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 \(\{(x_i, y_i)\}_{i=1}^n \sim P_{xy}\) (unknown), test
  \[H_0 : P_{xy} = P_x P_y,\]
  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 \(\{(x_i, y_i)\}_i^{n} \sim P_{xy}\) (unknown), test
  \[H_0 : P_{xy} = P_x P_y,\]
  \[vs. \ H_1 : P_{xy} \neq P_x P_y.\]

- 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 \(\{(x_i, y_i)\}_{i=1}^n \sim P_{xy}\) (unknown), test
  \[H_0 : P_{xy} = P_x P_y,\]
  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.
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.
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: $O(n^2)$ where $n =$ sample size.
- ✗ No systematic way to choose kernels.

Propose the Finite-Set Independence Criterion (FSIC).

1. Nonparametric.
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: $k(x, v) = \exp\left(-\frac{||x-v||^2}{2\sigma^2_x}\right)$.

2. Pick some feature $(v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

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

\[\text{FSIC}^2(X, Y) = \text{cov}^2_{(x, y) \sim p_{xy}} [k(x, v), l(y, w)].\]
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: \( k \) for \( X \), and \( l \) for \( Y \).
   - Gaussian kernel: \( k(x, v) = \exp\left(-\frac{\|x-v\|^2}{2\sigma_x^2}\right) \).

2. Pick some feature \( (v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \)

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

\[
FSIC^2(X, Y) = \text{cov}^2_{(x, y) \sim P_{xy}} [k(x, v), l(y, w)] .
\]
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: \( k \) for \( X \), and \( l \) for \( Y \).
   - Gaussian kernel: \( k(x, v) = \exp\left(-\frac{||x-v||^2}{2\sigma_x^2}\right) \).

2. Pick some feature \((v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}\)

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

\[ \text{FSIC}^2(X, Y) = \text{cov}^2_{(x,y) \sim P_{xy}} [k(x, v), l(y, w)] . \]
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: $k$ for $X$, and $l$ for $Y$.
   - Gaussian kernel: $k(x, v) = \exp\left(-\frac{||x-v||^2}{2\sigma^2_x}\right)$.

2. Pick some feature $(v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

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

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

$$\text{FSIC}^2(X, Y) = \text{cov}^2_{(x,y) \sim \mathcal{P}_{xy}} [k(x, v), l(y, w)].$$

**Data $(v, w)$**

- **Correlation**: 0.97
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: $k$ for $X$, and $l$ for $Y$.
   - Gaussian kernel: $\kappa(x, v) = \exp\left(-\frac{||x-v||^2}{2\sigma^2}\right)$.

2. Pick some **feature** $(v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

3. Transform $(x, y) \mapsto (\kappa(x, v), l(y, w))$ then measure covariance
   \[ \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R} \times \mathbb{R} \]

   \[ \text{FSIC}^2(X, Y) = \text{cov}^2_{(x, y) \sim P_{xy}} [\kappa(x, v), l(y, w)] . \]

   - Data
   - $(v, w)$

   ![Data and feature data](image)

   correlation: -0.47
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: $k$ for $X$, and $l$ for $Y$.
   - Gaussian kernel: $k(x, v) = \exp\left(-\frac{||x-v||^2}{2\sigma_x^2}\right)$.

2. Pick some feature $(v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

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

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

- Data $(v, w)$
- correlation: 0.33

\[ k(x, v) \]
\[ l(y, w) \]
\[ x \]
\[ y \]
\[ 0 \quad 0.5 \quad 1.0 \]
\[ 0 \quad 0.5 \quad 1.0 \]
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: $k$ for $X$, and $l$ for $Y$.
   - Gaussian kernel: $k(x, v) = \exp\left(-\frac{||x-v||^2}{2\sigma_x^2}\right)$.

