_y = 250$.



- 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 $\alpha := 0.05$, $d_x = d_y = 250$.



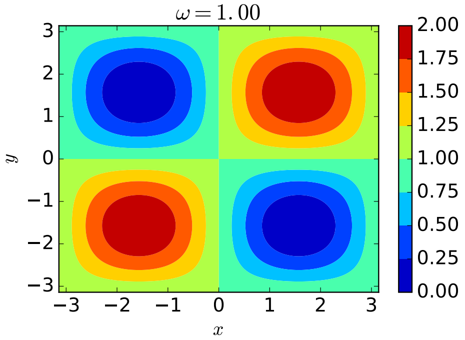
- Correct type-I errors (false positive rate).

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 $n = 4000$.

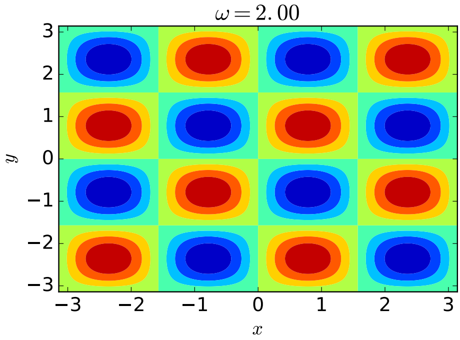
Toy Problem 2: Sinusoid

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



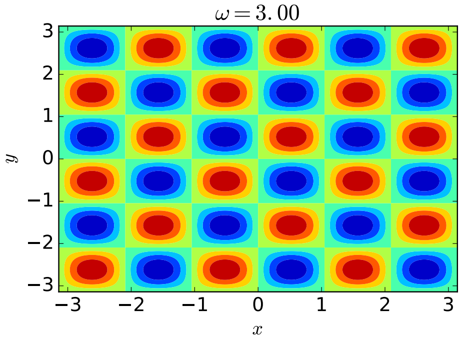
Toy Problem 2: Sinusoid

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



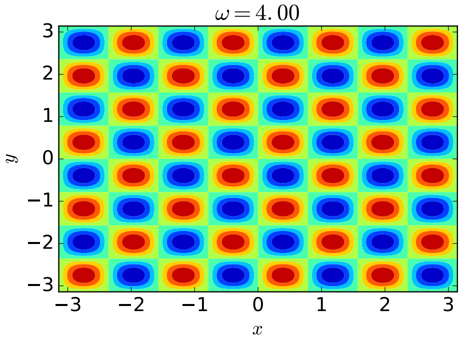
Toy Problem 2: Sinusoid

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



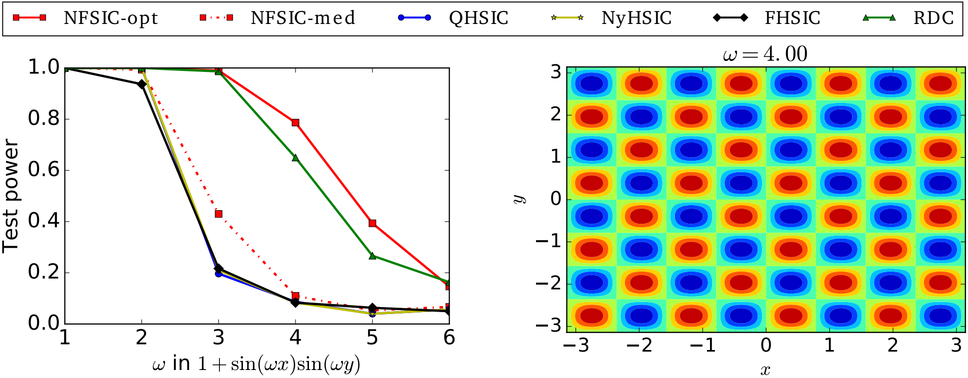
Toy Problem 2: Sinusoid

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



Toy Problem 2: Sinusoid

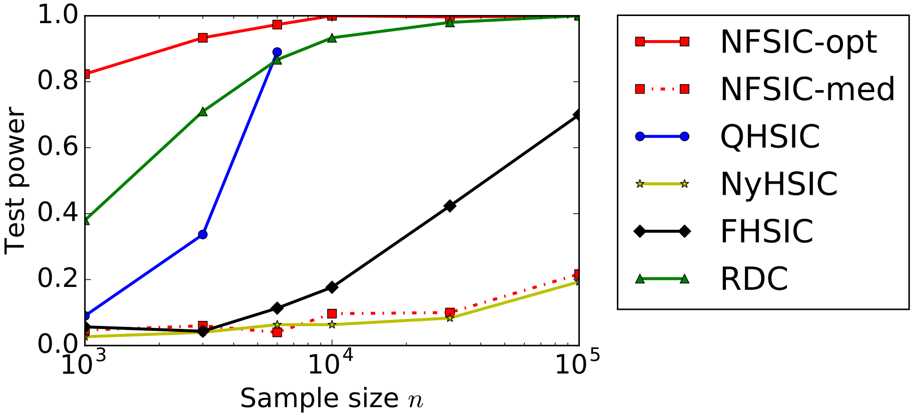
- $p_{xy}(x, y) \propto 1 + \sin(\omega x)\sin(\omega y)$ where $x, y \in (-\pi, \pi)$.
- Local changes between p_{xy} and $p_x p_y$.
- Set $n = 4000$.



