Nonparametric Independence Testing for Small Sample Sizes

Aaditya Ramdas, Leila Wehbe (IJCAI-2015)

Zoltán Szabó

Machine Learning Journal Club, Gatsby Unit

April 4, 2016
Goal: nonparametric independence testing.
Goal: nonparametric independence testing.

Idea:

large \( \text{cov}(X, Y) \Rightarrow \) declare dependence.
Goal: nonparametric independence testing.

Ideas:

1. Large $\text{cov}(X, Y) \Rightarrow$ declare dependence.
2. Large $\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \text{cov}(f(X), g(Y)) \Rightarrow$ dependence
   - nice asymptotic results.
Goal: nonparametric independence testing.

Ideas:
1. large $\text{cov}(X, Y) \Rightarrow$ declare dependence.
2. large $\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \text{cov}(f(X), g(Y)) \Rightarrow$ dependence
   - nice asymptotic results.

Focus:
- small sample size,
- small false positive regime: 'avoid' false dependence detection.
Goal: nonparametric independence testing.

Ideas:
1. large $\text{cov}(X, Y) \Rightarrow$ declare dependence.
2. large $\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \text{cov}(f(X), g(Y)) \Rightarrow$ dependence
   - nice asymptotic results.

Focus:
- small sample size,
- small false positive regime: 'avoid' false dependence detection.

Trick: introduce some bias to reduce variance - Stein.

large $\text{shrunk} \left[ \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \text{cov}(f(X), g(Y)) \right] \Rightarrow$ dependence
Given: \( \{(x_i, y_i)\}_{i=1}^{n} \sim_{i.i.d.} P_{XY} \).

Marginals of \( P_{XY} \): \( P_X, P_Y \).

Hypotheses:

\[
H_0 : P_{XY} = P_X \times P_Y, \quad H_1 : P_{XY} \neq P_X \times P_Y.
\]
Ingredients: independence testing problem

- **Given:** $\{(x_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{XY}$.
- **Marginals of** $P_{XY}$: $P_X, P_Y$.
- **Hypotheses:**
  \[
  H_0 : P_{XY} = P_X \times P_Y, \quad H_1 : P_{XY} \neq P_X \times P_Y.
  \]
- **Aim:**
  1. Low type-I error $= P(\text{detect dependence, when there isn't any}) \leq \alpha$,
  2. High power $= P(\text{detect dependence, when there is})$. 

Zoltán Szabó  Nonparametric Independence Testing for Small Sample Sizes
X ∈ (X, k), Y ∈ (Y, ℓ), k, ℓ: kernels. RKHSs: \( \mathcal{H}_k, \mathcal{H}_\ell \).
Ingredients: cross-covariance

- $X \in (\mathcal{X}, k)$, $Y \in (\mathcal{Y}, \ell)$, $k, \ell$: kernels. RKHSs: $\mathcal{H}_k$, $\mathcal{H}_\ell$.

- Mean embedding and its empirical counterpart:

$$
\mu_X = \mathbb{E}_{x \sim P_X} k(\cdot, x), \quad \mu_Y = \mathbb{E}_{y \sim P_Y} \ell(\cdot, y),
$$

$$
\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \quad \hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i).
$$
Ingredients: cross-covariance

- $X \in (\mathcal{X}, k)$, $Y \in (\mathcal{Y}, \ell)$, $k, \ell$: kernels. RKHSs: $\mathcal{H}_k, \mathcal{H}_\ell$.

- Mean embedding and its empirical counterpart:

  $$
  \mu_X = \mathbb{E}_{x \sim P_X} k(\cdot, x), \quad \mu_Y = \mathbb{E}_{y \sim P_Y} \ell(\cdot, y),
  $$

  $$
  \hat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \quad \hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i).
  $$

- Cross-covariance:

  $$
  \Sigma_{XY} = \mathbb{E}_{(x,y) \sim P_{XY}} \left[ \phi(x) - \mu_X \right] \otimes \left[ \psi(y) - \mu_Y \right] : \mathcal{H}_\ell \rightarrow \mathcal{H}_k,
  $$

  $$
  S_{XY} = \frac{1}{n} \sum_{i=1}^{n} \left[ \phi(x_i) - \hat{\mu}_X \right] \otimes \left[ \psi(y_i) - \hat{\mu}_Y \right].
  $$
Cross-covariance as an independence measure

Known: $\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_k} = \text{cov}(f(X), g(Y)), \forall g \in \mathcal{H}_\ell, f \in \mathcal{H}_k$.