2. Pick some feature $(v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

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

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

- Data
- $(v, w)$

\begin{itemize}
  \item correlation: 0.023
\end{itemize}
Proposal: The Finite-Set Independence Criterion (FSIC)

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

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

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

$$FSIC^2(X, Y) = \text{cov}^2_{(x,y) \sim P_{xy}} [k(x, v), l(y, w)].$$
Proposal: The Finite-Set Independence Criterion (FSIC)

1. Pick 2 positive definite kernels: $k$ for $X$, and $l$ for $Y$.
   - Gaussian kernel: $k(x, v) = \exp \left( -\frac{||x-v||^2}{2\sigma^2_x} \right)$.

2. Pick some feature $(v, w) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$

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

$\text{FSIC}^2(X, Y) = \text{cov}^2_{(x,y) \sim p_{xy}} [k(x, v), l(y, w)]$.

Correlation: 0.087
General Form of FSIC

\[
\text{FSIC}^2(X, Y) = \frac{1}{J} \sum_{j=1}^{J} \text{cov}^2_{(x,y) \sim P_{xy}} [k(x, v_j), l(y, w_j)],
\]

for \( J \) features \( \{(v_j, w_j)\}_{j=1}^{J} \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \).

**Proposition 1.**

Assume

1. Kernels \( k \) and \( l \) satisfy some conditions (e.g. Gaussian kernels).
2. Features \( \{(v_i, w_i)\}_{i=1}^{J} \) 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 \( H_0 : P_{xy} = P_x P_y \),

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

- Difficult to get \((1 - \alpha)\)-quantile for the threshold.
General Form of FSIC

\[
FSIC^2(X, Y) = \frac{1}{J} \sum_{j=1}^{J} \text{cov}^2_{(x,y) \sim P_{xy}} [k(x, v_j), l(y, w_j)],
\]

for J features \( \{(v_j, w_j)\}_{j=1}^{J} \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \).

**Proposition 1.**

Assume

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

Then, for any \( J \geq 1 \),

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

Under \( H_0 : P_{xy} = P_x P_y \),

\[
nFSIC^2 \sim \text{weighted sum of } J \text{ dependent } \chi^2 \text{ variables}.
\]

- Difficult to get \((1 - \alpha)\)-quantile for the threshold.
General Form of FSIC

\[
\text{FSIC}^2(X, Y) = \frac{1}{J} \sum_{j=1}^{J} \text{cov}_{(x,y) \sim P_{xy}}^2 [k(x, v_j), l(y, w_j)],
\]

for \( J \) features \( \{(v_j, w_j)\}_{j=1}^{J} \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \).

**Proposition 1.**

Assume

1. Kernels \( k \) and \( l \) satisfy some conditions (e.g. Gaussian kernels).
2. Features \( \{(v_i, w_i)\}_{i=1}^{J} \) 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 \( H_0 : P_{xy} = P_x P_y \),

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

- **Difficult** to get \( (1 - \alpha) \)-quantile for the threshold.
Normalized FSIC (NFSIC)

- Let $\hat{u} := \left( \text{cov}[k(x, v_1), l(y, w_1)], \ldots, \text{cov}[k(x, v_J), l(y, w_J)] \right)^\top \in \mathbb{R}^J$.

- Then, $\widehat{\text{FSIC}}^2 = \frac{1}{J} \hat{u}^\top \hat{u}$.

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

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

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

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 $T_\alpha$.
2. Under $H_1$, $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.

- Complexity: $O(J^3 + J^2 n + (d_x + d_y)Jn)$. Only need small $J$. 
Normalized FSIC (NFSIC)

- Let $\hat{u} := \left(\text{cov}[k(x, v_1), l(y, w_1)], \ldots, \text{cov}[k(x, v_J), l(y, w_J)]\right)^\top \in \mathbb{R}^J$.

- Then, $\text{FSIC}^2 = \frac{1}{J} \hat{u}^\top \hat{u}$.

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

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

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

**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 $T_\alpha$.
2. Under $H_1$, $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.

