Statistical Tests of Dependence and Conditional Dependence

Arthur Gretton

Gatsby Computational Neuroscience Unit,
University College London

Workshop on Hypothesis Testing, ICML 2026

July 13, 2026

Outline

Statistical tests of dependence ($X \perp\!\!\!\perp Y$?)

- Covaraince of smooth functions of data (COCO)
- Hilbert-Schmidt norm of feature covariance (HSIC)
- How to test significance?
- Feature covariance vs mutual information

Statistical tests of conditional dependence ($X \perp\!\!\!\perp Y | Z$?)

- Condition for conditional independence
- Test statistic (KCI)
- How to choose test parameters
- Why so many false positives, and what to do about it?

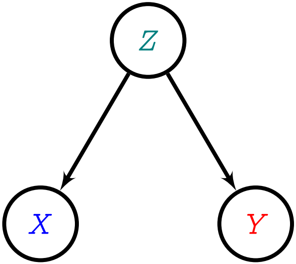
Application: causal structure discovery

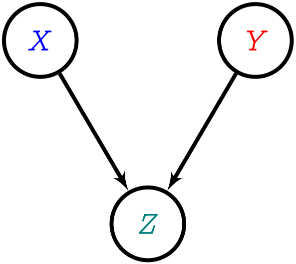
$X \not\!\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \mid Z$

$X \perp\!\!\!\perp Y$

$X \not\!\perp\!\!\!\perp Y \mid Z$





Application: as a loss/regularizer

Dependence

NeurIPS 2021

Self-Supervised Learning with Kernel Dependence Maximization

Yazhe Li* DeepMind and Gatsby Unit, UCL yazhe@google.com

Roman Pogodin* Gatsby Unit, UCL roman.pogodin.17@ucl.ac.uk

Danica J. Sutherland UBC and Amii† dsuth@cs.ubc.ca

Arthur Gretton Gatsby Unit, UCL arthur.gretton@gmail.com

Conditional dependence

ICLR 2023

EFFICIENT CONDITIONALLY INVARIANT REPRESENTATION LEARNING

Roman Pogodin* Gatsby Unit, UCL rmn.pogodin@gmail.com

Namrata Deka* UBC dnamrata@cs.ubc.ca

Yazhe Li* DeepMind & Gatsby Unit, UCL yazhe@google.com

Danica J. Sutherland UBC & Amii dsuth@cs.ubc.ca

Victor Veitch UChicago & Google Brain victorveitch@google.com

Arthur Gretton Gatsby Unit, UCL arthur.gretton@gmail.com

Dependence testing

Illustration: dependence $\neq$correlation

- Given: Samples from a distribution P_{XY}
- Goal: Are X and Y dependent?

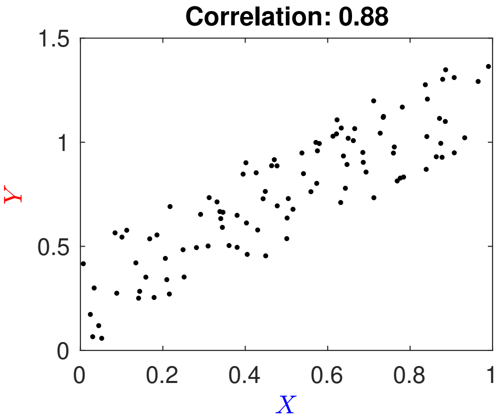


Illustration: dependence $\neq$correlation

- Given: Samples from a distribution P_{XY}
- Goal: Are X and Y dependent?

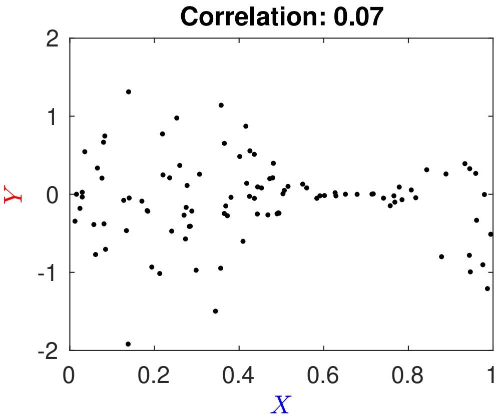
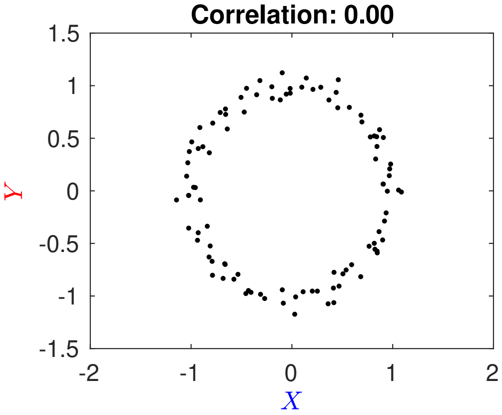


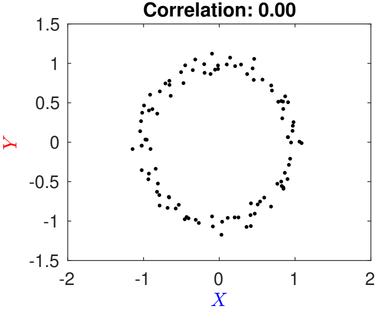
Illustration: dependence $\neq$correlation

- Given: Samples from a distribution P_{XY}
- Goal: Are X and Y dependent?



Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.

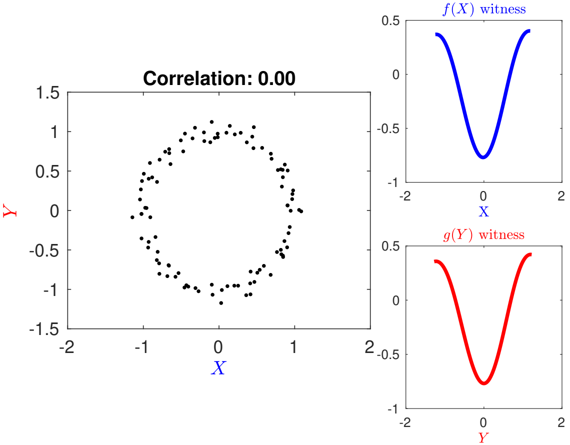


;

G, Smola, Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Scheolkopf, Logothetis (AISTATS 2005)

Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.

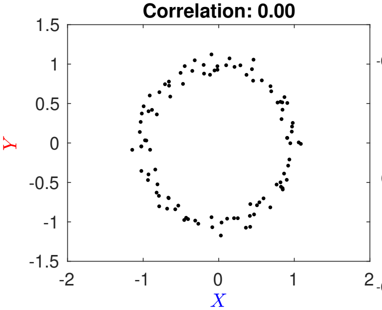


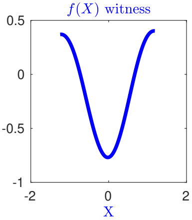
;

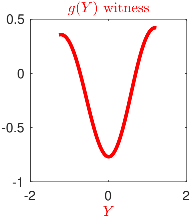
G, Smola, Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Scheolkopf, Logothetis (AISTATS 2005)

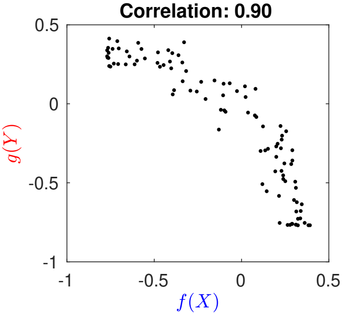
Finding covariance with smooth transformations

Illustration: two variables with no correlation but strong dependence.









