Kernel measures of (conditional) dependence

a foundational statistical problem

Danica J. Sutherland (she/her) – University of British Columbia & Amii

Based on work with **Zheng He** (UBC), **Roman Pogodin** (UCL → McGill/Mila → Google), Antonin Schrab (UCL → Cambridge), Yazhe Li (UCL/DeepMind → Microsoft), Namrata Deka (UBC → CMU), Arthur Gretton (UCL/DeepMind), **Nathaniel Xu** (UBC), Feng Liu (Melbourne), **Aaron Wei** (UBC).

He et al. (2026), merging He et al. (2025) and Pogodin et al. (2024), in prep.
Pointwise testing results (Sutherland 2026): on arXiv very soon.
Mentioned extensions (Nate & Feng; Aaron): very early work.

BIRS workshop on Kernel Approximation and Gaussian Processes – Sep 2026

Why care about conditional independence?

Does X tell us anything about Y once we know Z?

$$X \perp\!\!\!\perp Y \mid Z \qquad \text{iff} \qquad P_{XY|Z} = P_{X|Z} \otimes P_{Y|Z},\ Z\text{-almost surely}$$

Comparing distributions

A problem in its own right: “two-sample testing” or Mu Zhu’s talk yesterday.

For this talk, a building block.

Maximum Mean Discrepancy

Assume a kernel k on $\mathcal{X}$ with RKHS $\mathcal{H}$ and feature map $\varphi(x) = k(x, \cdot)$.

Mean embedding of a distribution P (Berlinet and Thomas-Agnan 2004):

$$\mu_P := P\varphi = \mathbb{E}_{X\sim P}\,\varphi(X) = \int k(x,\cdot)\,\mathrm{d}P(x) \in \mathcal{H}, \quad \langle \mu_P, f\rangle_{\mathcal{H}} = Pf \text{ for } f \in \mathcal{H}.$$

Can use this to define a distance on distributions (Gretton et al. 2012):

$$\mathrm{MMD}(P,Q) := \|\mu_P - \mu_Q\|_{\mathcal{H}} = \sup_{f:\|f\|_{\mathcal{H}}=1} \langle \mu_P - \mu_Q, f\rangle = \sup_{f:\|f\|_{\mathcal{H}}=1} Pf - Qf.$$

The **power function** from kernel interpolation has $P = \delta_x, Q = \sum_i u_i(x)\delta_{x_i}$.

For a **characteristic** kernel (Sriperumbudur et al. 2011) (e.g. Gaussian, …):
$P \mapsto \mu_P$ is injective, so $\mathrm{MMD}(P,Q) = 0$ iff $P = Q$.
Characteristic kernels exist on any Polish space (Ziegel et al. 2024).

MMD in GP form

$$\mathrm{MMD}^2(P, Q) = \mathbb{E}_{\substack{X, X' \sim P \\ Y, Y' \sim Q}} \Big[k(X, X') + k(Y, Y') - 2k(X, Y) \Big]$$

$$= \mathbb{E}_{f \sim \mathcal{GP}(0, k)} \Big[\Big(Pf - Qf \Big)^2 \Big]$$

Worst case over the RKHS ball (kernel side) = **average case** under the GP prior (Ritter 2000; Kanagawa et al. 2018, sec. 6.1)

Statistics > Machine Learning

[Submitted on 6 Jul 2018]

Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, Bharath K Sriperumbudur

This paper is an attempt to bridge the conceptual gaps between researchers working on the two widely used approaches based on positive definite kernels: Bayesian learning or inference using Gaussian

Testing with MMD

- Null hypothesis: $P = Q$.
- Reject the null hypothesis if $\mathrm{MMD}(\hat{P}, \hat{Q})$ is big.
- Test has **level** α if probability of rejecting is $\leq \alpha$ when null hypothesis holds.
- Want a test with high **power**: probability of rejecting when $P \neq Q$.
- Can decide on "big" with a **permutation test**:
  - If $P = Q$, samples in X and Y are exchangeable.
  - Bunch of times: randomly re-assign to $\tilde{P}, \tilde{Q}$ and compute $\mathrm{MMD}(\tilde{P}, \tilde{Q})$.
  - Reject if $\mathrm{MMD}(\hat{P}, \hat{Q}) >$ the $1 - \alpha$ quantile of (distribution above plus real assignment).
  - Gives a level α test (Hemerik and Goeman 2018).
- You should learn a (deep) kernel! (Sutherland et al. 2017; Liu et al. 2020; Wei et al. 2026)

Testing independence

Getting closer to the problem we care about!