- Complexity: $O(J^3 + J^2 n + (d_x + d_y)Jn)$. Only need small $J$. 
Normalized FSIC (NFSIC)

Let \( \hat{u} := \left( \text{cov}[k(x, v_1), l(y, w_1)], \ldots, \text{cov}[k(x, v_J), l(y, w_J)] \right)^\top \in \mathbb{R}^J. \)

Then, \( \hat{\text{FSIC}}^2 = \frac{1}{J} \hat{u}^\top \hat{u}. \)

\[
\hat{\text{NFSIC}}^2(X, Y) = \lambda_n := n \hat{u}^\top \left( \hat{\Sigma} + \gamma_n I \right)^{-1} \hat{u},
\]

with a regularization parameter \( \gamma_n \geq 0. \)

\( \hat{\Sigma}_{ij} = \text{covariance of } \hat{u}_i \text{ and } \hat{u}_j. \)

**Theorem 1 (NFSIC test is consistent).**

Assume \( \gamma_n \rightarrow 0, \) and same conditions on \( k \) and \( l \) as before.

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

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

- Let \( \hat{u} := \left( \text{cov}[k(x, v_1), l(y, w_1)], \ldots, \text{cov}[k(x, v_J), l(y, w_J)] \right)^\top \in \mathbb{R}^J \).

- Then, \( \widehat{\text{FSIC}}^2 = \frac{1}{J} \hat{u}^\top \hat{u} \).

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

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

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

- Let $\hat{u} := \left( \overline{\text{cov}}[k(x, v_1), l(y, w_1)], \ldots, \overline{\text{cov}}[k(x, v_J), l(y, w_J)] \right)^\top \in \mathbb{R}^J$.

- Then, $\overline{\text{FSIC}}^2 = \frac{1}{J} \hat{u}^\top \hat{u}$.

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

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

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

**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 $T_\alpha$.
2. Under $H_1$, $\mathbb{P}(\text{reject } H_0) \to 1$ as $n \to \infty$.

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

6/10
Tuning Features and Kernels

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

Procedure:

1. Choose \( \{ (v_i, w_i) \}_{i=1}^J \) and Gaussian widths by maximizing \( \hat{\lambda}_n^{(tr)} \) (i.e., computed on the training set). Gradient ascent.

2. Reject \( H_0 \) if \( \hat{\lambda}_n^{(te)} > (1 - \alpha) \)-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 \( \alpha \).
Tuning Features and Kernels

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

**Procedure:**

1. Choose \( \{(v_i, w_i)\}_{i=1}^{J} \) and Gaussian widths by maximizing \( \hat{\lambda}_n^{(tr)} \) (i.e., computed on the training set). Gradient ascent.

2. Reject \( H_0 \) if \( \hat{\lambda}_n^{(te)} > (1 - \alpha) \)-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 \( \alpha \).
Tuning Features and Kernels

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

**Procedure:**

1. Choose \( \{(v_i, w_i)\}_{i=1}^{J} \) and Gaussian widths by maximizing \( \hat{\lambda}_{n}^{(tr)} \) (i.e., computed on the training set). Gradient ascent.
2. Reject \( H_0 \) if \( \hat{\lambda}_{n}^{(te)} > (1 - \alpha) \)-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 \( \alpha \).
Simulation Settings

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 NFSIC-opt</td>
<td>NFSIC with optimization. ( O(n) ).</td>
</tr>
<tr>
<td>QHSIC</td>
<td>State-of-the-art HSIC. ( O(n^2) ).</td>
</tr>
<tr>
<td>2 [Gretton et al., 2005]</td>
<td></td>
</tr>
<tr>
<td>3 NFSIC-med</td>
<td>NFSIC with random features.</td>
</tr>
<tr>
<td>4 NyHSIC</td>
<td>Linear-time HSIC with Nystrom approx.</td>
</tr>
<tr>
<td>5 FHSIC</td>
<td>Linear-time HSIC with random Fourier features</td>
</tr>
<tr>
<td>6 RDC</td>
<td>Canonical Correlation Analysis with cosine basis.</td>
</tr>
<tr>
<td>8/10</td>
<td></td>
</tr>
</tbody>
</table>

- \( J = 10 \) in NFSIC.
Youtube Video ($X$) vs. Caption ($Y$).