;

G, Smola, Bousquet, Herbrich, Belitski, Augath, Murayama, Pauls, Scheolkopf, Logothetis (AISTATS 2005)

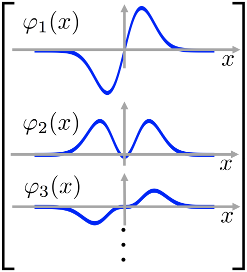
Define two spaces, one for each witness

Function in $\mathcal{F}$

$$f(x) = \sum_{\ell=1}^{\infty} f_\ell \varphi_\ell(x)$$

Feature map

$$\varphi(x) =$$



Kernel for RKHS $\mathcal{F}$ on $\mathcal{X}$:

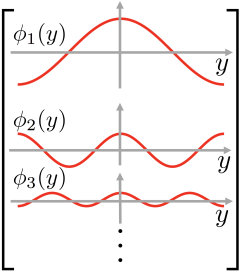
$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$$

Function in $\mathcal{G}$

$$g(y) = \sum_{\ell=1}^{\infty} g_\ell \phi_\ell(y)$$

Feature map

$$\phi(y) =$$



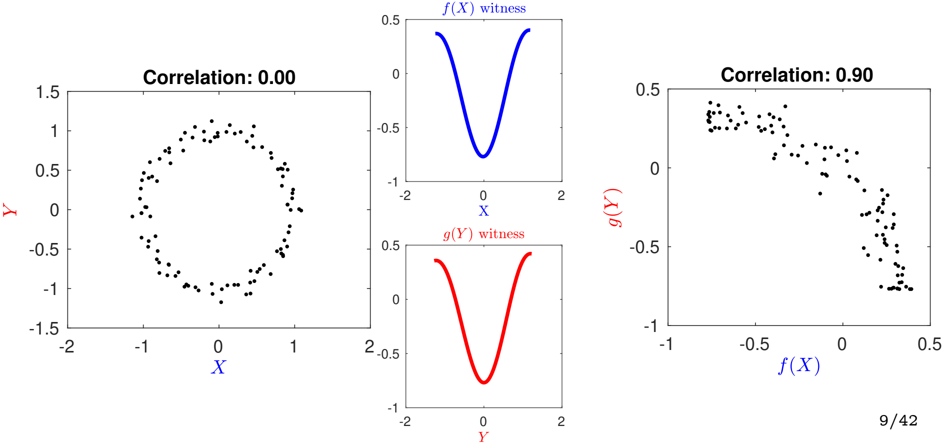
Kernel for RKHS $\mathcal{G}$ on $\mathcal{Y}$:

$$l(y, y') = \langle \phi(y), \phi(y') \rangle_{\mathcal{G}}$$

The constrained covariance

The constrained covariance is

$$\mathrm{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \mathrm{cov}[f(X)g(Y)]$$



The constrained covariance

The constrained covariance is

$$\mathrm{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \mathrm{cov}\left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(X)\right)\left(\sum_{j=1}^{\infty} g_j \phi_j(Y)\right)\right]$$

The constrained covariance

The constrained covariance is

$$\mathrm{COCO}(P_{XY}) = \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \mathrm{cov}\left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(X)\right)\left(\sum_{j=1}^{\infty} g_j \phi_j(Y)\right)\right]$$

$$= \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \left\langle f, \underbrace{\mathrm{E}\left[\varphi^{\mu}(X) \otimes \phi^{\mu}(Y)\right]}_{C_{XY}} g \right\rangle$$

Centering: $\varphi^{\mu}(x) = \varphi(x) - \underbrace{\mathrm{E}_X \varphi(X)}_{\mu_X}$ and $\phi^{\mu}(y) = \phi(y) - \underbrace{\mathrm{E}_Y \phi(Y)}_{\mu_Y}$.

The constrained covariance

The constrained covariance is

$$\begin{aligned}\mathrm{COCO}(P_{XY}) &= \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \mathrm{cov}\left[\left(\sum_{j=1}^{\infty} f_j \varphi_j(X)\right)\left(\sum_{j=1}^{\infty} g_j \phi_j(Y)\right)\right] \\ &= \sup_{\substack{\|f\|_{\mathcal{F}} \leq 1 \\ \|g\|_{\mathcal{G}} \leq 1}} \left\langle f, \underbrace{\mathrm{E}\left[\varphi^{\mu}(X) \otimes \phi^{\mu}(Y)\right]}_{C_{XY}} g \right\rangle\end{aligned}$$

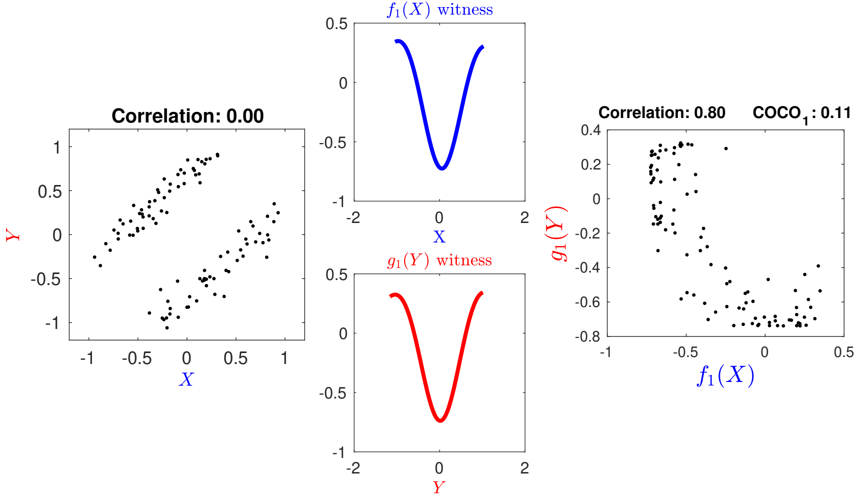
Centering: $\varphi^{\mu}(x) = \varphi(x) - \underbrace{\mathrm{E}_X \varphi(X)}_{\mu_X}$ and $\phi^{\mu}(y) = \phi(y) - \underbrace{\mathrm{E}_Y \phi(Y)}_{\mu_Y}$.

$\boxed{\text{COCO: max singular value of feature covariance } C_{XY}}$

Can we do better than COCO?

A second example with zero correlation.

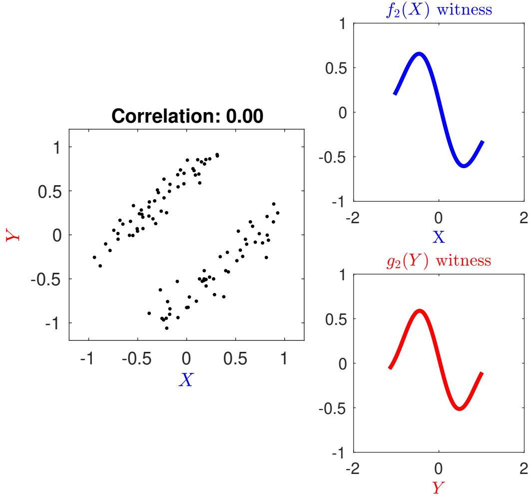
First singular value of feature covariance C_{XY}:



Can we do better than COCO?

A second example with zero correlation.

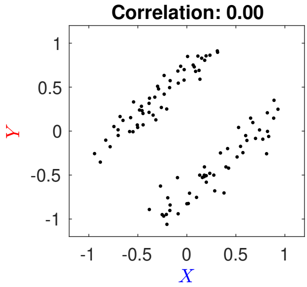
Second singular value of feature covariance C_{XY}:

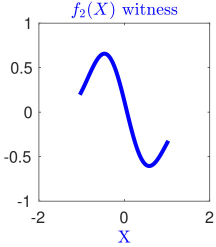


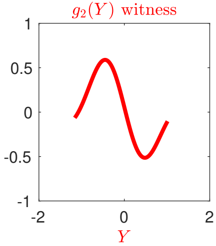
Can we do better than COCO?

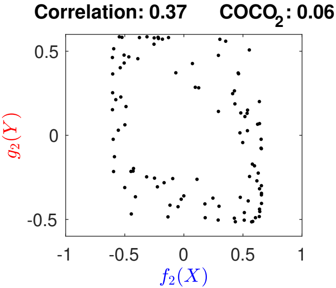
A second example with zero correlation.

