Nonparametric Independence Testing for Small Sample Sizes

Aaditya Ramdas, Leila Wehbe (IJCAI-2015)

Zoltán Szabó

Machine Learning Journal Club, Gatsby Unit

April 4, 2016

One-page summary

- **Goal**: nonparametric independence testing.

One-page summary

- **Goal**: nonparametric independence testing.
- Idea:
  1. large $cov(X, Y)$ $\Rightarrow$ declare dependence.

One-page summary

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

One-page summary

- **Goal**: nonparametric independence testing.
- Ideas:
  1. large $cov(X, Y)$ $\Rightarrow$ declare dependence.
  2. large $\sup_{f \in \mathcal{F}, g \in \mathcal{G}} cov(f(X), g(Y))$ $\Rightarrow$ dependence
     - nice asymptotic results.
- Focus:
  - small sample size,
  - small false positive regime: 'avoid' false dependence detection.

One-page summary

- **Goal**: nonparametric independence testing.
- Ideas:
  1. large $cov(X, Y) \Rightarrow$ declare dependence.
  2. large $\sup_{f \in \mathcal{F}, g \in \mathcal{G}} cov(f(X), g(Y)) \Rightarrow$ 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 shrunk$[\sup_{f \in \mathcal{F}, g \in \mathcal{G}} cov(f(X), g(Y))] \Rightarrow$ dependence

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:

$$H_0 : P_{XY} = P_X \times P_Y, \qquad H_1 : P_{XY} \neq P_X \times P_Y.$$

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:

$$H_0 : P_{XY} = P_X \times P_Y, \qquad H_1 : P_{XY} \neq P_X \times P_Y.$$

- Aim:
  1. Low type-I error $= \mathbb{P}(\underbrace{\text{detect dependence, when there isn't any}}_{\text{false positive}}) \leq \alpha$,
  2. High power $= \mathbb{P}(\text{detect dependence, when there is})$.

Ingredients: cross-covariance

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

Ingredients: cross-covariance

- $X \in (\mathcal{X}, k)$, $Y \in (\mathcal{Y}, \ell)$, k, ℓ: kernels. RKHSs: $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- Mean embedding and its empirical counterpart:

$$\mu_X = \mathbb{E}_{x \sim \mathbb{P}_X} \underbrace{k(\cdot, x)}_{=:\phi(x)}, \qquad \mu_Y = \mathbb{E}_{y \sim \mathbb{P}_Y} \underbrace{\ell(\cdot, y)}_{=:\psi(y)},$$

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

Ingredients: cross-covariance

- $X \in (\mathcal{X}, k)$, $Y \in (\mathcal{Y}, \ell)$, k, ℓ: kernels. RKHSs: $\mathcal{H}_k$, $\mathcal{H}_\ell$.
- Mean embedding and its empirical counterpart:

$$\mu_X = \mathbb{E}_{x \sim \mathbb{P}_X} \underbrace{k(\cdot, x)}_{=:\phi(x)}, \qquad \mu_Y = \mathbb{E}_{y \sim \mathbb{P}_Y} \underbrace{\ell(\cdot, y)}_{=:\psi(y)},$$

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

- Cross-covariance:

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

$$S_{XY} = \frac{1}{n} \sum_{i=1}^{n} [\phi(x_i) - \hat{\mu}_X] \otimes [\psi(y_i) - \hat{\mu}_Y].$$

Cross-covariance as an independence measure

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

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

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0$.

Cross-covariance as an independence measure

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

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

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

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0$.

Cross-covariance as an independence measure

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

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

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

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0$.

Cross-covariance as an independence measure

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

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

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

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0$.

Cross-covariance as an independence measure

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

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

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

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

Yes $\Rightarrow$ Test: $\Sigma_{XY} = 0$.

Side-note

$\Sigma_{XY} \in HS(\mathcal{H}_\ell, \mathcal{H}_k) =: HS(\mathcal{G}, \mathcal{F})$. What does this mean?
Extension of Frobenious norm.

$$\|C\|_F^2 = \sum_{i,j} {C_{ij}}^2$$

Side-note

$\Sigma_{XY} \in HS(\mathcal{H}_\ell, \mathcal{H}_k) =: HS(\mathcal{G}, \mathcal{F})$. What does this mean?
Extension of Frobenious norm.

$$\|C\|_F^2 = \sum_{i,j} C_{ij}{}^2,$$

$$\|C\|_{HS}^2 = \sum_{i,j} \langle C g_j, f_i \rangle_{\mathcal{F}}^2 < \infty,$$

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: $\mathbf{f}\mathbf{g}^T$. $(\mathbf{f}\mathbf{g}^T)\mathbf{u} = \mathbf{f}\underbrace{(\mathbf{g}^T\mathbf{u})}_{=\langle \mathbf{g},\mathbf{u}\rangle}$.