Are $\mathcal{H}_\ell$ and $\mathcal{H}_k$ enough for the independence testing of $X$ and $Y$?

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0$. 
Known: \( \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_k} = \text{cov}(f(X), g(Y)), \forall g \in \mathcal{H}_\ell, f \in \mathcal{H}_k. \)

Are \( \mathcal{H}_\ell \) and \( \mathcal{H}_k \) enough for the independence testing of \( X \) and \( Y \)?

- \( C_b(X) \) and \( C_b(Y) \) would be sufficient: Jacod and Protter 2000.

Yes \( \Rightarrow \) Test: \( \Sigma_{XY} = 0. \)
Cross-covariance as an independence measure

Known: $\langle f, \Sigma_{XY} g \rangle_{H_k} = \text{cov}(f(X), g(Y)), \forall g \in H_\ell, f \in H_k$.

Are $H_\ell$ and $H_k$ enough for the independence testing of $X$ and $Y$?

- $C_b(X)$ and $C_b(Y)$ would be sufficient: Jacod and Protter 2000.
- Trick [Gretton et al. ’05]: guarantee the denseness of $H_k$ in $C_b(X)$, $H_\ell$ in $C_b(Y)$.

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0$. 
Cross-covariance as an independence measure

Known: \( \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_k} = \text{cov}(f(X), g(Y)), \forall g \in \mathcal{H}_\ell, f \in \mathcal{H}_k. \)

Are \( \mathcal{H}_\ell \) and \( \mathcal{H}_k \) enough for the independence testing of \( X \) and \( Y \)?

- \( C_b(X) \) and \( C_b(Y) \) would be sufficient: Jacod and Protter 2000.
- **Trick** [Gretton et al. ’05]: guarantee the denseness of \( \mathcal{H}_k \) in \( C_b(X) \), \( \mathcal{H}_\ell \) in \( C_b(Y) \).
- Space: compact metric, kernel: universal ✓

Yes \( \Rightarrow \) Test: \( \Sigma_{XY} = 0. \)
Cross-covariance as an independence measure

Known: $\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_k} = \text{cov}(f(X), g(Y)), \forall g \in \mathcal{H}_\ell, f \in \mathcal{H}_k.$

Are $\mathcal{H}_\ell$ and $\mathcal{H}_k$ enough for the independence testing of $X$ and $Y$?

- $C_b(X)$ and $C_b(Y)$ would be sufficient: Jacod and Protter 2000.
- **Trick** [Gretton et al. '05]: guarantee the *denseness* of $\mathcal{H}_k$ in $C_b(X)$, $\mathcal{H}_\ell$ in $C_b(Y)$.
- Space: compact metric, kernel: universal ✓
- Examples:

  $$k(x, x') = e^{-\gamma \|x-x'\|_2^2}, \quad k(x, x') = e^{-\gamma \|x-x'\|_1}.$$

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0.$
\[ \Sigma_{XY} \in HS(H_\ell, H_k) =: HS(G, F). \text{ What does this mean?} \]

Extension of Frobenious norm.

\[
\| C \|_F^2 = \sum_{i,j} C_{ij}^2
\]
\[ \sum_{XY} \in HS(\mathcal{H}_\ell, \mathcal{H}_k) =: HS(\mathcal{G}, \mathcal{F}). \] What does this mean?

Extension of Frobenious norm.

\[ \| C \|_F^2 = \sum_{i,j} C_{ij}^2, \]
\[ \| C \|_{HS}^2 = \sum_{i,j} \langle Cg_j, f_i \rangle_{\mathcal{F}}^2 < \infty, \]

where

- \( C : \mathcal{G} \to \mathcal{F} \) bounded linear operator.
- \( \mathcal{G}, \mathcal{F} \) are separable Hilbert spaces with ONBs \( \{g_j\}_j, \{f_i\}_i \).
HS operator example: \( f \otimes g \)

Intuition: \( fg^T \cdot (fg^T)u = f(g^Tu) \cdot (g^Tu) = \langle g, u \rangle \)
HS operator example: $f \otimes g$

- **Intuition:** $fg^T \cdot (fg^T)u = f(g^Tu) = \langle g, u \rangle$

- **Outer product:** $f \otimes g$ ($f \in \mathcal{F}, g \in \mathcal{G}$)

  $$(f \otimes g)(u) = f \langle g, u \rangle_{\mathcal{G}}, \forall u \in \mathcal{G}.$$
HS operator example: \( f \otimes g \)