Second singular value of feature covariance C_{XY}:









The Hilbert-Schmidt Independence Criterion

Writing the ith singular value of the feature covariance C_{XY} as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC) [A,B]

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = \sum_{i=1}^{\infty} \gamma_i^2$$

$$= E_{XY} E_{X'Y'}[k(X, X')l(Y, Y')]$$
$$+ E_{XX'}[k(X, X')]E_{YY'}[l(Y, Y')]$$
$$- 2E_{XY}\left[E_{X'}[k(X, X')]E_{Y'}[l(Y, Y')]\right]$$

[A] G, Bousquet, Smola, Schoelkopf (ALT 2005)
[B] G, Fukumizu, Teo, Song, Schoelkopf, Smola (NeurIPS 2007)
[C] G. ArXiv, 2015
[D] Lyons (Annals of Probability 2013)

The Hilbert-Schmidt Independence Criterion

Writing the ith singular value of the feature covariance C_{XY} as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC) [A,B]

$$\begin{aligned} HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) &= \sum_{i=1}^{\infty} \gamma_i^2 \\ &= E_{XY} E_{X'Y'}[k(X, X')l(Y, Y')] \\ &\quad + E_{XX'}[k(X, X')] E_{YY'}[l(Y, Y')] \\ &\quad - 2E_{XY}\left[E_{X'}[k(X, X')] E_{Y'}[l(Y, Y')]\right] \end{aligned}$$

[A] G, Bousquet, Smola, Schoelkopf (ALT 2005)
[B] G, Fukumizu, Teo, Song, Schoelkopf, Smola (NeurIPS 2007)
[C] G. ArXiv, 2015
[D] Lyons (Annals of Probability 2013)

The Hilbert-Schmidt Independence Criterion

Writing the ith singular value of the feature covariance C_{XY} as

$$\gamma_i := COCO_i(P_{XY}; \mathcal{F}, \mathcal{G}),$$

define Hilbert-Schmidt Independence Criterion (HSIC) [A,B]

$$\begin{aligned} HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) &= \sum_{i=1}^{\infty} \gamma_i^2 \\ &= E_{XY} E_{X'Y'}[k(X, X') l(Y, Y')] \\ &\quad + E_{XX'}[k(X, X')] E_{YY'}[l(Y, Y')] \\ &\quad - 2E_{XY}\left[E_{X'}[k(X, X')] E_{Y'}[l(Y, Y')]\right] \end{aligned}$$

When k, l characteristic kernels [C, following from D]

$$HSIC^2(P_{XY}; \mathcal{F}, \mathcal{G}) = 0 \quad \text{iff} \quad P_{XY} = P_X P_Y$$

[A]; G, Bousquet, Smola, Schoelkopf (ALT 2005)
[B] G, Fukumizu, Teo, Song, Schoelkopf, Smola (NeurIPS 2007)
[C] G. ArXiv, 2015
[D] Lyons (Annals of Probability 2013)

A statistical test using HSIC

- Empirical HSIC (simplified biased estimator)

$$\widehat{HSIC}^2 = \frac{1}{n^2}\mathrm{trace}(KHLH)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ $\quad$ $(H = I_n - \frac{1}{n}1_n 1_n^\top)$

A statistical test using HSIC

- Empirical HSIC (simplified biased estimator)

$$\widehat{HSIC}^2 = \frac{1}{n^2}\mathrm{trace}(K H L H)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ $\quad (H = I_n - \frac{1}{n}1_n 1_n^\top)$

Statistical hypothesis test:

- Null hypothesis $\mathcal{H}_0$ when $P_{XY} = P_X P_Y$
  - should see $\widehat{HSIC}^2$ close to zero.
- Alternative hypothesis $\mathcal{H}_1$ when $P_{XY} \neq P_X P_Y$
  - should see $\widehat{HSIC}^2$ far from zero.

A statistical test using HSIC

- Empirical HSIC (simplified biased estimator)

$$\widehat{HSIC}^2 = \frac{1}{n^2}\text{trace}(K\,H\,L\,H)$$

$K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i y_j)$ $\qquad (H = I_n - \frac{1}{n}1_n 1_n^\top)$

Statistical hypothesis test:

- Null hypothesis $\mathcal{H}_0$ when $P_{XY} = P_X P_Y$
  - should see $\widehat{HSIC}^2$ close to zero.
- Alternative hypothesis $\mathcal{H}_1$ when $P_{XY} \neq P_X P_Y$
  - should see $\widehat{HSIC}^2$ far from zero.

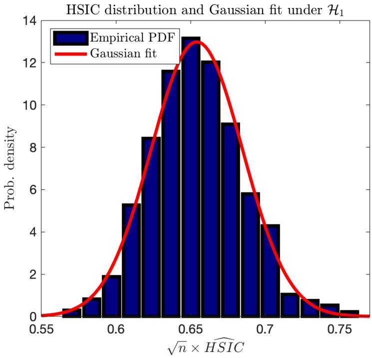
Want threshold c_α for $\widehat{HSIC}^2$ to get false positive rate α

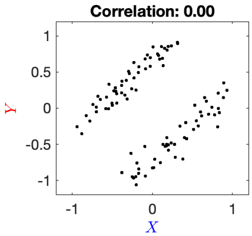
Asymptotics of $\widehat{HSIC}^2$ when $P_{XY} \neq P_X P_Y$

When $P_{XY} \neq P_X P_Y$, statistic is asymptotically normal,

$$\sqrt{n}\left(\widehat{\mathrm{HSIC}}^2 - \mathrm{HSIC}^2(P_{XY})\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2(P_{XY}))$$

with variance $\sigma^2(P_{XY})$.

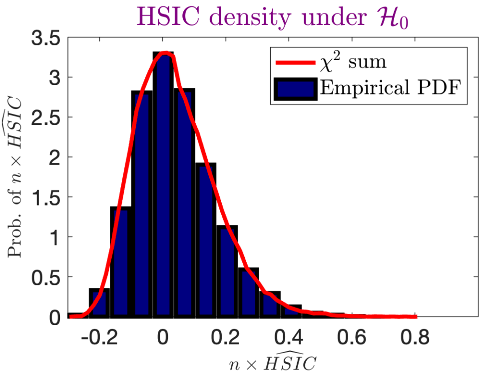




Asymptotics of $\widehat{HSIC}^2$ when $P_{XY} = P_X P_Y$

Where $P_{XY} = P_X P_Y$, statistic has asymptotic distribution

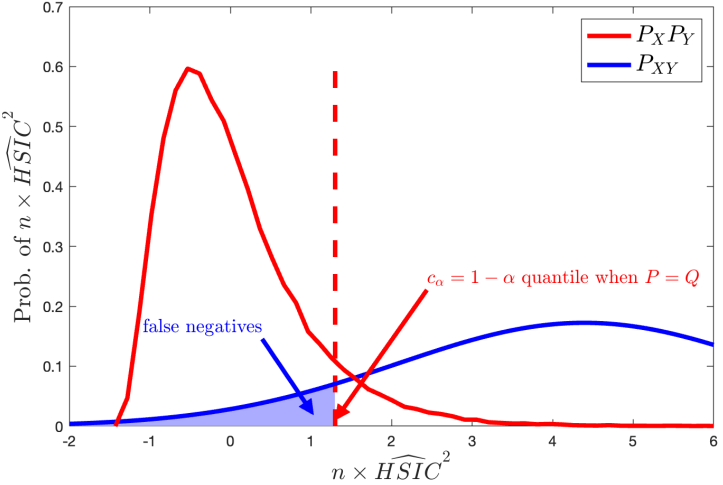
$$n\widehat{HSIC}^2 \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i \left(z_i^2 - 1\right), \qquad z_i \sim \mathcal{N}(0,1)\text{i.i.d.}$$



$\{\lambda_i\}_{i=1}^{\infty}$ a complicated function of P_{XY}, k, l.

A statistical test

Test construction



A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α (prob. of false positive)?