- $Y \in \mathbb{R}^{1878}$: 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).**

- **Y ∈ ℝ^{1878}:** Bag of words. Term frequency.
- **α = 0.01.**

For large \( n \), NFSIC is comparable to HSIC.
Conclusions

- Proposed The Finite Set Independence Criterion (FSIC).
- Independence 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
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).

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

- Analytic: Taylor series about \( x_0 \) converges for all \( x_0 \in \mathcal{X} \).
- \( \implies k \) is infinitely differentiable.

Definition 2 (Characteristic kernels).

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

\( k \) is said to be characteristic if \( \mu_P \) is unique for distinct \( P \). Equivalently, \( P \mapsto \mu_P \) is injective.
Optimization Objective = Power Lower Bound

- Recall $\hat{\lambda}_n := n\hat{u}^\top (\hat{\Sigma} + \gamma_n I)^{-1} \hat{u}$.
- 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 $\mathbb{P}_{H_1} (\hat{\lambda}_n \geq T_\alpha) \geq L(\lambda_n)$ where

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

   where $\xi_1, \ldots, \xi_4, c_3 > 0$ are constants.

2. For large $n$, $L(\lambda_n)$ is increasing in $\lambda_n$. 
Recall $\hat{\lambda}_n := n\hat{u}^\top (\hat{\Sigma} + \gamma_n I)^{-1} \hat{u}$.  

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

\begin{center}
\begin{tabular}{|c|}
\hline
\textbf{Theorem 3 (A lower bound on the test power).} \\
\hline
\end{tabular}
\end{center}

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

\begin{align*}
L(\lambda_n) &= 1 - 62 e^{-\xi_1 \gamma_n^2 (\lambda_n - T_\alpha)^2 / n} - 2 e^{-0.5 n (\lambda_n - T_\alpha)^2 / [\xi_2 n^2]} \\
&\quad - 2 e^{-(\lambda_n - T_\alpha) \g n(n-1)/3 - \xi_3 \n - c_3 \gamma_n n(n-1) \g^2 / [\xi_4 n^2 (n-1)]},
\end{align*}

where $\xi_1, \ldots, \xi_4, c_3 > 0$ are constants.

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


Optimization Objective = Power Lower Bound

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

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

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

   where $\xi_1, \ldots, \xi_4, c_3 > 0$ are constants.

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


Optimization Objective = Power Lower Bound

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

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

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

   $$L(\lambda_n) = 1 - 62e^{-\xi_1 \gamma_n^2 (\lambda_n - T_{\alpha})^2 / n} - 2e^{-0.5n(\lambda_n - T_{\alpha})^2 / [\xi_2 n^2]}$$

   $$- 2e^{-[(\lambda_n - T_{\alpha})\gamma_n(n-1)/3 - \xi_3 n - c_3 \gamma_n^2 n(n-1)]^2 / [\xi_4 n^2 (n-1)]},$$

   where $\xi_1, \ldots, \xi_4, c_3 > 0$ are constants.

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

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

14/10
An Estimator of NFSIC\textsuperscript{2}

\[ \hat{\lambda}_n := n\hat{u}^\top (\hat{\Sigma} + \gamma_n I)^{-1} \hat{u}, \]

- \( J \) test locations \( \{(v_i, w_i)\}_{i=1}^J \sim \eta. \)
- \( K = [k(v_i, x_j)] \in \mathbb{R}^{J \times n} \)
- \( L = [l(w_i, y_j)] \in \mathbb{R}^{J \times n}. \) (No \( n \times n \) Gram matrix.)