- **Intuition:** \( fg^T \). \((fg^T)u = f(g^Tu).
  \[
  = \langle g, u \rangle
  \]

- **Outer product:** \( f \otimes g \) \((f \in \mathcal{F}, g \in \mathcal{G})\)
  \[
  (f \otimes g)(u) = f \langle g, u \rangle_{\mathcal{G}}, \forall u \in \mathcal{G}.
  \]

- **HS norm of** \( f \otimes g \):
  \[
  \|f \otimes g\|_{HS}^2 = \langle f, f \rangle_{\mathcal{F}} \langle g, g \rangle_{\mathcal{G}}.
  \]

  Cross-covariance: made of \( f \otimes g \)-type quantities.
It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =: \text{HSIC}$. 

\[
\text{HSIC} = \|\Sigma_{XY}\|_{HS}^2 = \left\langle \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \frac{1}{n} \sum_{j=1}^{n} \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}
\]
It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =: \text{HSIC}$.

$$\text{HSIC} = \|\Sigma_{XY}\|_{HS}^2 = \left\langle \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \frac{1}{n} \sum_{j=1}^{n} \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}$$

$$= \frac{1}{n^2} \sum_{i,j=1}^{n} \left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}$$

$$= \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}_k} \langle \tilde{\psi}(y_i), \tilde{\psi}(y_j) \rangle_{\mathcal{H}_\ell} = \tilde{K}_{ij} \tilde{L}_{ij}$$
It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =: \text{HSIC}$. 

\[
\text{HSIC} = \|\Sigma_{XY}\|_{HS}^2 = \left\langle \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \frac{1}{n} \sum_{j=1}^{n} \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}
\]

\[= \frac{1}{n^2} \sum_{i,j=1}^{n} \left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}
\]

\[= \frac{1}{n^2} \left\langle \tilde{K}, \tilde{L} \right\rangle_F.
\]
It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =: \text{HSIC}$.

\[
\text{HSIC} = \|\Sigma_{XY}\|_{HS}^2 = \left\langle \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \frac{1}{n} \sum_{j=1}^{n} \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}
\]

\[
= \frac{1}{n^2} \sum_{i,j=1}^{n} \left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}
= \frac{1}{n^2} \left\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \right\rangle_F.
\]

\[
\tilde{\mathbf{K}} = \mathbf{HKH}, \quad \mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1} \mathbf{1}^T, \quad \tilde{\mathbf{L}} = \mathbf{HLH}.
\]

Zoltán Szabó

Nonparametric Independence Testing for Small Sample Sizes
Independence test using HSIC [Gretton et al. 2005]

- Given: samples and $\alpha \in (0, 1)$.
- Test statistics: $T = HSIC = \|\Sigma_{XY}\|_{HS}^2$.
- Simulated null distribution of $T$: via $\{y_1, \ldots, y_n\}$ permutations $\Rightarrow t_\alpha$.
- Decision: reject $H_0$ if $t_\alpha < T$. 
Observation

- $S_{XY}$ is unbiased estimator of $\Sigma_{XY}$: $\mathbb{E}[S_{XY}] = \Sigma_{XY}$.
- **Issue**: $S_{XY}$ can have high variance for small sample numbers.
Observation

- $S_{XY}$ is unbiased estimator of $\Sigma_{XY}$: $\mathbb{E}[S_{XY}] = \Sigma_{XY}$.
- **Issue**: $S_{XY}$ can have **high variance** for small sample numbers.
- **Idea** [Stein, 1956]: decrease the variance by adding some bias.
• $S_{XY}$ is unbiased estimator of $\Sigma_{XY}$: $\mathbb{E}[S_{XY}] = \Sigma_{XY}$.

• Issue: $S_{XY}$ can have high variance for small sample numbers.

• Idea [Stein, 1956]: decrease the variance by adding some bias.

• [Maundet et al. 2014]: 2 shrinkage based estimators.
Observation

- $S_{XY}$ is unbiased estimator of $\Sigma_{XY}$: $\mathbb{E}[S_{XY}] = \Sigma_{XY}$.
- **Issue**: $S_{XY}$ can have high variance for small sample numbers.
- Idea [Stein, 1956]: decrease the variance by adding some bias.
- [Maundet et al. 2014]: 2 shrinkage based estimators.