- Original sample:

$$\begin{array}{cccccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 & Y_8 & Y_9 & Y_{10} \end{array}$$

- Permutation:

$$\begin{array}{cccccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \\ Y_7 & Y_3 & Y_9 & Y_2 & Y_4 & Y_8 & Y_5 & Y_1 & Y_6 & Y_{10} \end{array}$$

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α (prob. of false positive)?

- Original sample:

$$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}$$
$$Y_1\ Y_2\ Y_3\ Y_4\ Y_5\ Y_6\ Y_7\ Y_8\ Y_9\ Y_{10}$$

- Permutation:

$$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}$$
$$Y_7\ Y_3\ Y_9\ Y_2\ Y_4\ Y_8\ Y_5\ Y_1\ Y_6\ Y_{10}$$

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold c_α is $1 - \alpha$ quantile of empirical CDF

A statistical test

- Given $P_{XY} = P_X P_Y$, what is the threshold c_α such that $P(\widehat{HSIC}^2 > c_\alpha) < \alpha$ for small α (prob. of false positive)?

- Original sample:

$$\begin{array}{cccccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 & Y_8 & Y_9 & Y_{10} \end{array}$$

- Permutation:

$$\begin{array}{cccccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \\ Y_7 & Y_3 & Y_9 & Y_2 & Y_4 & Y_8 & Y_5 & Y_1 & Y_6 & Y_{10} \end{array}$$

- Null distribution via permutation
  - Compute HSIC for $\{x_i, y_{\pi(i)}\}_{i=1}^n$ for random permutation π of indices $\{1, \ldots, n\}$. This gives HSIC for independent variables.
  - Repeat for many different permutations, get empirical CDF
  - Threshold c_α is $1 - \alpha$ quantile of empirical CDF

Application: dependence detection across languages

Testing task: detect dependence between English and French text

X	Y
Honourable senators, I have a question for the Leader of the Government in the Senate	Honorables sénateurs, ma question s’adresse au leader du gouvernement au Sénat
No doubt there is great pressure on provincial and municipal governments	Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions
In fact, we have increased federal investments for early childhood development.	Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes
⋮	⋮

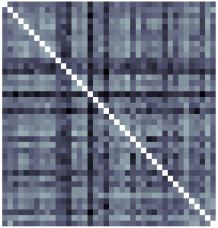
Text from the aligned hansards of the 36th parliament of canada,
https://www.isi.edu/natural-language/download/hansard/

Application: dependence detection across languages

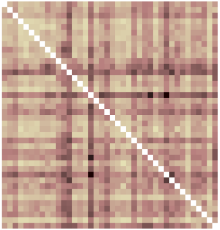
Testing task: detect dependence between English and French text

k-spectrum kernel, $k = 10$, sample size $n = 10$

X	Y
Honourable senators, I have a question for the Leader of the Government in the Senate	Honorables sénateurs, ma question s'adresse au leader du gouvernement au Sénat
No doubt there is great pressure on provincial and municipal governments	Les ordres de gouvernements provinciaux et municipaux subissent de fortes pressions
In fact, we have increased federal investments for early childhood development.	Au contraire, nous avons augmenté le financement fédéral pour le développement des jeunes
⋮	⋮



K



L

$$\widehat{HSIC}^2 = \frac{1}{n^2}\mathrm{trace}(K H L H)$$

$H = I_n - \frac{1}{n}1_n 1_n^\top$

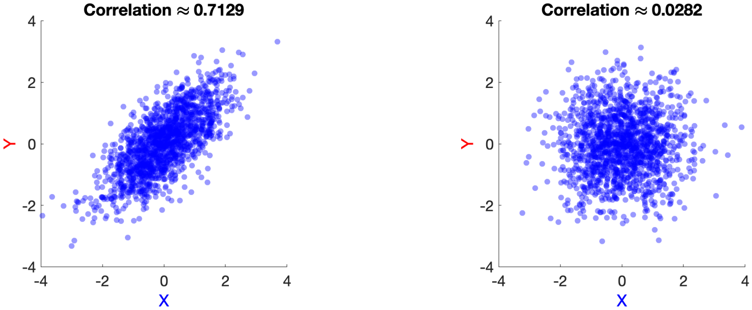
Application:Dependence detection across languages

Results (for $\alpha = 0.05$)

- k-spectrum kernel: average Type II error 0
- Bag of words kernel: average Type II error 0.18

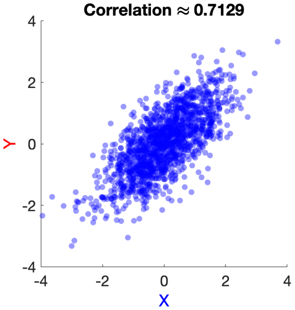
Settings: Five line extracts, averaged over 300 repetitions, for "Agriculture" transcripts. Similar results for Fisheries and Immigration transcripts.

What about the mutual information?

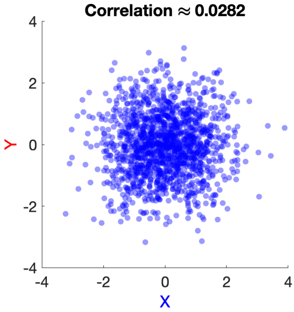


What about the mutual information?

Mutual information: 0.5 bits



Mutual information: ??

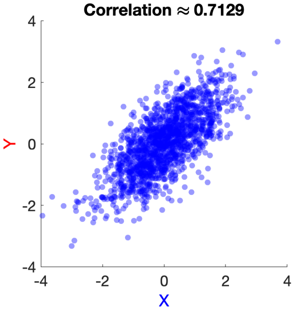


$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

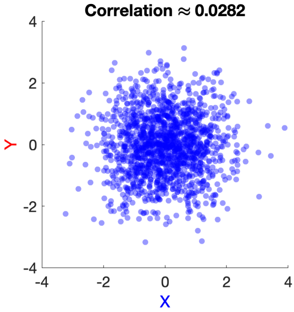
$$\rho = 0.5$$

What about the mutual information?

Mutual information: 0.5 bits



Mutual information: 6.6 bits



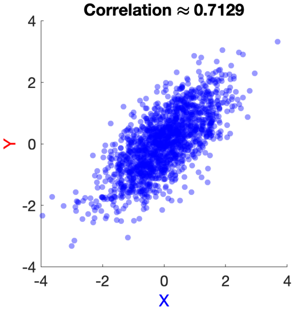
$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

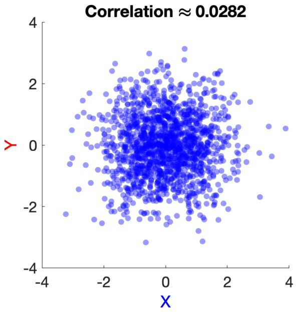
$\rho = 0.5$

What about the mutual information?

Mutual information: 0.5 bits

Mutual information: 6.6 bits





$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$\rho = 0.5$

$x \sim \mathcal{N}(0, 1)$

y is independent Gaussian noise in the leading 5 digits, then digits of x with tiny noise added.

What about the mutual information?

- The MI has no unbiased estimator [Paninski, 2003, Proposition 8]
  - HSIC has an unbiased estimator and well known asymptotics

What about the mutual information?

- The MI has no unbiased estimator [Paninski, 2003, Proposition 8]
  - HSIC has an unbiased estimator and well known asymptotics

In continuous spaces:

- The MI is probably not what you want (eg as a loss for learning simple NN features)

What about the mutual information?

- The MI has no unbiased estimator [Paninski, 2003, Proposition 8]
  - HSIC has an unbiased estimator and well known asymptotics

In continuous spaces:

- The MI is probably not what you want (eg as a loss for learning simple NN features)
- What you're computing is not the MI
  - but it might still have useful properties!
- Better: solve the problem you mean to solve!

What about the mutual information?

- The MI has no unbiased estimator [Paninski, 2003, Proposition 8]
    - HSIC has an unbiased estimator and well known asymptotics

In continuous spaces:

- The MI is probably not what you want (eg as a loss for learning simple NN features)
- What you're computing is not the MI
    - but it might still have useful properties!
- Better: solve the problem you mean to solve!