Estimators

1. \( \hat{u} = \frac{(K \circ L)1_n}{n-1} - \frac{(K1_n \circ (L1_n))}{n(n-1)}. \)
2. \( \hat{\Sigma} = \frac{\Gamma \Gamma^\top}{n} \) where \( \Gamma := (K - n^{-1}K1_n1_n^\top) \circ (L - n^{-1}L1_n1_n^\top) - \hat{u}1_n^\top. \)

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

Main Point: Linear in \( n. \) Cubic in \( J \) (small).
An Estimator of NFSIC$^2$

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

- $J$ test locations $\{(v_i, w_i)\}_{i=1}^J \sim \eta$.
- $K = [k(v_i, x_j)] \in \mathbb{R}^{J \times n}$
- $L = [l(w_i, y_j)] \in \mathbb{R}^{J \times n}$. (No $n \times n$ Gram matrix.)

Estimators

1. $\hat{\mu} = \frac{(K \circ L) 1_n}{n-1} - \frac{(K 1_n) \circ (L 1_n)}{n(n-1)}$.

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

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

Main Point: Linear in $n$. Cubic in $J$ (small).
An Estimator of $\text{NFSIC}^2$

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

- **$J$ test locations** $\{(v_i, w_i)\}_{i=1}^J \sim \eta$.
- $K = [k(v_i, x_j)] \in \mathbb{R}^{J \times n}$
- $L = [l(w_i, y_j)] \in \mathbb{R}^{J \times n}$. (No $n \times n$ Gram matrix.)

**Estimators**

1. $\hat{\mu} = \frac{(K \circ L)1_n}{n-1} - \frac{(K 1_n) \circ (L 1_n)}{n(n-1)}$.

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

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

**Main Point:** Linear in $n$. Cubic in $J$ (small).
Alternative View of the Witness $u(v, w)$

The witness $u(v, w)$ can be rewritten as

$$
\begin{align*}
    u(v, w) & := \mu_{xy}(v, w) - \mu_x(v)\mu_y(w) \\
            & = \mathbb{E}_{xy}[k(x, v)l(y, w)] - \mathbb{E}_x[k(x, v)]\mathbb{E}_y[l(y, w)], \\
            & = \text{cov}_{xy}[k(x, v), l(y, w)].
\end{align*}
$$

1. Transforming $x \mapsto k(x, v)$ and $y \mapsto l(y, w)$ (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 $u(v, w)$

The witness $u(v, w)$ can be rewritten as

$$u(v, w) := \mu_{xy}(v, w) - \mu_x(v) \mu_y(w)$$
$$= \mathbb{E}_{xy}[k(x, v)l(y, w)] - \mathbb{E}_x[k(x, v)]\mathbb{E}_y[l(y, w)],$$
$$= \text{cov}_{xy}[k(x, v), l(y, w)].$$

1. Transforming $x \mapsto k(x, v)$ and $y \mapsto l(y, w)$ (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 $\hat{u}(v, w)$

- Recall $\overline{FSIC^2} = \frac{1}{J} \sum_{i=1}^{J} \hat{u}(v_i, w_i)^2$

- Let $\hat{\mu}_x \hat{\mu}_y(v, w)$ be an unbiased estimator of $\mu_x(v) \mu_y(w)$.

- $\hat{\mu}_x \hat{\mu}_y(v, w) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, v) l(y_j, w)$.