Hilbert-Schmidt Independence Criterion

$$X \perp\!\!\!\perp Y \qquad \text{iff} \qquad P_{XY} = P_X \otimes P_Y \qquad \mathrm{MI}(X;Y) = \mathrm{KL}(P \,\|\, P_X \otimes P_Y)$$

Take kernels k_X on $\mathcal{X}$, k_Y on $\mathcal{Y}$; $k_X \otimes k_Y$ is a kernel on $\mathcal{X} \times \mathcal{Y}$.
Its RKHS is $\mathcal{H}_X \otimes \mathcal{H}_Y \cong \mathrm{HS}(\mathcal{H}_Y, \mathcal{H}_X)$. Then (Gretton et al. 2005, 2012; Kanagawa et al. 2018)

$$\begin{aligned}
\mathrm{HSIC}(X,Y) &:= \mathrm{MMD}^2_{k_X \otimes k_Y}\Big(P_{XY},\ P_X \otimes P_Y\Big) \\
&= \Big\| \underbrace{P_{XY}(\varphi_X \otimes \varphi_Y) - P_X\varphi_X \otimes P_Y\varphi_Y}_{C_{XY} \text{ has } \langle f, C_{XY} g\rangle = \mathrm{Cov}(f(X), g(Y))} \Big\|^2_{\mathrm{HS}} \\
&= \mathbb{E}_{\substack{f \sim \mathcal{GP}(0,k_X) \\ g \sim \mathcal{GP}(0,k_Y)}} \Big[\mathrm{Cov}(f(X), g(Y))^2 \Big]
\end{aligned}$$

Testing with HSIC

- $\widehat{\mathbf{HSIC}}_n = \frac{1}{n^2}\langle K_X, HK_YH\rangle_F$ for centring matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$: $O(n^2)$ time.
- Can use same permutation idea: shuffle the Ys. Test has level α.
- If $X \perp\!\!\!\perp Y$, then $\widehat{\mathbf{HSIC}}_n = O_p(1/n)$.
- If $\mathbf{HSIC} > 0$, then $\widehat{\mathbf{HSIC}}_n \to \mathbf{HSIC} > 0$ so power $\to 1$.
- In practice, you should learn a (deep) kernel! (Xu et al. 2026)

Testing conditional independence

The problem we care about… and where things get tough.

Kernel Conditional Independence criterion

$$X \perp\!\!\!\perp Y \mid Z \qquad \text{iff} \qquad P_{XYZ} = \Pi_{\mathrm{CI}}(P_{XYZ})$$
$$\text{for} \quad \Pi_{\mathrm{CI}}(P)(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) := P_Z(\mathrm{d}z)\, P_{X|Z=z}(\mathrm{d}x)\, P_{Y|Z=z}(\mathrm{d}y)$$

Π_{CI} is the unique CI law matching (X, Z) and (Y, Z) marginals (Dawid and Lauritzen 1993; Massa and Lauritzen 2010). It minimizes $\mathrm{KL}(P_{XYZ} \,\|\, Q)$ over CI laws Q, and

$$\mathrm{CMI}(X; Y \mid Z) = \mathrm{KL}(P \,\|\, \Pi_{\mathrm{CI}} P).$$

We can define (this version from He et al. (2026); original quantity from Zhang et al. (2011))

$$\mathrm{KCI}(X, Y \mid Z) := \mathrm{MMD}^2_{k_X \otimes k_Y \otimes k_Z}\Big(P_{XYZ},\ \Pi_{\mathrm{CI}}(P_{XYZ})\Big)$$
$$= \mathbb{E}_{\substack{f \sim \mathcal{GP}(0, k_X) \\ g \sim \mathcal{GP}(0, k_Y) \\ w \sim \mathcal{GP}(0, k_Z)}} \Big[\Big(\mathbb{E}_{Z \sim P_Z} \Big[w(Z)\ \mathrm{Cov}_{P_{XY|Z}}(f(X), g(Y) \mid Z) \Big] \Big)^2 \Big]$$

KCI characterizes CI

Theorem (He et al. 2026) For bounded continuous kernels on Polish spaces, the following are equivalent:

- $\mathrm{KCI} = 0$ if and only if $X \perp\!\!\!\perp Y \mid Z$.
- k_X, k_Y are characteristic, and k_Z is integrally strictly positive definite.

- For bounded kernels, $\mathrm{KCI} \leq c \cdot \mathrm{CMI}$.
- If we take k_X, k_Y linear and $k_Z \equiv 1$, we get (square of) the population target of GCM, $\mathbb{E}\,\mathrm{Cov}(X, Y \mid Z)$ (Shah and Peters 2020).
- If we use $k_Z = \omega \otimes \omega$, get weighted GCM (Scheidegger et al. 2022).