Self-Supervised Learning with Kernel Dependence Maximization

Yazhe Li*
DeepMind and Gatsby Unit, UCL
yazhe@google.com

Roman Pogodin*
Gatsby Unit, UCL
roman.pogodin.17@ucl.ac.uk

Danica J. Sutherland
UBC and Amii†
dsuth@cs.ubc.ca

Arthur Gretton
Gatsby Unit, UCL
arthur.gretton@gmail.com

- Proves that InfoNCE (a lower bound on MI) approximates SSL-HSIC with a variance-based regularization,
- SSL-HSIC achieves performance on par with SimCLR, BYOL, etc on SSL benchmarks (NeurIPS2021).

Testing conditional independence

Conditional Independence Testing

arXiv > cs > arXiv:2402.13196

Search... Help | Adv

Computer Science > Machine Learning

[Submitted on 20 Feb 2024 (v1), last revised 19 Sep 2025 (this version, v2)]

Practical Kernel Tests of Conditional Independence

Roman Pogodin, Antonin Schrab, Yazhe Li, Danica J. Sutherland, Arthur Gretton

JMLR (in revision)

arXiv > stat > arXiv:2512.14000

Search... Help | Adv

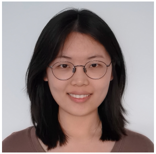
Statistics > Machine Learning

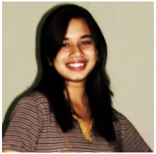
[Submitted on 16 Dec 2025]

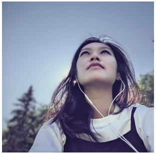
On the Hardness of Conditional Independence Testing In Practice

Zheng He, Roman Pogodin, Yazhe Li, Namrata Deka, Arthur Gretton, Danica J. Sutherland

NeurIPS (2025)







Conditional independence testing...

Kernel-based Conditional Independence Test and Application in Causal Discovery

Kun Zhang **Jonas Peters** **Dominik Janzing** **Bernhard Schölkopf**
Max Planck Institute for Intelligent Systems
Spemannstr. 38, 72076 Tübingen
Germany

KCI, from UAI 2011

…is impossible…

The Annals of Statistics
2020, Vol. 48, No. 3, 1514–1538
https://doi.org/10.1214/19-AOS1857

THE HARDNESS OF CONDITIONAL INDEPENDENCE TESTING AND THE GENERALISED COVARIANCE MEASURE

BY RAJEN D. SHAH[1] AND JONAS PETERS[2]

Annals of Statistics, 2020

Why is CI testing so hard in practice

- It's tricky to tune parameters
- You need way more samples than you think (and not just for parameter tuning)
- If you don't have enough samples, your test will fail in a very bad way (elevated false positives)

Conditional (in)dependence

[A, Theorem 2.2, following from B]

$X \perp\!\!\!\perp Y \mid Z$ iff for all $f \in L^2_X$, $g \in L^2_Y$, $w \in L^2_Z$

$$\mathrm{E}_Z\left[\underbrace{w(Z)}_{\text{TBD}} \underbrace{\mathop{\mathbb{E}}_{XY}\left[\left(f(X) - \mathrm{E}\left[f(X) \mid Z\right]\right)\left(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\right] = 0.$$

[A] He, Pogodin, Li, Deka, G, Sutherland (2025)
[B] Daudin (1980)

Conditional (in)dependence

[A, Theorem 2.2, following from B]

$X \perp\!\!\!\perp Y \mid Z$ iff for all $f \in L^2_X$, $g \in L^2_Y$, $w \in L^2_Z$

$$\mathrm{E}_Z\Bigg[\underbrace{w(Z)}_{\text{TBD}}\, \underbrace{\mathop{\mathbb{E}}_{XY}\left[\left(f(X) - \mathrm{E}[f(X) \mid Z]\right)\left(g(Y) - \mathrm{E}[g(Y) \mid Z]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\Bigg] = 0.$$

Let's kernelize! Replace $f \in L^2_X$ with $f \in \mathcal{F}$

$$f(X) - \mathrm{E}[f(X) \mid Z] = \langle f, \varphi(X) - \underbrace{\mathrm{E}[\varphi(X)|Z]}_{\mu_{X|Z}} \rangle_{\mathcal{F}}$$

[A] He, Pogodin, Li, Deka, G, Sutherland (2025)
[B] Daudin (1980)

Conditional (in)dependence

[A, Theorem 2.2, following from B]

$X \perp\!\!\!\perp Y \mid Z$ iff for all $f \in L^2_X$, $g \in L^2_Y$, $w \in L^2_Z$

$$\mathrm{E}_Z\Big[\underbrace{w(Z)}_{\text{TBD}} \underbrace{\mathop{\mathbb{E}}_{XY}\left[\left(f(X) - \mathrm{E}\left[f(X) \mid Z\right]\right)\left(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\Big] = 0.$$

Let's kernelize! Replace $f \in L^2_X$ with $f \in \mathcal{F}$

$$f(X) - \mathrm{E}\left[f(X) \mid Z\right] = \langle f, \varphi(X) - \underbrace{\mathrm{E}[\varphi(X)|Z]}_{\mu_{X|Z}} \rangle_{\mathcal{F}}$$

The conditional cross-covariance operator:

$$\mathfrak{C}_{XY|z} = \mathrm{E}\left[\left(\varphi(X) - \mu_{X|Z}\right) \otimes \left(\varphi(Y) - \mu_{Y|Z}\right) \mid Z = z\right].$$

Covariance of feature residuals after regressing on Z.

[A] He, Pogodin, Li, Deka, G, Sutherland (2025)
[B] Daudin (1980)

Kernel test statistic

Define three kernels:

$$k(x, x'), \quad l(y, y'), \quad m(z, z') = \langle \theta(z), \theta(z') \rangle_{\mathcal{M}}.$$

[A] Zhang, Peters, Janzing, Schoelkopf (2011)

Kernel test statistic

Define three kernels:

$$k(x, x'), \quad l(y, y'), \quad m(z, z') = \langle \theta(z), \theta(z') \rangle_{\mathcal{M}}.$$

KCI operator [A]: averages $\mathfrak{C}_{X\,Y|z}$ over Z: $\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{X\,Y|Z} \otimes \theta(Z)\right]$,

$$\begin{aligned}&\langle f \otimes g, \mathfrak{C}_{KCI}\, w\rangle_{\mathrm{HS}} \\ &= \mathrm{E}\left[w(Z)\mathrm{E}\left[(f(X) - \mathrm{E}\left[f(X) \mid Z\right])(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]) \mid Z\right]\right]\end{aligned}$$

[A] Zhang, Peters, Janzing, Schoelkopf (2011)

Kernel test statistic

Define three kernels:

$$k(x,x'),\quad l(y,y'),\quad m(z,z') = \langle \theta(z), \theta(z') \rangle_{\mathcal{M}}.$$

KCI operator [A]: averages $\mathfrak{C}_{X\,Y|z}$ over Z: $\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{X\,Y|Z} \otimes \theta(Z)\right]$,

$$\begin{aligned}
&\langle f \otimes g, \mathfrak{C}_{KCI}\, w \rangle_{\mathrm{HS}} \\
&= \mathrm{E}\left[w(Z)\mathrm{E}\left[(f(X) - \mathrm{E}\left[f(X) \mid Z\right])(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]) \mid Z\right]\right]
\end{aligned}$$

KCI definition

$$\begin{aligned}
KCI^2 &:= \|\mathfrak{C}_{KCI}\|^2_{HS} \\
&= \mathrm{E}\left[m(Z,Z')\, k^{\mu}((X,Z),(X',Z'))\, l^{\mu}((Y,Z),(Y',Z'))\right]
\end{aligned}$$

with centred kernel

$$k^{\mu}((X,Z),(X',Z')) = \langle \varphi(X) - \mu_{X|Z}, \varphi(X') - \mu_{X|Z'} \rangle_{\mathcal{F}}$$

[A] Zhang, Peters, Janzing, Schoelkopf (2011)