**Questions**

1. How do they perform in independence testing?
2. Optimality?
Variations: shrinking towards zero

Recall: $S_{XY} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) \Rightarrow$

$$S_{XY} = \arg \min_{Z \in HS(\mathcal{H}_\ell, \mathcal{H}_k)} \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - Z \right\|^2_{HS}.$$
Variations: shrinking towards zero

Recall: \( S_{XY} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) \Rightarrow \)

\[
S_{XY} = \arg \min_{Z \in HS(\mathcal{H}_\ell, \mathcal{H}_k)} \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - Z \right\|_{HS}^2.
\]

SCOSE (simple covariance shrinkage estimator, \( \lambda > 0 \)):

\[
S_{XY}^S = \arg \min_{Z \in HS(\mathcal{H}_\ell, \mathcal{H}_k)} \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - Z \right\|_{HS}^2 + \lambda \left\| Z \right\|_{HS}^2.
\]
Variations: shrinking towards zero

Recall: $S_{XY} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i)$ \Rightarrow

$$S_{XY} = \arg \min_{Z \in HS(\mathcal{H}_\ell, \mathcal{H}_k)} \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - Z \right\|_{HS}^2.$$  

**SCOSE (simple covariance shrinkage estimator, $\lambda > 0$):**

$$S_{XY}^{S} = \arg \min_{Z \in HS(\mathcal{H}_\ell, \mathcal{H}_k)} \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - Z \right\|_{HS}^2 + \lambda \left\| Z \right\|_{HS}^2.$$  

**FCOSE (flexible covariance shrinkage estimator):**

$$S_{XY}^{F} = \sum_{j=1}^{n} \frac{\beta_j}{n} \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j),$$

$$\beta = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - \sum_{j=1}^{n} \frac{\beta_j}{n} \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\|_{HS}^2 + \lambda \left\| \beta \right\|_{2}^2.$$
In both cases: $\lambda$ is chosen via leave-one-out CV.

- **SCOSE**: analytical formula for $\lambda^*$

$$
HSIC^S = \left\| S^S_{XY} \right\|_{HS}^2 = \left( 1 - \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} - HSIC \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} (n-2)HSIC + \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} \right)^2 + HSIC.
$$
In both cases: $\lambda$ is chosen via leave-one-out CV.

- **SCOSE**: analytical formula for $\lambda^*$
  \[
  HSIC^S = \left\| S^S_{XY} \right\|_{HS}^2 = \left( 1 - \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} - HSIC \right) \frac{(n - 2)HSIC + \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii}}{n} \right)^2 + HSIC.
  \]

- **FCOSE**: after SVD of $\tilde{K} \circ \tilde{L} [O(n^3)]$, '/$\lambda$': $O(n^2)$.  

Zoltán Szabó  Nonparametric Independence Testing for Small Sample Sizes
In both cases: $\lambda$ is chosen via leave-one-out CV.

- **SCOSE**: analytical formula for $\lambda^*$

$$
HSIC^S = \left\llbracket S^S_{XY} \right\rrbracket^2_{HS} = \left(1 - \frac{\frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} - HSIC}{(n - 2) HSIC + \frac{\frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii}}{n}} \right)^2 + HSIC.
$$

- **FCOSE**: after SVD of $\tilde{K} \circ \tilde{L}$ $[O(n^3)]$, '/$\lambda'$: $O(n^2)$.

**Statement**

SCOSE is (essentially) the oracle linear shrinkage estimator w.r.t. the quadratic loss.
Oracle estimator: linear shrinkage, quadratic loss

Proposition

\[(S^*, \rho^*) := \arg \min_{Z \in HS(\mathcal{H}_\ell, \mathcal{H}_k), Z = (1-\rho)S_{XY}, \rho \in [0,1]} \mathbb{E} \|Z - \Sigma_{XY}\|^2_{HS}.\]

\[S^* = (1 - \rho^*)S_{XY},\]

\[\rho^* = \frac{\mathbb{E} \|S_{XY} - \Sigma_{XY}\|^2_{HS}}{\mathbb{E} \|S_{XY}\|^2_{HS}}.\]