- An unbiased estimator of $u(v, w)$ is

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

  $$ = \frac{2}{n(n-1)} \sum_{i < j} h_{(v, w)}((x_i, y_i), (x_j, y_j)),$$

  where

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

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

- Recall $\text{FSIC}^2 = \frac{1}{J} \sum_{i=1}^{J} \hat{u}(v_i, w_i)^2$
- Let $\hat{\mu_x \mu_y}(v, w)$ be an unbiased estimator of $\mu_x(v) \mu_y(w)$.
- $\hat{\mu_x \mu_y}(v, w) := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, v) l(y_j, w)$.
- An unbiased estimator of $u(v, w)$ is
  \[
  \hat{u}(v, w) = \hat{\mu}_{xy}(v, w) - \hat{\mu}_x \hat{\mu}_y(v, w)
  = \frac{2}{n(n-1)} \sum_{i < j} h_{(v, w)}((x_i, y_i), (x_j, y_j)),
  \]

  where

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

- $\hat{u}(v, w)$ is a one-sample 2nd-order U-statistic, given $(v, w)$. 
Independence Test with HSIC [Gretton et al., 2005]

- **Hilbert-Schmidt Independence Criterion.**

  \[
  \text{HSIC}(X, Y) = \text{MMD}(P_{xy}, P_xP_y) = \|u\|_{\text{RKHS}}
  \]

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

- **Empirical witness:**

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

  where \(\hat{\mu}_{xy}(v, w) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v)l(y_i, w)\).

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

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

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

- **Empirical witness:**

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

where \(\hat{\mu}_{xy}(v, w) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) l(y_i, w)\).

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

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

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

- **Empirical witness:**

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

where \(\hat{\mu}_{xy}(v, w) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v)l(y_i, w)\).

\[
\hat{\mu}_{xy}(v, w) - \hat{\mu}_x(v)\hat{\mu}_y(w)
\]

- **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:**
  \[
  \hat{u}(v, w) = \hat{\mu}_{xy}(v, w) - \hat{\mu}_x(v)\hat{\mu}_y(w)
  \]
  where \( \hat{\mu}_{xy}(v, w) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v)l(y_i, w) \).

\[
\begin{align*}
\hat{\mu}_{xy}(v, w) & \quad - \quad \hat{\mu}_x(v)\hat{\mu}_y(w) \\
= & \quad \text{Witness } \hat{u}(v, w)
\end{align*}
\]

- **Key:** Can we measure the flatness by other way that costs only \( \mathcal{O}(n^2) \)?
**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:**

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

  where \( \hat{\mu}_{xy}(v, w) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v)l(y_i, w) \).

- \( \text{HSIC}(X, Y) = 0 \) if and only if \( X \) and \( Y \) are independent.

- Test statistic = \( \| \hat{u} \|_{\text{RKHS}} \) ("flatness" of \( \hat{u} \)). Complexity: \( O(n^2) \).

**Key:** Can we measure the flatness by other way that costs only \( O(n) \)?
Independence Test with HSIC [Gretton et al., 2005]

- **Hilbert-Schmidt Independence Criterion.**

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

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

- **Empirical witness:**

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

  where

  \[
  \hat{\mu}_{xy}(v, w) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v) l(y_i, w).
  \]

  \[
  \hat{\mu}_x(v) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v),
  \hat{\mu}_y(w) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, w).
  \]

- **Witness \( \hat{u}(v, w) \)**

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

  \[
  \text{HSIC}(X, Y) = 0 \text{ if and only if } X \text{ and } Y \text{ are independent.}
  \]

  \[
  \text{Test statistic } = \| \hat{u} \|_{\text{RKHS}} \text{ ("flatness" of } \hat{u}). \text{ Complexity: } \mathcal{O}(n^2).
  \]
Independence Test with HSIC \cite{Gretton2005}

- **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:**
  \[
  \hat{u}(v, w) = \hat{\mu}_{xy}(v, w) - \hat{\mu}_x(v)\hat{\mu}_y(w)
  \]
  where \(\hat{\mu}_{xy}(v, w) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, v)l(y_i, w)\).

- \(\text{HSIC}(X, Y) = 0\) 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 $\hat{u}^2(v, w)$ at only finitely many test locations.