Kernel test statistic

Define three kernels:

$$k(x, x'), \quad l(y, y'), \quad m(z, z') = \langle \theta(z), \theta(z') \rangle_{\mathcal{M}}.$$

KCI operator [A]: averages $\mathfrak{C}_{XY|z}$ over Z: $\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z} \otimes \theta(Z)\right]$,

$$\begin{aligned}&\langle f \otimes g, \mathfrak{C}_{KCI}\, w \rangle_{\text{HS}} \\ &= \mathrm{E}\left[w(Z)\mathrm{E}\left[(f(X) - \mathrm{E}\left[f(X) \mid Z\right])(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]) \mid Z\right]\right]\end{aligned}$$

KCI definition

$$\begin{aligned}KCI^2 &:= \|\mathfrak{C}_{KCI}\|^2_{HS} \\ &= \mathrm{E}\left[m(Z, Z')\, k^{\mu}((X, Z), (X', Z'))\, l^{\mu}((Y, Z), (Y', Z'))\right]\end{aligned}$$

with centred kernel

$$k^{\mu}((X, Z), (X', Z')) = \langle \varphi(X) - \mu_{X|Z}, \varphi(X') - \mu_{X|Z'} \rangle_{\mathcal{F}}$$

For characteristic k, l, m,

$$KCI^2 = 0 \quad \text{iff} \quad X \perp\!\!\!\perp Y \mid Z$$

[A] Zhang, Peters, Janzing, Schoelkopf (2011)

Testing Conditional Independence

From last slide:

$$KCI^2 = \mathrm{E}\left[m(Z, Z')\, k^{\mu}((X, Z), (X', Z'))\, l^{\mu}((Y, Z), (Y', Z'))\right]$$

where

$$k^{\mu}((X, Z), (X', Z')) = \langle \varphi(X) - \mu_{X|Z}, \varphi(X') - \mu_{X|Z'} \rangle_{\mathcal{F}}$$
$$l^{\mu}((Y, Z), (Y', Z')) = \langle \varphi(Y) - \mu_{Y|Z}, \varphi(Y' - \mu_{Y|Z'} \rangle_{\mathcal{G}}$$

[A] Zhang, Peters, Janzing, Schoelkopf (2011)

Testing Conditional Independence

From last slide:

$$KCI^2 = \mathrm{E}\left[m(Z, Z')\; k^\mu((X, Z), (X', Z'))\; l^\mu((Y, Z), (Y', Z'))\right]$$

where

$$k^\mu((X, Z), (X', Z')) = \langle \varphi(X) - \mu_{X|Z}, \varphi(X') - \mu_{X|Z'} \rangle_{\mathcal{F}}$$
$$l^\mu((Y, Z), (Y', Z')) = \langle \varphi(Y) - \mu_{Y|Z}, \varphi(Y' - \mu_{Y|Z'} \rangle_{\mathcal{G}}$$

U-statistic estimator

$$\widehat{\mathrm{KCI}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} m(z_i, z_j)\; k^\mu((x_i, z_i), (x_j, z_j))\; l^\mu((y_i, z_i), (y_j, z_j))$$

[A] Zhang, Peters, Janzing, Schoelkopf (2011)

Testing Conditional Independence

From last slide:

$$KCI^2 = \mathrm{E}\left[m(Z, Z')\, k^{\mu}((X, Z), (X', Z'))\, l^{\mu}((Y, Z), (Y', Z'))\right]$$

where

$$k^{\mu}((X, Z), (X', Z')) = \langle \varphi(X) - \mu_{X|Z}, \varphi(X') - \mu_{X|Z'} \rangle_{\mathcal{F}}$$
$$l^{\mu}((Y, Z), (Y', Z')) = \langle \varphi(Y) - \mu_{Y|Z}, \varphi(Y' - \mu_{Y|Z'} \rangle_{\mathcal{G}}$$

U-statistic estimator

$$\widehat{\mathrm{KCI}}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} m(z_i, z_j)\, k^{\mu}((x_i, z_i), (x_j, z_j))\, l^{\mu}((y_i, z_i), (y_j, z_j))$$

Challenges:

- What does $m(Z, Z')$ do? How to choose its bandwidth?
- We don't know $\mu_{X|z} = \mathrm{E}_X[\varphi(X)|Z = z]$ or $\mu_{Y|z} = \mathrm{E}_Y[\phi(Y)|Z = z]$ - need to learn from data!

[A] Zhang, Peters, Janzing, Schoelkopf (2011)

Purpose of $m(Z, Z') = \langle \theta(Z), \theta(Z') \rangle_{\mathcal{M}}$

- A toy model:

$$X = f_X(Z) + \tau\, r_X, \quad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sin(\beta Z) \\ \sin(\beta Z) & 1 \end{bmatrix} \right),$$

Fine print: assume perfect $\mu_{X|z} = \mathrm{E}_X\left[\varphi(X) | Z = z\right]$ and $\mu_{Y|z} = \mathrm{E}_Y\left[\phi(Y) | Z = z\right]$.

Purpose of $m(Z, Z') = \langle \theta(Z), \theta(Z') \rangle_{\mathcal{M}}$

- A toy model:

$$X = f_X(Z) + \tau\, r_X, \qquad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sin(\beta Z) \\ \sin(\beta Z) & 1 \end{bmatrix} \right),$$

- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$

Fine print: assume perfect $\mu_{X|z} = \mathrm{E}_X\left[\varphi(X) | Z = z\right]$ and $\mu_{Y|z} = \mathrm{E}_Y\left[\phi(Y) | Z = z\right]$.

Purpose of $m(Z, Z') = \langle \theta(Z), \theta(Z') \rangle_{\mathcal{M}}$

- A toy model:

$$X = f_X(Z) + \tau\, r_X, \qquad Y = f_Y(Z) + \tau\, r_Y,$$

where f_X, f_Y are fixed functions, $\tau > 0$, $\beta \geq 0$, and the residual terms (r_X, r_Y) follow

$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sin(\beta Z) \\ \sin(\beta Z) & 1 \end{bmatrix} \right),$$

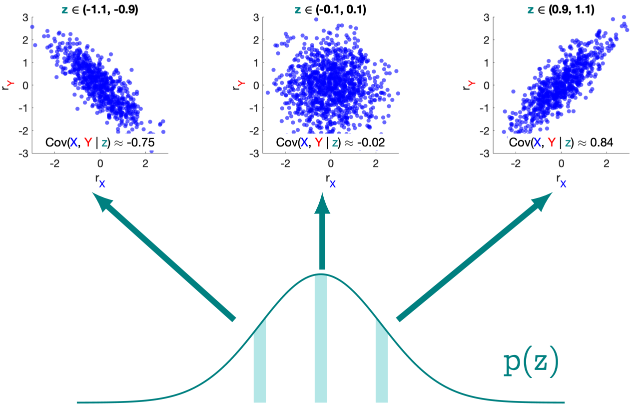
- Here $X \perp\!\!\!\perp Y \mid Z$ iff $\beta = 0$
- We take linear kernels on X and Y, and define kernel on z with lengthscale ℓ,

$$m_\ell(z, z') = \exp\left(-\ell^{-2} \|z - z'\|^2 \right)$$

Fine print: assume perfect $\mu_{X|z} = \mathrm{E}_X \left[\varphi(X) | Z = z \right]$ and $\mu_{Y|z} = \mathrm{E}_Y \left[\phi(Y) | Z = z \right]$.