Intuition: we shrink \(S_{XY}\) towards zero, optimally in quadratic sense.
Using $\mathbb{E}[S_{XY}] = \Sigma_{XY}$:

$$\mathbb{E} \left\| Z - \Sigma_{XY} \right\|_{HS}^2 = \mathbb{E} \left\| (1 - \rho)S_{XY} - \Sigma_{XY} \right\|_{HS}^2 =$$

$$= \mathbb{E} \left\| -\rho S_{XY} + (S_{XY} - \Sigma_{XY}) \right\|_{HS}^2$$
Indeed

Using $\mathbb{E}[S_{XY}] = \Sigma_{XY}$:

\[
\mathbb{E} \left\| Z - \Sigma_{XY} \right\|_{HS}^2 = \mathbb{E} \left\| (1 - \rho)S_{XY} - \Sigma_{XY} \right\|_{HS}^2 = \\
= \mathbb{E} \left\| -\rho S_{XY} + (S_{XY} - \Sigma_{XY}) \right\|_{HS}^2 \\
= \rho^2 \mathbb{E} \left\| S_{XY} \right\|_{HS}^2 + \mathbb{E} \left\| S_{XY} - \Sigma_{XY} \right\|_{HS}^2 - 2\rho \mathbb{E} \left\langle S_{XY}, S_{XY} - \Sigma_{XY} \right\rangle_{HS}
\]
Indeed

Using $\mathbb{E}[S_{XY}] = \Sigma_{XY}$:

$$
\mathbb{E} \| Z - \Sigma_{XY} \|_{HS}^2 = \mathbb{E} \| (1 - \rho)S_{XY} - \Sigma_{XY} \|_{HS}^2 = \\
= \mathbb{E} \| -\rho S_{XY} + (S_{XY} - \Sigma_{XY}) \|_{HS}^2 \\
= \rho^2 \mathbb{E} \| S_{XY} \|_{HS}^2 + \mathbb{E} \| S_{XY} - \Sigma_{XY} \|_{HS}^2 - 2\rho \mathbb{E} \langle S_{XY}, S_{XY} - \Sigma_{XY} \rangle_{HS} \\
= \rho^2 \mathbb{E} \| S_{XY} \|_{HS}^2 + (1 - 2\rho) \mathbb{E} \| S_{XY} - \Sigma_{XY} \|_{HS}^2 =: J(\rho).
$$
Using $\mathbb{E}[S_{XY}] = \Sigma_{XY}$:

\[
\mathbb{E} \left\| Z - \Sigma_{XY} \right\|^2_{HS} = \mathbb{E} \left\| (1 - \rho)S_{XY} - \Sigma_{XY} \right\|^2_{HS} = \\
= \mathbb{E} \left\| -\rho S_{XY} + (S_{XY} - \Sigma_{XY}) \right\|^2_{HS} \\
= \rho^2 \mathbb{E} \left\| S_{XY} \right\|^2_{HS} + \mathbb{E} \left\| S_{XY} - \Sigma_{XY} \right\|^2_{HS} - 2\rho \mathbb{E} \left\langle S_{XY}, S_{XY} - \Sigma_{XY} \right\rangle_{HS} \\
= \rho^2 \mathbb{E} \left\| S_{XY} \right\|^2_{HS} + (1 - 2\rho) \mathbb{E} \left\| S_{XY} - \Sigma_{XY} \right\|^2_{HS} =: J(\rho).
\]

Optimizing in $\rho$:

\[
0 = J'(\rho) = 2\rho \mathbb{E} \left\| S_{XY} \right\|^2_{HS} - 2\mathbb{E} \left\| S_{XY} - \Sigma_{XY} \right\|^2_{HS} \Rightarrow \\
\rho^* = \frac{\mathbb{E} \left\| S_{XY} - \Sigma_{XY} \right\|^2_{HS}}{\mathbb{E} \left\| S_{XY} \right\|^2_{HS}}.
\]
Plug-in estimator

\[ \rho^* = \frac{\mathbb{E} \| S_{XY} - \Sigma_{XY} \|_{HS}^2}{\mathbb{E} \| S_{XY} \|_{HS}^2} = \frac{\beta}{\delta}, \]
Plug-in estimator

\[
\rho^* = \frac{\mathbb{E} \| S_{XY} - \Sigma_{XY} \|^2_{HS}}{\mathbb{E} \| S_{XY} \|^2_{HS}} = \frac{\beta}{\delta}, \quad \hat{\delta} = \| S_{XY} \|^2_{HS} = HSIC,
\]
Plug-in estimator