- A set of random $J$ locations: $\{(v_1, w_1), \ldots, (v_J, w_J)\}$
- $\text{FSIC}^2(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(v_i, w_i)$

- Complexity: $O((d_x + d_y) Jn)$. Linear time.
- Can $\text{FSIC}^2(X, Y) = 0$ even if $X$ and $Y$ are dependent?
Proposal: The Finite Set Independence Criterion (FSIC)

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

- A set of random $J$ locations: $\{(v_1, w_1), \ldots, (v_J, w_J)\}$
- $\overline{\text{FSIC}}^2(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(v_i, w_i)$

![Diagram](image)

- Complexity: $\mathcal{O}((d_x + d_y) J n)$. Linear time.
- Can $\text{FSIC}^2(X, Y) = 0$ even if $X$ and $Y$ are dependent??
Proposal: The Finite Set Independence Criterion (FSIC)

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

- A set of random $J$ locations: $\{(v_1, w_1), \ldots, (v_J, w_J)\}$
- $\text{FSIC}^2(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(v_i, w_i)$

Complexity: $\mathcal{O}((d_x + d_y)Jn)$. Linear time.

Can $\text{FSIC}^2(X, Y) = 0$ even if $X$ and $Y$ are dependent?
Proposal: The Finite Set Independence Criterion (FSIC)

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

- A set of random $J$ locations: $\{(v_1, w_1), \ldots, (v_J, w_J)\}$
- $\widehat{\text{FSIC}}^2(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(v_i, w_i)$

**Complexity:** $\mathcal{O}((d_x + d_y) J n)$. Linear time.

- Can $\text{FSIC}^2(X, Y) = 0$ even if $X$ and $Y$ are dependent?
Proposal: The Finite Set Independence Criterion (FSIC)

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

- A set of random $J$ locations: $\{(v_1, w_1), \ldots, (v_J, w_J)\}$
- $\overline{\text{FSIC}}^2(X, Y) = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(v_i, w_i)$

![Image of a 3D graph with scattered points]

- Complexity: $\mathcal{O}((d_x + d_y) J n)$. Linear time.
- Can $\overline{\text{FSIC}}^2(X, Y) = 0$ even if $X$ and $Y$ are dependent??
- No. Population $\text{FSIC}(X, Y) = 0$ iff $X \perp Y$, almost surely.
**HSIC vs. FSIC**

Recall the witness

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

**HSIC** [Gretton et al., 2005]
\[ = ||\hat{u}||_{RKHS} \]

**Good** when difference between \( p_{xy} \) and \( p_x p_y \) is spatially diffuse.

- \( \hat{u} \) is almost flat.

**FSIC** [proposed]
\[ = \frac{1}{J} \sum_{i=1}^{J} \hat{u}^2(v_i, w_i) \]

**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

- $X \sim \mathcal{N}(0, \mathbf{I}_{d_x})$ and $Y \sim \mathcal{N}(0, \mathbf{I}_{d_y})$.
- Independent $X$, $Y$. So, $H_0$ holds.
- Set $\alpha := 0.05$, $d_x = d_y = 250$. 
Toy Problem 1: Independent Gaussians

- \( X \sim \mathcal{N}(0, I_{d_x}) \) and \( Y \sim \mathcal{N}(0, I_{d_y}) \).
- 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

- $X \sim \mathcal{N}(0, I_{d_x})$ and $Y \sim \mathcal{N}(0, I_{d_y})$.
- 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 $n = 4000$. 

![Image of the sinusoid function with a grid and contours]
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 $x \sim \mathcal{N}(0, I_{d_y})$ 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

- \( y = |Z| \prod_{i=1}^{d_x} \text{sign}(x_i) \), where \( x \sim \mathcal{N}(0, I_{d_y}) \) 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 $\hat{\lambda}_n = n\hat{u}^T (\hat{\Sigma} + \gamma_n I)^{-1} \hat{u}$ 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.

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

Break \((X, Y)\) pairs to simulate \(H_0\).


Action recognition with improved trajectories.
In *IEEE International Conference on Computer Vision (ICCV)*,
pages 3551–3558.