HS operator example: $f \otimes g$

- Intuition: $\mathbf{f}\mathbf{g}^T$. $(\mathbf{f}\mathbf{g}^T)\mathbf{u} = \mathbf{f}\underbrace{(\mathbf{g}^T\mathbf{u})}_{=\langle \mathbf{g},\mathbf{u}\rangle}$.
- 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: $\mathbf{f}\mathbf{g}^T$. $(\mathbf{f}\mathbf{g}^T)\mathbf{u} = \mathbf{f}\,\underbrace{(\mathbf{g}^T\mathbf{u})}_{=\langle \mathbf{g},\mathbf{u}\rangle}$.
- 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.

HSIC

It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =:$ HSIC.

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

HSIC

It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =:$ HSIC.

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

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} \underbrace{\left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}}_{\left\langle \tilde{\phi}(x_i),\tilde{\phi}(x_j) \right\rangle_{\mathcal{H}_k} \left\langle \tilde{\psi}(y_i),\tilde{\psi}(y_j) \right\rangle_{\mathcal{H}_\ell} = \tilde{K}_{ij}\tilde{L}_{ij}}$$

HSIC

It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =:$ HSIC.

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

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} \underbrace{\left\langle \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i), \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\rangle_{HS}}_{\left\langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \right\rangle_{\mathcal{H}_k} \left\langle \tilde{\psi}(y_i), \tilde{\psi}(y_j) \right\rangle_{\mathcal{H}_\ell} = \tilde{K}_{ij}\tilde{L}_{ij}}$$

$$= \frac{1}{n^2}\left\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \right\rangle_F.$$

It is easy to compute $\|\Sigma_{XY}\|_{HS}^2 =:$ HSIC.

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

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} \underbrace{\left\langle \tilde{\phi}(x_i)\otimes\tilde{\psi}(y_i), \tilde{\phi}(x_j)\otimes\tilde{\psi}(y_j) \right\rangle_{HS}}_{\left\langle\tilde{\phi}(x_i),\tilde{\phi}(x_j)\right\rangle_{\mathcal{H}_k}\left\langle\tilde{\psi}(y_i),\tilde{\psi}(y_j)\right\rangle_{\mathcal{H}_\ell}=\tilde{K}_{ij}\tilde{L}_{ij}}$$

$$= \frac{1}{n^2}\left\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \right\rangle_F.$$

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

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 Σ_{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 Σ_{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.

Observation

- S_{XY} is unbiased estimator of Σ_{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 Σ_{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: $S_{XY} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\phi}(x_i)\otimes\tilde{\psi}(y_i) \Rightarrow$

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

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, $\lambda > 0$):

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

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):

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

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^n \left\| \tilde{\phi}(x_i) \otimes \tilde{\psi}(y_i) - \sum_{j=1}^n \frac{\beta_j}{n} \tilde{\phi}(x_j) \otimes \tilde{\psi}(y_j) \right\|_{HS}^2 + \lambda \left\| \boldsymbol{\beta} \right\|_2^2 .$$

SCOSE vs FCOSE

In both cases: λ is chosen via leave-one-out CV.

- SCOSE: analytical formula for λ_*

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

SCOSE vs FCOSE

In both cases: λ is chosen via leave-one-out CV.

- SCOSE: analytical formula for λ_*

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

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

SCOSE vs FCOSE

In both cases: λ is chosen via leave-one-out CV.

- SCOSE: analytical formula for λ_*

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

- 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

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

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

$$\rho^* = \frac{\mathbb{E} \left\| S_{XY} - \Sigma_{XY} \right\|_{HS}^2}{\mathbb{E} \left\| S_{XY} \right\|_{HS}^2}.$$

Intuition: we shrink S_{XY} towards zero, optimally in quadratic sense.

Indeed

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

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

Indeed

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

$$\begin{aligned}\mathbb{E}\left\|Z - \Sigma_{XY}\right\|_{HS}^2 = \mathbb{E}\left\|(1-\rho)S_{XY} - \Sigma_{XY}\right\|_{HS}^2 = \\ = \quad & \mathbb{E}\left\|-\rho S_{XY} + (S_{XY} - \Sigma_{XY})\right\|_{HS}^2 \\ = \quad & \rho^2 \mathbb{E}\left\|S_{XY}\right\|_{HS}^2 + \underbrace{\mathbb{E}\left\|S_{XY} - \Sigma_{XY}\right\|_{HS}^2}_{\mathbb{E}\|S_{XY}\|_{HS}^2 - \|\Sigma_{XY}\|_{HS}^2} - 2\rho \underbrace{\mathbb{E}\left\langle S_{XY}, S_{XY} - \Sigma_{XY}\right\rangle_{HS}}_{\mathbb{E}\|S_{XY}\|_{HS}^2 - \|\Sigma_{XY}\|_{HS}^2}\end{aligned}$$