Choosing bandwidth ℓ of $m_\ell(z, z')$

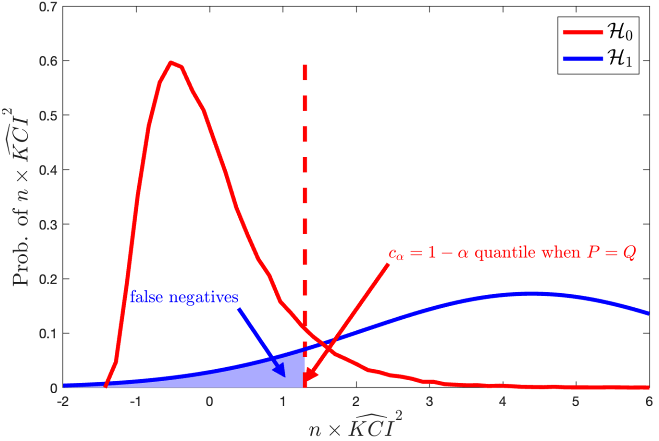
$m_\ell(z, z') \implies$ neighbourhood of Z for which $\mathfrak{C}_{X\,Y|Z} \approx$ constant.



Distribution of residuals: $(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \sin(\beta Z)\\ \sin(\beta Z) & 1\end{bmatrix}\right)$

Asymptotics of KCI

Choose bandwidth ℓ of $m_\ell(Z, Z')$ to maximize test power



Fine print: assume perfect $\mu_{X|z} = \mathbf{E}_X[\varphi(X)|Z = z]$ and $\mu_{Y|z} = \mathbf{E}_Y[\phi(Y)|Z = z]$.

Choose ℓ to maximize test power proxy

The power of our test:

$$\Pr_{\mathcal{H}_1}\left(n\widehat{\mathrm{KCI}}^2 > \hat{q}^{\ell}_{1-\alpha}\right)$$

Fine print: assume perfect $\mu_{X|z} = \mathbb{E}_X\left[\varphi(X)|Z=z\right]$ and $\mu_{Y|z} = \mathbb{E}_Y\left[\phi(Y)|Z=z\right]$.

Choose ℓ to maximize test power proxy

The power of our test:

$$\Pr_{\mathcal{H}_1}\left(n\widehat{\mathrm{KCI}}^2 > \hat{q}^{\ell}_{1-\alpha}\right)$$
$$\to \Phi\left(\frac{\mathrm{KCI}^2}{\sqrt{n^{-1}\sigma^2_{\ell}}} - \frac{\hat{q}^{\ell}_{1-\alpha}}{n\sqrt{n^{-1}\sigma^2_{\ell}}}\right)$$

where

- Φ is the CDF of the standard normal distribution.
- σ^2_{ℓ} is asymptotic variance of $\sqrt{n}\left(\widehat{\mathrm{KCI}}^2 - \mathrm{KCI}^2\right)$
- $\hat{q}^{\ell}_{1-\alpha}$ is an estimate of $q^{\ell}_{1-\alpha}$ test threshold.

Fine print: assume perfect $\mu_{X|z} = \mathrm{E}_X\left[\varphi(X)|Z = z\right]$ and $\mu_{Y|z} = \mathrm{E}_Y\left[\phi(Y)|Z = z\right]$.

Choose ℓ to maximize test power proxy

The power of our test:

$$\Pr_{\mathcal{H}_1}\left(n\widehat{\mathrm{KCI}}^2 > \hat{q}^{\ell}_{1-\alpha}\right) \to \Phi\left(\underbrace{\frac{\mathrm{KCI}^2}{\sqrt{n^{-1}\sigma_\ell^2}}}_{O(n^{1/2})} - \underbrace{\frac{\hat{q}^{\ell}_{1-\alpha}}{n\sqrt{n^{-1}\sigma_\ell^2}}}_{O(n^{-1/2})}\right)$$

For large n, second term negligible!

Fine print: assume perfect $\mu_{X|z} = \mathbf{E}_X[\varphi(X)|Z=z]$ and $\mu_{Y|z} = \mathbf{E}_Y[\phi(Y)|Z=z]$.

Choose ℓ to maximize test power proxy

The power of our test:

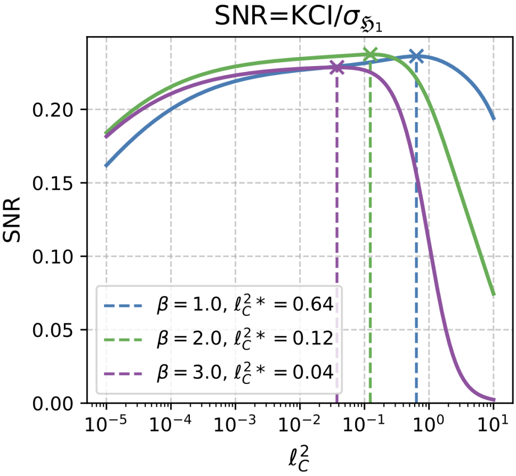
$$\begin{aligned}
&\Pr_{\mathcal{H}_1}\left(n\widehat{\mathrm{KCI}}^2 > \hat{q}^{\ell}_{1-\alpha}\right) \\
&\to \Phi\left(\frac{\mathrm{KCI}^2}{\sqrt{n^{-1}\sigma^2_{\ell}}} - \frac{\hat{q}^{\ell}_{1-\alpha}}{n\sqrt{n^{-1}\sigma^2_{\ell}}}\right)
\end{aligned}$$

To maximize test power, maximize (on held-out set)

$$\ell^{\star} := \operatorname*{argmax}_{\ell \in \mathcal{L}} \frac{\widehat{\mathrm{KCI}}^2}{\sqrt{\hat{\sigma}^2_{\ell}}}$$

Fine print: assume perfect $\mu_{X|z} = \mathrm{E}_X\left[\varphi(X)|Z = z\right]$ and $\mu_{Y|z} = \mathrm{E}_Y\left[\phi(Y)|Z = z\right]$.

Choose ℓ to maximize SNR



$$(r_X, r_Y) \mid Z \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sin(\beta Z) \\ \sin(\beta Z) & 1 \end{bmatrix}\right) \qquad m_\ell(z, z') = \exp\left(-\ell^{-2}\|z - z'\|^2\right)$$

Fine print: assume perfect $\mu_{X|z} = \mathbf{E}_X[\varphi(X)|Z = z]$ and $\mu_{Y|z} = \mathbf{E}_Y[\phi(Y)|Z = z]$.

CI: the regression challenge

Recall the test statistic

$\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z} \otimes \theta(Z)\right]$ so

$$\langle f \otimes g, \mathfrak{C}_{KCI}\, w\rangle$$
$$= \mathrm{E}_Z\left[w(Z) \underbrace{\mathop{\mathbb{E}}_{XY}\left[\left(f(X) - \mathrm{E}\left[f(X) \mid Z\right]\right)\left(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\right] = 0.$$

Residual term for linear kernel on X:

$$\varphi(X) - \mu_{X|Z} = X - \mathrm{E}[X|Z]$$

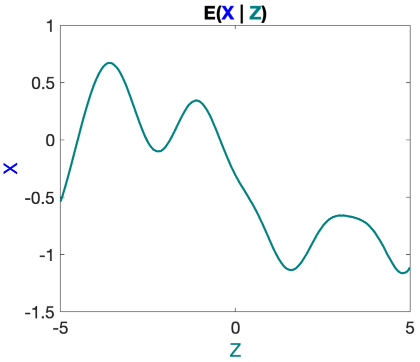
Regression errors:

$$\Delta_{X|Z} := \widehat{\mathrm{E}}[X|Z] - \mathrm{E}[X|Z]$$
$$\Delta_{Y|Z} := \widehat{\mathrm{E}}[Y|Z] - \mathrm{E}[Y|Z]$$

$\hat{\mathrm{E}}[X|Z] - \mathrm{E}[X|Z]$ for kernel ridge regression

$\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z} \otimes \theta(Z)\right]$ so

$$\langle f \otimes g, \mathfrak{C}_{KCI}\, w\rangle$$
$$= \mathrm{E}_Z\left[w(Z)\, \underbrace{\mathbb{E}_{XY}\left[\left(f(X) - \mathrm{E}\left[f(X) \mid Z\right]\right)\left(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\right] = 0.$$