Conditional mean embeddings (CMEs)

Have $\Pi_{\mathrm{CI}}(P)(\varphi_X \otimes \varphi_Y \otimes \varphi_Z) = \mathbb{E}_Z\Big[\mu_{X|Z}(Z) \otimes \mu_{Y|Z}(Z) \otimes \varphi_Z(Z)\Big]$,

using the **conditional mean embedding** (Song et al. 2009; Grünewälder et al. 2012; Park and Muandet 2020)

$$\mu_{X|Z}(z) := \mathbb{E}[\varphi_X(X) \mid Z = z] \in \mathcal{H}_X, \qquad \langle \mu_{X|Z}(z), f \rangle = \mathbb{E}[f(X) \mid Z = z].$$

So, subtracting and conditioning on Z,

$$\mathrm{KCI} = \Big\| P_{XYZ}\Big[(\varphi_X(X) - \mu_{X|Z}(Z)) \otimes (\varphi_Y(Y) - \mu_{Y|Z}(Z)) \otimes \varphi_Z(Z)\Big] \Big\|^2.$$

Using $(K_X^c)_{ij} = \langle \varphi_X(x_i) - \mu_{X|Z}(z_i), \varphi_X(x_j) - \mu_{X|Z}(z_j) \rangle$, natural estimator

$$\mathrm{KCI}_n = \frac{1}{n(n-1)} \sum_{i \neq j} (K_Z)_{ij} (K_X^c)_{ij} (K_Y^c)_{ij}.$$

Testing with KCI

- Unfortunately we can't do exact permutation tests: only have one sample per Z.
- Under $H_0, \mathrm{KCI}_n = O_p(1/n)$; test threshold has scale $1/n$.
  - Can estimate threshold with **wild bootstrap** (or use concentration inequalities).
- If $\mathrm{KCI} > 0$, then $\mathrm{KCI}_n \to \mathrm{KCI} > 0$: the test is consistent.
- So, all good… except that **we don't know the CMEs!**

Estimating CMEs

$$\mu_{X|Z}(z) = \mathbb{E}[\varphi_X(X) \mid Z = z]$$

One option: conditional generative model to sample from $X \mid Z = z$; then use Monte Carlo. Called (approximate) "Model-X" (Candès et al. 2018).

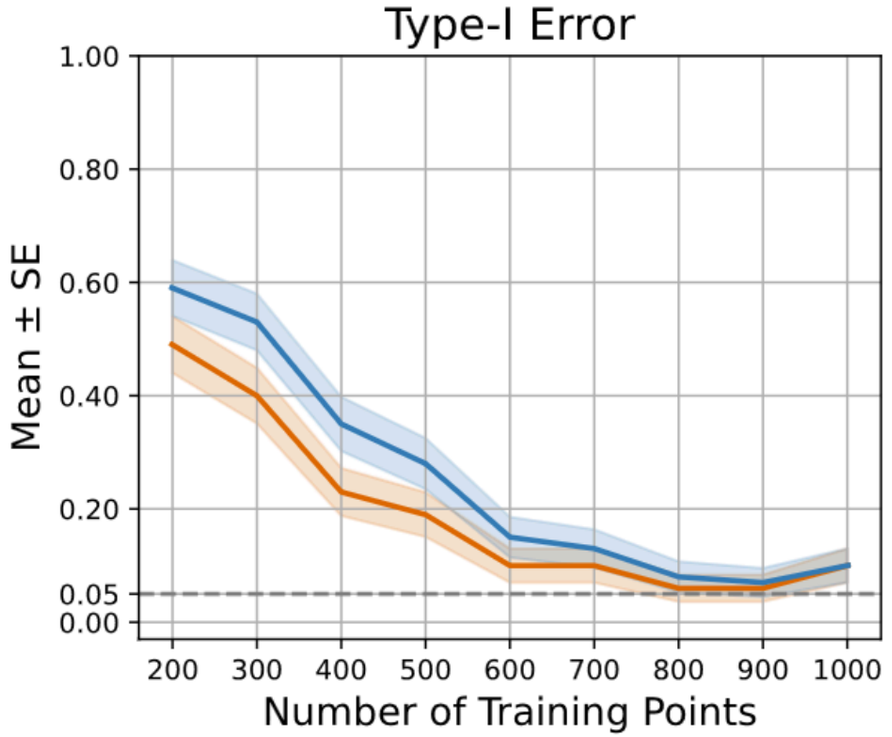
Another: regression from $z \in \mathcal{Z}$ to $\varphi_X(X) \in \mathcal{H}_X$ (Grünewälder et al. 2012).

$$\hat{\mu}_{X|Z}(z) = \sum_{i=1}^{m} \beta_i(z)\, \varphi_X(x_i), \qquad \beta(z) = (K_Z + m\lambda I)^{-1} k_Z(z).$$

- We only need $\langle \hat{\mu}(z), \hat{\mu}(z') \rangle = \beta(z)^\top K_X \beta(z')$: can evaluate everything.
- **Rates** (Li et al. 2022, 2024): In the **best case** (with characteristic k_X) convergence is $m^{-1/4}$; worst case can be arbitrarily slow.