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

Toy Problem 3: Gaussian Sign

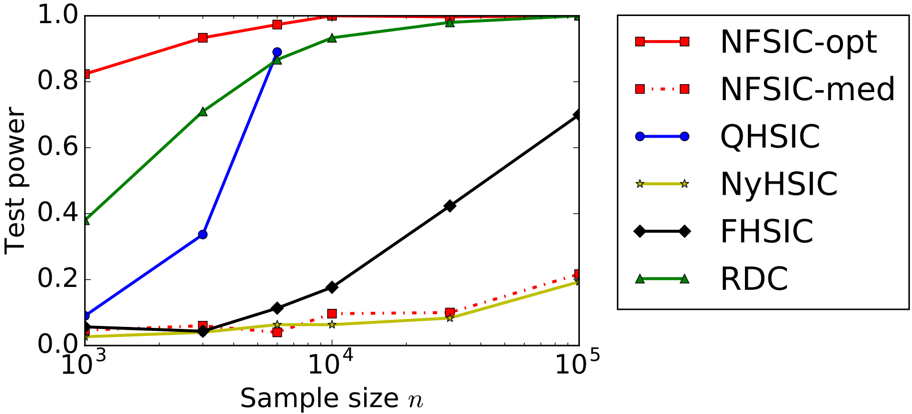
- $y = |Z| \prod_{i=1}^{d_x} \mathrm{sign}(x_i)$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ and $Z \sim \mathcal{N}(0, 1)$ (noise).
- Full interaction among $x_1, \ldots, x_{d_x}$.
- Need to consider all $x_1, \ldots, x_d$ to detect the dependency.



Main Point: NFSIC can handle feature interaction.

Toy Problem 3: Gaussian Sign

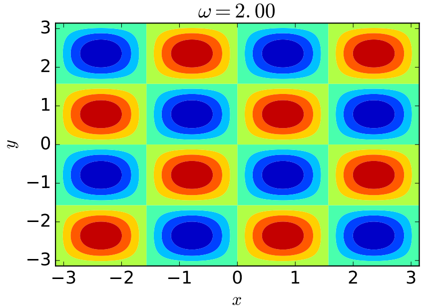
- $y = |Z| \prod_{i=1}^{d_x} \mathrm{sign}(x_i)$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_y})$ and $Z \sim \mathcal{N}(0, 1)$ (noise).
- Full interaction among $x_1, \ldots, x_{d_x}$.
- Need to consider all $x_1, \ldots, x_d$ to detect the dependency.

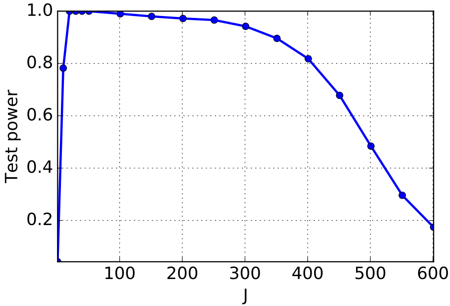


Main Point: NFSIC can handle feature interaction.

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.

Real Problem: Million Song Data

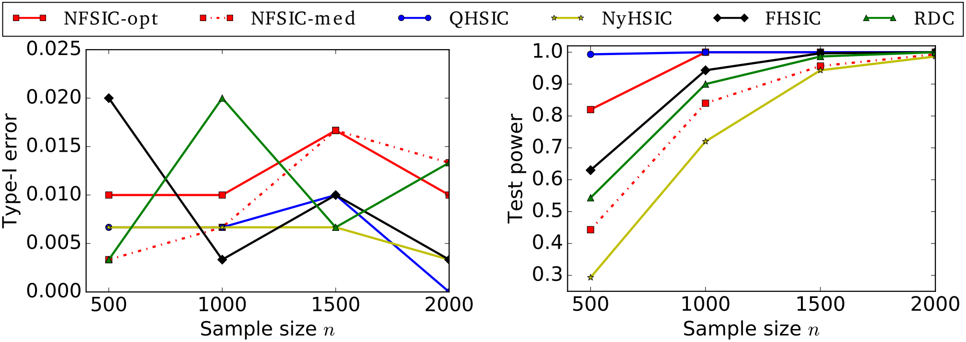
Song (X) vs. year of release (Y).

- 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: 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.

NFSIC-opt has the highest power among the linear-time tests.

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

Lopez-Paz, D., Hennig, P., and Schölkopf, B. (2013).
The Randomized Dependence Coefficient.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9.

References II

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