Indeed

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

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

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

$$= \rho^2\mathbb{E}\left\|S_{XY}\right\|_{HS}^2 + \underbrace{\mathbb{E}\left\|S_{XY} - \Sigma_{XY}\right\|_{HS}^2}_{\mathbb{E}\|S_{XY}\|_{HS}^2 - \|\Sigma_{XY}\|_{HS}^2} - 2\rho\underbrace{\mathbb{E}\left\langle S_{XY}, S_{XY} - \Sigma_{XY}\right\rangle_{HS}}_{\mathbb{E}\|S_{XY}\|_{HS}^2 - \|\Sigma_{XY}\|_{HS}^2}$$

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

Indeed

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

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

Optimizing in ρ:

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

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

$$\rho^* = \frac{\mathbb{E}\left\|S_{XY} - \Sigma_{XY}\right\|_{HS}^2}{\mathbb{E}\left\|S_{XY}\right\|_{HS}^2} = \frac{\beta}{\delta}, \ \widehat{\delta} = \left\|S_{XY}\right\|_{HS}^2 = HSIC,$$

Plug-in estimator

$$\rho^* = \frac{\mathbb{E}\left\|S_{XY} - \Sigma_{XY}\right\|_{HS}^2}{\mathbb{E}\left\|S_{XY}\right\|_{HS}^2} = \frac{\beta}{\delta},\ \widehat{\delta} = \left\|S_{XY}\right\|_{HS}^2 = HSIC,$$

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

Plug-in estimator

$$\rho^* = \frac{\mathbb{E}\,\|S_{XY} - \Sigma_{XY}\|_{HS}^2}{\mathbb{E}\,\|S_{XY}\|_{HS}^2} = \frac{\beta}{\delta},\ \widehat{\delta} = \|S_{XY}\|_{HS}^2 = HSIC,$$

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

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

Plug-in estimator

$$\rho^* = \frac{\mathbb{E}\left\|S_{XY} - \Sigma_{XY}\right\|_{HS}^2}{\mathbb{E}\left\|S_{XY}\right\|_{HS}^2} = \frac{\beta}{\delta}, \ \widehat{\delta} = \left\|S_{XY}\right\|_{HS}^2 = HSIC,$$

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

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

$$= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{K}_{ii}\tilde{L}_{ii} + \underbrace{\left\|S_{XY}\right\|_{HS}^2 - 2\left\|S_{XY}\right\|_{HS}^2}_{-\left\|S_{XY}\right\|_{HS}^2 = -HSIC}\right] =: \widehat{\beta},$$

Plug-in estimator

$$\rho^* = \frac{\mathbb{E}\left\|S_{XY} - \Sigma_{XY}\right\|_{HS}^2}{\mathbb{E}\left\|S_{XY}\right\|_{HS}^2} = \frac{\beta}{\delta},\ \widehat{\delta} = \left\|S_{XY}\right\|_{HS}^2 = HSIC,$$

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

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

$$= \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{K}_{ii}\tilde{L}_{ii} + \underbrace{\left\|S_{XY}\right\|_{HS}^2 - 2\left\|S_{XY}\right\|_{HS}^2}_{-\left\|S_{XY}\right\|_{HS}^2=-HSIC}\right] =: \widehat{\beta},$$

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

Comparison

- SCOSE:

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

- Oracle estimator with plug-in:

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

SCOSE ≈ oracle with perturbed plug-in.

Numerical experiments

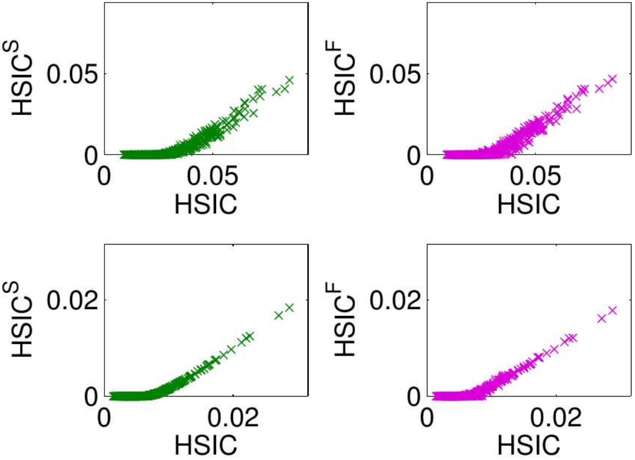
- Shrinkage usually improves power.

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 $\rightarrow$ non-linear shrinkage?, non-quadratic loss?
- Soft HSIC shrinkage:



Thank you for the attention!