KCI with estimated CMEs

Just plug $\hat{\mu}_{X|Z}$ and $\hat{\mu}_{Y|Z}$ into KCI_n: call this $\widehat{\mathrm{KCI}}_n$. Use wild bootstrap.



Need the right k_Z for good power – trying to learn it makes the problem *way worse*.

Is this task even possible?

No free lunch (Shah and Peters 2020)

- $X, Y, Z \sim \mathbf{Unif}[0, 1]$ independent
- D := 100th decimal digit of Z
- X' := X but first digit D; Y' likewise
- $X' \not\perp\!\!\!\perp Y'$, but $X' \perp\!\!\!\perp Y' \mid Z$

Z	0.47182…9316…
X	0.82915…7140…
X'	0.32915…7140…
Y	0.10483…0204…
Y'	0.30483…0204…
Z'	0.47182…9716…

- Z' := Z with the 100th digit re-randomized; $X' \not\perp\!\!\!\perp Y' \mid Z'$

No free lunch (Shah and Peters 2020)

Theorem (Shah and Peters 2020, Corollary 3) For $M \in (0, \infty]$, let $\mathcal{E}$ be the laws on $(-M, M)^s$ with a Lebesgue density. Let $\mathcal{P} \subset \mathcal{E}$ be those with $X \perp\!\!\!\perp Y \mid Z$, and $\mathcal{Q} = \mathcal{E} \setminus \mathcal{P}$ the rest.

If (ψ_n) has uniformly asymptotic level α over $\mathcal{P}$, then $\limsup_n Q^n \psi_n \leq \alpha$ for **every** $Q \in \mathcal{Q}$.

Okay, but, who cares about the 100th decimal place?

- This is obviously a pathological construction.
- But what *is* it, really? It's a conditional law $P_{X|Z}$ that regression can't learn from finite data. Hard to estimate $\mu_{X|Z}$ means hard to test.
- For KCI this is exactly the failure mode: the **only** unknowns are $\mu_{X|Z}, \mu_{Y|Z}$. With them, validity is one line of Hoeffding; without them, nothing is guaranteed.
- If we assume $P_{X|Z}$ known (Model-X), everything is fine. If it's wrong, Type I error inflates by the conditional total variation error (Berrett et al. 2020).
- So, let's think a little more about what "possible" means.

It’s actually not *so* bad

Three kinds of level

Tests are $\psi_n : \mathcal{V}^n \to [0,1]$ = rejection probability. Over a null class $\mathcal{P}$, (ψ_n) has

- **finite-sample level α:** $\sup_n \sup_{P \in \mathcal{P}} P^n \psi_n \leq \alpha$
- **uniformly asymptotic level α:** $\limsup_n \sup_{P \in \mathcal{P}} P^n \psi_n \leq \alpha$
- **pointwise asymptotic level α:** $\sup_{P \in \mathcal{P}} \limsup_n P^n \psi_n \leq \alpha$

Each implies the next. Shah and Peters (2020) rule out useful tests with the first two.

Practical tests almost always only claim the third.

A not-so-scary problem

Testing $H_0 : \mathbb{E}W = 0$ among laws with finite variance.

- Bahadur and Savage (1956) show that any test with *uniformly* asymptotic level α has $\limsup_n Q^n \psi_n \leq \alpha$ at every alternative.
- The t-test has *pointwise* asymptotic level α and is consistent against every alternative.

- **Maybe CI testing is like this?**
  - Györfi and Walk (2012) claimed yes, but proof was wrong (Neykov et al. 2021).
  - Zhang et al. (2011, Prop. 5) would imply yes, but proof sketch was wrong.
  - Stated as open by Györfi et al. (2023) and Boeken et al. (2026).

CI testing is scarier…

Theorem (Sutherland 2026) Fix $M \in (0, \infty]$ and $L \geq L_0(M, s)$. Let $\mathcal{P}$ be the CI laws with density $\leq L$ on $(-M, M)^s$, and $\mathcal{Q}$ the conditionally dependent ones. If (ψ_n) has pointwise asymptotic level $\alpha < 1$ over $\mathcal{P}$,

$$\inf_{Q \in \mathcal{Q}} \limsup_{n \to \infty} Q^n \psi_n \leq \alpha.$$

The offending null can have a continuous density and a uniformly continuous conditional law; the alternatives can be taken C^∞.