\[
\rho^* = \frac{E \| S_{XY} - \Sigma_{XY} \|^2_{HS}}{E \| S_{XY} \|^2_{HS}} = \frac{\beta}{\delta}, \quad \hat{\delta} = \| S_{XY} \|^2_{HS} = \text{HSIC},
\]

\[
\beta = E \left\| \frac{1}{n} \sum_{i=1}^{n} \left[ \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right] \right\|^2_{HS} = \frac{1}{n} E \left\| \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right\|^2_{HS}
\]
\[
\rho^* = \frac{\mathbb{E} \left\| S_{XY} - \Sigma_{XY} \right\|_{HS}^2}{\mathbb{E} \left\| S_{XY} \right\|_{HS}^2} = \frac{\beta}{\delta}, \quad \hat{\delta} = \left\| S_{XY} \right\|_{HS}^2 = \text{HSIC},
\]

\[
\beta = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[ \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right] \right\|_{HS}^2 = \frac{1}{n} \mathbb{E} \left\| \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right\|_{HS}^2
\]

\[
\approx \frac{1}{n^2} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - S_{XY} \right\|_{HS}^2 \approx \frac{1}{n^2} \sum_{i=1}^{n} \left[ \tilde{K}_{ii} \tilde{L}_{ii} + \left\| S_{XY} \right\|_{HS}^2 - 2 \left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), S_{XY} \right\rangle_{HS} \right]
\]
Plug-in estimator

\[ \rho^* = \frac{\mathbb{E} \| S_{XY} - \Sigma_{XY} \|^2_{HS}}{\mathbb{E} \| S_{XY} \|^2_{HS}} = \frac{\beta}{\delta}, \quad \hat{\delta} = \| S_{XY} \|^2_{HS} = \text{HSIC}, \]

\[ \beta = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[ \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right] \right\|_{HS}^2 = \frac{1}{n} \mathbb{E} \left\| \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right\|_{HS}^2 \]

\[ \approx \frac{1}{n^2} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - S_{XY} \right\|_{HS}^2 - \| S_{XY} \|^2_{HS} - 2 \langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), S_{XY} \rangle_{HS} \]

\[ = \frac{1}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} + \| S_{XY} \|^2_{HS} - 2 \| S_{XY} \|^2_{HS} \right] =: \hat{\beta}, \]
Plug-in estimator

\[
\rho^* = \frac{\mathbb{E} \| S_{XY} - \Sigma_{XY} \|^2_{HS}}{\mathbb{E} \| S_{XY} \|^2_{HS}} = \frac{\beta}{\delta}, \quad \hat{\delta} = \| S_{XY} \|^2_{HS} = \text{HSIC},
\]

\[
\beta = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[ \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right] \right\|^2_{HS} = \frac{1}{n} \mathbb{E} \left\| \tilde{\phi}(x_i) \otimes \tilde{\phi}(y_i) - \Sigma_{XY} \right\|^2_{HS}
\]

\[
\geq \frac{1}{n^2} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - S_{XY} \right\|^2_{HS} - 2 \left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), S_{XY} \right\rangle_{HS}
\]

\[
= \frac{1}{n} \left[ \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} + \| S_{XY} \|^2_{HS} - 2 \left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), S_{XY} \right\rangle_{HS} \right] =: \hat{\beta},
\]

\[
\Rightarrow \text{HSIC}^* = \left( 1 - \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} - \text{HSIC} \right)^2 \text{HSIC}.
\]
Comparison

- **SCOSE:**

\[
HSIC^S = \left(1 - \frac{\frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} - HSIC}{(n-2)HSIC + \frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii}}\right)^2 + HSIC.
\]

- **Oracle estimator with plug-in:**

\[
\hat{HSIC}^* = \left(1 - \frac{\frac{1}{n} \sum_{i=1}^{n} \tilde{K}_{ii} \tilde{L}_{ii} - HSIC}{nHSIC}\right)^2 HSIC.
\]

SCOSE \approx \text{oracle with perturbed plug-in.}
Numerical experiments

- Shrinkage usually improves power.
Numerical experiments

- Shrinkage usually improves power.
- FCOSE: often achieves better power $\rightarrow$ non-linear shrinkage?, non-quadratic loss?
Numerical experiments

- Shrinkage usually improves power.
- FCOSE: often achieves better power → non-linear shrinkage?, non-quadratic loss?
- Soft HSIC shrinkage:
Thank you for the attention!