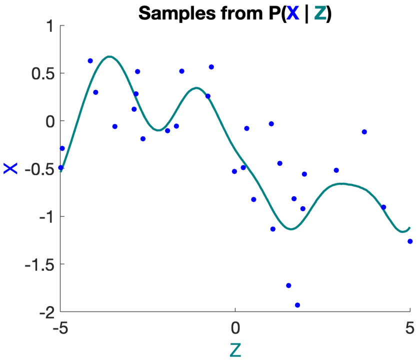
Residual term for linear kernel:

$$X - \mathrm{E}[X|Z]$$

$\hat{\mathrm{E}}[X|Z] - \mathrm{E}[X|Z]$ for kernel ridge regression

$\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z} \otimes \theta(Z)\right]$ so

$$\langle f \otimes g, \mathfrak{C}_{KCI}\, w\rangle$$
$$= \mathrm{E}_Z\left[w(Z) \underbrace{\mathop{\mathbb{E}}_{XY}\left[\left(f(X) - \mathrm{E}\left[f(X) \mid Z\right]\right)\left(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\right] = 0.$$



Residual term for linear kernel:

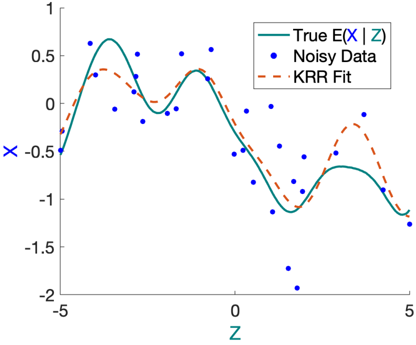
$$X - \mathrm{E}[X|Z]$$

$\hat{\mathrm{E}}[X|Z] - \mathrm{E}[X|Z]$ for kernel ridge regression

$\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z} \otimes \theta(Z)\right]$ so

$$\langle f \otimes g, \mathfrak{C}_{KCI}\, w \rangle$$

$$= \mathrm{E}_Z\left[w(Z) \underbrace{\mathbb{E}_{XY}\left[\left(f(X) - \mathrm{E}\left[f(X) \mid Z\right]\right)\left(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\right] = 0.$$



Residual term for linear kernel:

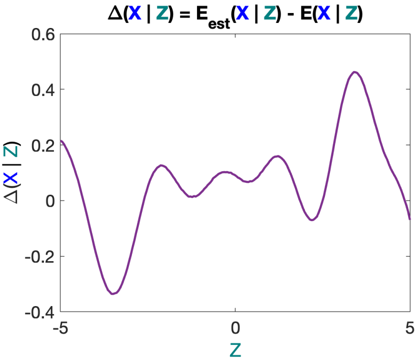
$$X - \mathrm{E}[X|Z]$$

$\hat{\mathrm{E}}[X|Z] - \mathrm{E}[X|Z]$ for kernel ridge regression

$$\mathfrak{C}_{KCI} = \mathbb{E}\left[\mathfrak{C}_{XY|Z} \otimes \theta(Z)\right] \text{ so}$$

$$\langle f \otimes g, \mathfrak{C}_{KCI}\, w\rangle$$

$$= \mathrm{E}_Z\left[w(Z) \underbrace{\mathop{\mathbb{E}}_{XY}\left[\left(f(X) - \mathrm{E}\left[f(X) \mid Z\right]\right)\left(g(Y) - \mathrm{E}\left[g(Y) \mid Z\right]\right) \mid Z\right]}_{\text{conditional covariance of residuals}}\right] = 0.$$



Residual term for linear kernel:

$$X - \mathrm{E}[X|Z]$$

Regression errors:

$$\Delta_{X|Z} := \hat{\mathrm{E}}[X|Z] - \mathrm{E}[X|Z]$$
$$\Delta_{Y|Z} := \hat{\mathrm{E}}[Y|Z] - \mathrm{E}[Y|Z]$$

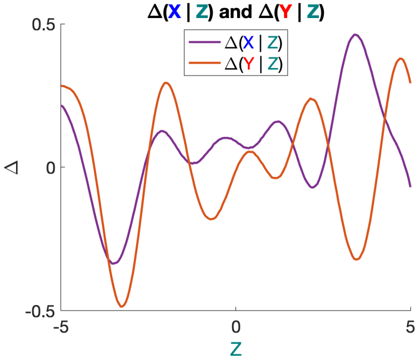
Recall the test statistic

KCI^2_Δ is population statistic with regression errors:

$$\Delta_{X|Z} := \widehat{\mathrm{E}}[X|Z] - \mathrm{E}[X|Z] \qquad \Delta_{Y|Z} := \widehat{\mathrm{E}}[Y|Z] - \mathrm{E}[Y|Z]$$

Then under conditional independence [A, eq. (10)]

$$\mathrm{KCI}^2_\Delta = \mathrm{E}\left[m(Z, Z')\left[\Delta_{X|Z}\Delta_{X|Z'}\right]\left[\Delta_{Y|Z}\Delta_{Y|Z'}\right]\right] \neq 0.$$



[A] Pogodin, Schrab, Li, Sutherland, G (2026)

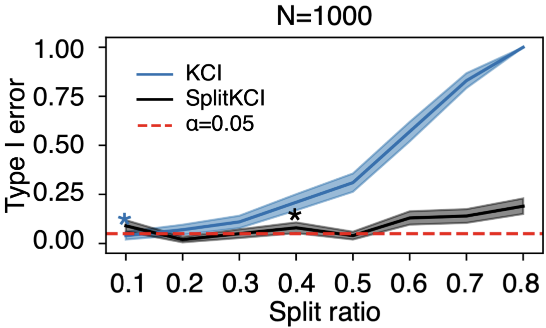
Mitigations

- SplitKCI: compute $\mu^{(1)}_{X|Z}, \mu^{(2)}_{X|Z}$ on two <u>different</u> splits, substitute

$$\tilde{k}^{\mu}((X, Z), (X', Z')) = \langle \varphi(X) - \mu^{(1)}_{X|Z}, \varphi(X') - \mu^{(2)}_{X|Z'} \rangle_{\mathcal{F}}$$

  in KCI (same for $\tilde{l}^{\mu}((Y, Z), (Y', Z'))$)
- Use enough samples for $\mu_{X|Z}, \mu_{Y|Z}$ to converge

Synthetic neuroscience experiment:



[A, Definition 4] Pogodin, Schrab, Li, Sutherland, G (2026)

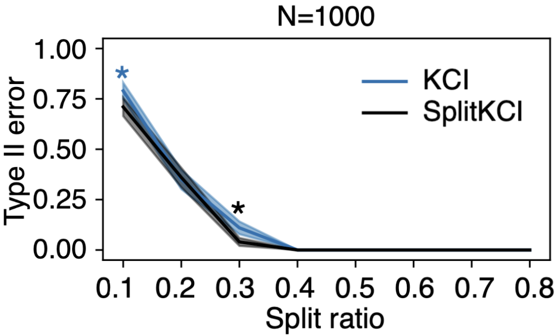
Mitigations

- SplitKCI: compute $\mu^{(1)}_{X|Z}, \mu^{(2)}_{X|Z}$ on two different splits, substitute

$$\tilde{k}^{\mu}((X, Z), (X', Z')) = \langle \varphi(X) - \mu^{(1)}_{X|Z}, \varphi(X') - \mu^{(2)}_{X|Z'} \rangle_{\mathcal{F}}$$

  in KCI (same for $\tilde{l}^{\mu}((Y, Z), (Y', Z'))$)
- Use enough samples for $\mu_{X|Z}, \mu_{Y|Z}$ to converge

Synthetic neuroscience experiment:



[A, Definition 4] Pogodin, Schrab, Li, Sutherland, G (2026)

Conclusions

Statistical tests of dependence

- Hilbert-Schmidt norm of feature covariance (HSIC)
- Hypothesis test

Statistical tests of conditional dependence

- Condition for conditional independence
- Test statistic (KCI)
- Need held-out data to learn test parameters
- Need more held-out data to learn regression functions

Research support

Work supported by:

The Gatsby Charitable Foundation

Google Deepmind

Questions?


THE KERNEL
BREWERY
LONDON
BALTIC PORTER
THE KERNEL
EXPORT STOUT