Corollary. If $Q^n \psi_n \to 1$ for every alternative, then some null P_* has $\limsup_n P_*^n \psi_n = 1$.

…but not *too* scary

Proposition (Sutherland 2026) Take any $D \geq 0$ with $D(P) > 0$ iff $X \not\perp\!\!\!\perp Y \mid Z$, and estimators $\hat{D}_n \to D(P)$ in probability for **every** P. A partition estimate will do, or KCI with universally consistent CMEs (Tamás and Csáji 2024).

Then $\psi_n = \alpha + (1-\alpha)\min\{1, \hat{D}_n/\delta\}$ has pointwise asymptotic level α, limiting power $\alpha + (1-\alpha)\min\{1, D(Q)/\delta\} > \alpha$ at **every** alternative, and $Q^n\psi_n \to 1$ whenever $D(Q) \geq \delta$.

- Uniform-level CI tests have power against *no* alternative (Shah and Peters 2020)
- Pointwise-level CI tests can have power 1 against *some* alternatives
- …but not against *all* alternatives, unlike the t-test

But in practice, it’s still pretty hard.

“Eventually does the right thing” is a *really* weak guarantee!

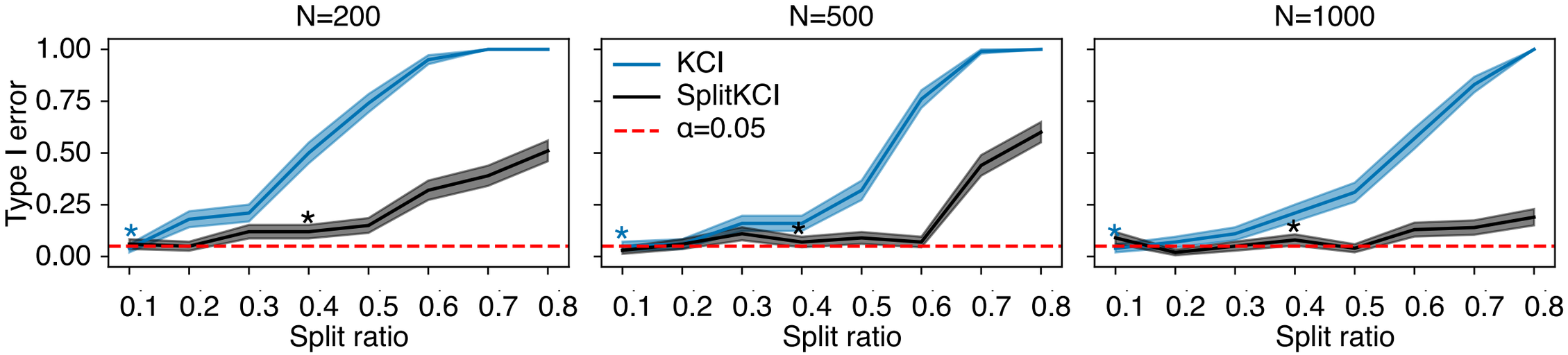
Estimated CMEs break the null calibration

Fit $\hat{\mu}_{X|Z} = \mu_{X|Z} + \Delta_X$ and $\hat{\mu}_{Y|Z} = \mu_{Y|Z} + \Delta_Y$ on independent data, plug in. Under H_0, the population value of the plug-in statistic is

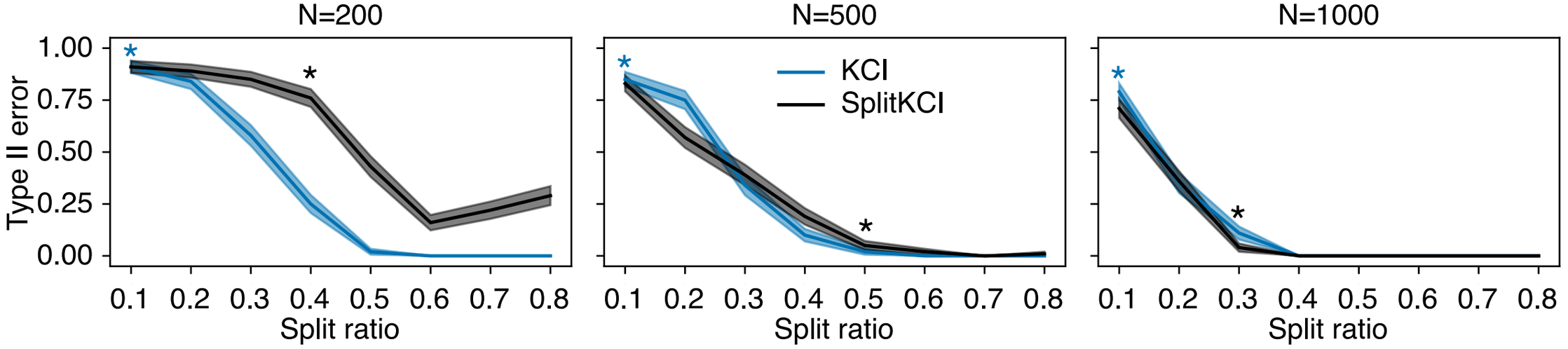
$$\widehat{\mathrm{KCI}} = \mathbb{E}\Big[k_Z(Z, Z')\,\langle \Delta_X(Z), \Delta_X(Z')\rangle_{\mathcal{H}_X}\,\langle \Delta_Y(Z), \Delta_Y(Z')\rangle_{\mathcal{H}_Y}\Big] \neq 0$$

- $\hat{\mu}_{X|Z}$ is smooth in k_Z; hopefully $\mu_{X|Z}$ is too or everything is hopeless.
- So Δ_X, Δ_Y are smooth functions in k_Z… and so $\widehat{\mathrm{KCI}}$ might be big.
- Our test calibration is based on $\mathrm{KCI}_n = O_p(1/n)$.
- But if $\widehat{\mathrm{KCI}} > 0$, then $\widehat{\mathrm{KCI}}_n = \widehat{\mathrm{KCI}} + O_p(1/\sqrt{n})$.
- For fixed regressions, **bigger n makes everything worse.**
- Can’t tell regression error from real dependence!

How bad? Pretty bad!



Synthetic rat hippocampal data. $N \cdot \textbf{split ratio}$ for training CMEs, rest for testing.



Reducing bias with SplitKCI

Fit *two independent* CMEs per variable and use (Pogodin et al. 2024)

$$(K_X^c)_{ij} = \left\langle \varphi_X(x_i) - \hat{\mu}^{(1)}_{X|Z}(z_i), \varphi_X(x_j) - \hat{\mu}^{(2)}_{X|Z}(z_j) \right\rangle.$$

Changes the bias under the null hypothesis to the (smaller)

$$\mathbb{E}\left[k_Z(Z, Z') \, \langle \Delta^{(1)}_X(Z), \Delta^{(2)}_X(Z') \rangle_{\mathcal{H}_X} \, \langle \Delta^{(1)}_Y(Z), \Delta^{(2)}_Y(Z') \rangle_{\mathcal{H}_Y}\right].$$

Can prove rates work out if the regression is **much** more precise than the test.

Heuristic to determine split: smallest one that says $X \perp\!\!\!\perp Z \mid Z, Y \perp\!\!\!\perp Z \mid Z$.

Manages the regression error. Doesn't fix it.

Ongoing work: "leave-two-out" extension that allows using the *same data* for training CMEs and the KCI estimate. Highlights that just throwing away data for KCI is weird…

Fundamentally, what we’re missing is the uncertainty in a Hilbert-valued regression.

Everything is about $\Delta = \hat{\mu} - \mu$

- **Shah and Peters (2020):** an adversary puts $P_{X|Z}$ where regression can't see it.
- **Pointwise:** even for one fixed law, regression error at small scales can be made to decay arbitrarily slowly.
- **Practice:** our null calibration can't account for regression errors. SplitKCI helps, doesn't solve.
- The only kind of validity that is *possible* – finite-sample or uniform level under an explicit smoothness class, or pointwise level with an explicit separation δ – is also exactly what uncertainty quantification for the regression would give us.

We need epistemic uncertainty quantification for Hilbert-valued kernel regression.

Hilbert-valued GPs to the rescue?

- We're starting to try modelling $\mu_{X|Z} \sim \mathcal{GP}$: inputs $z_i \in \mathcal{Z}$, outputs $\varphi_X(x_i) \in \mathcal{H}_X$
- Calibrate $\widehat{\mathrm{KCI}}_n$ under the Δ posterior, instead of assuming $\Delta = 0$
- A priori constraints on e.g. $0 < [\mu_{X|Z}(z)](x) \leq 1$ for all x, z (Holger's talk yesterday)
- Choosing infinite-dimensional priors/likelihoods is hard! Reasonable epistemic uncertainty in $\mathbb{E}[\varphi_X(X) \mid Z = z]$ needs reasonable aleatoric covariance $\Sigma_{X|Z}$
  - λI (a) isn't even a covariance (not trace-class) and (b) is super wrong
  - Definitely varies with z in realistic settings
  - Estimating it is another regression, at least as hard as the first
- Maybe there's a worst-case kernel approximation result that gives a finite-sample-valid CI test, under some explicit smoothness class?

Summary

- $\mathrm{MMD}(P, \text{projection of } P \text{ onto } H_0)$; unconditional independence is easy
- Conditional independence (KCI) needs a Hilbert-valued regression and is hard
  - Can't have universally uniform level (Shah and Peters 2020)
  - Can have pointwise level and consistency against any separated class (new)
  - Ignoring rates, only slightly harder than testing a mean
- In practice, the problem is the regression error
  - SplitKCI helps manage, definitely does not solve
  - Would probably fix it: honest epistemic uncertainty for Hilbert-valued kernel regression

Consider helping us out! 😇 · djsutherland.ml

References

Angwin, Julia, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2017. “Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas with the Same Risk.” *ProPublica*, April. https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk.

Bahadur, R. R., and Leonard J. Savage. 1956. “The Nonexistence of Certain Statistical Procedures in Nonparametric Problems.” *Annals of Mathematical Statistics* 27 (4): 1115–22. https://doi.org/10.1214/aoms/1177728077.

Berlinet, Alain, and Christine Thomas-Agnan. 2004. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer. https://link.springer.com/book/10.1007/978-1-4419-9096-9.

Berrett, Thomas B., Yi Wang, Rina Foygel Barber, and Richard J. Samworth. 2020. “The Conditional Permutation Test for Independence While Controlling for Confounders.” *Journal of the Royal Statistical Society Series B* 82 (1): 175–97. https://doi.org/10.1111/rssb.12340.

Boeken, Philip, Eduardo Skapinakis, Konstantin Genin, and Joris M. Mooij. 2026. *Topological Criteria for Hypothesis Testing with Finite-Precision Measurements*. arXiv. https://arxiv.org/abs/2601.13946.

Candès, Emmanuel, Yingying Fan, Lucas Janson, and Jinchi Lv. 2018. “Panning for Gold: ‘Model-X’ Knockoffs for High Dimensional Controlled Variable Selection.” *Journal of the Royal Statistical Society Series B* 80 (3): 551–77. https://doi.org/10.1111/rssb.12265.

Dawid, A. Philip, and Steffen L. Lauritzen. 1993. “Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models.” *Annals of Statistics* 21 (3): 1272–317. https://doi.org/10.1214/aos/1176349260.

Gao, Ming, Yuhao Wang, and Bryon Aragam. 2026. “Optimal Structure Learning and Conditional Independence Testing.” *International Conference on Machine Learning (ICML)*. https://openreview.net/forum?id=uurjpxrLc5.

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. “A Kernel Two-Sample Test.” *Journal of Machine Learning Research* 13 (25): 723–73. https://jmlr.org/papers/v13/gretton12a.html.

Gretton, Arthur, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. “Measuring Statistical Dependence with Hilbert-Schmidt Norms.” *Algorithmic Learning Theory (ALT)*, 63–77. https://doi.org/10.1007/11564089_7.

Grünewälder, Steffen, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. 2012. “Conditional Mean Embeddings as Regressors.” *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/1205.4656.

Györfi, László, Tamás Linder, and Harro Walk. 2023. “Lossless Transformations and Excess Risk Bounds in Statistical Inference.” *Entropy* 25 (10): 1394. https://doi.org/10.3390/e25101394.

Györfi, László, and Harro Walk. 2012. “Strongly Consistent Nonparametric Tests of Conditional Independence.” *Statistics & Probability Letters* 82 (6): 1145–50. https://doi.org/10.1016/j.spl.2012.02.023.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. “Equality of Opportunity in Supervised Learning.” *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1610.02413.

He, Zheng, Roman Pogodin, Yazhe Li, Namrata Deka, Arthur Gretton, and Danica J. Sutherland. 2025. “On the Hardness of Conditional Independence Testing in Practice.” *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2512.14000.

He, Zheng, Roman Pogodin, Antonin Schrab, et al. 2026. “On Measuring and Testing Conditional Independence.” Unpublished manuscript.

Hemerik, Jesse, and Jelle Goeman. 2018. “Exact Testing with Random Permutations.” *TEST* 27 (4): 811–25. https://doi.org/10.1007/s11749-017-0571-1.

Jiang, Yibo, and Victor Veitch. 2022. “Invariant and Transportable Representations for Anti-Causal Domain Shifts.” *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*. https://arxiv.org/abs/2207.01603.

Kanagawa, Motonobu, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. 2018. *Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences*. arXiv. https://arxiv.org/abs/1807.02582.

Li, Zhu, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. 2022. “Optimal Rates for Regularized Conditional Mean Embedding Learning.” *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2208.01711.

Li, Zhu, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. 2024. “Towards Optimal Sobolev Norm Rates for the Vector-Valued Regularized Least-Squares Algorithm.” *Journal of Machine Learning Research* 25 (181): 1–51. https://jmlr.org/papers/v25/23-1663.html.

Liu, Feng, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. 2020. “Learning Deep Kernels for Non-Parametric Two-Sample Tests.” *International Conference on Machine Learning (ICML)*. http://proceedings.mlr.press/v119/liu20m.html.

Massa, M. Sofia, and Steffen L. Lauritzen. 2010. “Combining Statistical Models.” In *Algebraic Methods in Statistics and Probability II*, vol. 516. Contemporary Mathematics. American Mathematical Society. https://doi.org/10.1090/conm/516/10179.

Neykov, Matey, Sivaraman Balakrishnan, and Larry Wasserman. 2021. “Minimax Optimal Conditional Independence Testing.” *Annals of Statistics* 49 (4): 2151–77. https://doi.org/10.1214/20-AOS2030.

Park, Junhyung, and Krikamol Muandet. 2020. “A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings.” *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2002.03689.

Pogodin, Roman, Antonin Schrab, Yazhe Li, Danica J. Sutherland, and Arthur Gretton. 2024. *Practical Kernel Tests of Conditional Independence*. arXiv. https://arxiv.org/abs/2402.13196.

Ritter, Klaus. 2000. *Average-Case Analysis of Numerical Problems*. Vol. 1733. Lecture Notes in Mathematics. Springer. https://doi.org/10.1007/BFb0103934.

Scheidegger, Cyrill, Julia Hörrmann, and Peter Bühlmann. 2022. "The Weighted Generalised Covariance Measure." *Journal of Machine Learning Research* 23 (273): 1–68. https://jmlr.org/papers/v23/21-1328.html.

Shah, Rajen D., and Jonas Peters. 2020. "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure." *Annals of Statistics* 48 (3): 1514–38. https://doi.org/10.1214/19-AOS1857.

Song, Le, Jonathan Huang, Alex Smola, and Kenji Fukumizu. 2009. "Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems." *International Conference on Machine Learning (ICML)*, 961–68. https://doi.org/10.1145/1553374.1553497.

Sriperumbudur, Bharath K., Kenji Fukumizu, and Gert R. G. Lanckriet. 2011. "Universality, Characteristic Kernels and RKHS Embedding of Measures." *Journal of Machine Learning Research* 12 (70): 2389–410. https://jmlr.org/papers/v12/sriperumbudur11a.html.

Sutherland, Danica J. 2026. "Conditional Independence Is Not Pointwise Testable." Unpublished manuscript.

Sutherland, Danica J., Hsiao-Yu Tung, Heiko Strathmann, et al. 2017. "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy." *International Conference on Learning Representations (ICLR)*. http://openreview.net/forum?id=HJWHIKqgl.

Szabó, Zoltán, and Bharath K. Sriperumbudur. 2018. "Characteristic and Universal Tensor Product Kernels." *Journal of Machine Learning Research* 18 (233): 1–29. https://jmlr.org/papers/v18/17-492.html.

Tamás, Ambrus, and Balázs Csanád Csáji. 2024. "Recursive Estimation of Conditional Kernel Mean Embeddings." *Journal of Machine Learning Research* 25 (264): 1–35. https://jmlr.org/papers/v25/23-0168.html.

Wei, Aaron, Milad Jalali Asadabadi, and Danica J. Sutherland. 2026. "Maximum Mean Discrepancy with Unequal Sample Sizes via Generalized u-Statistics." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=KjXW75GHHF.

Xu, Nathaniel, Feng Liu, and Danica J. Sutherland. 2026. "Learning Representations for Independence Testing." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=pDvKoXRsnW.

Zhang, Kun, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2011. "Kernel-Based Conditional Independence Test and Application in Causal Discovery." *Uncertainty in Artificial Intelligence (UAI)*, 804–13. https://arxiv.org/abs/1202.3775.

Ziegel, Johanna, David Ginsbourger, and Lutz Dümbgen. 2024. "Characteristic Kernels on Hilbert Spaces, Banach Spaces, and on Sets of Measures." *Bernoulli* 30 (2): 1441–57. https://doi.org/10.3150/23-BEJ1639.