On the Hardness of Conditional Independence Testing In Practice

Zheng He[1], Roman Pogodin[2,3], Yazhe Li[4], Namrata Deka[5], Arthur Gretton[6,7], Danica J. Sutherland[1,8]

[1]UBC, [2]McGill, [3]Mila, [4]Microsoft AI, [5]CMU, [6]Gatsby Unit, [7]UCL, [8]Amii

Abstract

- **Motivation:** Conditional independence tests (KCI, GCM) are widely used in causal discovery, scientific modeling, fairness, OOD generalization, …, but are unstable and behave poorly. Why?
- **Key Insight:** Regression errors in estimating conditionals cause spurious dependence under the null, violating asymptotic assumptions and leading to null calibration failure. Choosing the conditioning variable kernel helps power but makes bias worse.

CI Testing is Impossible

Shah and Peters (2020): for any CI test and any conditionally dependent continuous distribution, there is an indistinguishable distribution which is conditionally independent. (“Hide” dependence in the conditional.)
Us: solely because of estimating conditional feature means from data.

Characterizing CI

Theorem (extending Daudin, 1980). $A \perp\!\!\!\perp B \mid C$ if and only if

$$\mathop{\mathbb{E}}_{C}\left[w(C) \mathop{\mathbb{E}}_{AB|C}\left[\left(f(A) - \mathbb{E}[f(A) \mid C]\right)\left(g(B) - \mathbb{E}[g(B) \mid C]\right) \mid C\right]\right] = 0,$$

for all square-integrable functions $f \in L^2_A$, $g \in L^2_B$, and $w \in L^2_C$.

Kernel-based CI Measure

An RKHS $\mathcal{H}_\mathcal{A}$ contains functions that are linear w.r.t. some feature map ϕ_A: $f(a) = \langle w, \phi_A(a)\rangle$.

The **conditional mean embedding** $\mu_{A|C}(c) = \mathbb{E}[\phi_A(A) \mid C = c]$ gives $\langle \mu_{A|C}(c), f\rangle_{\mathcal{H}_\mathcal{A}} = \mathbb{E}[f(A) \mid C = c]$.

Can make a **conditional cross-covariance operator** $\mathfrak{C}_{AB|C}$ that captures the dependence between A and B given C:

$$\mathfrak{C}_{AB|C}(c) \coloneqq \mathop{\mathbb{E}}_{AB|C}\left[\left(\phi_A(A) - \mu_{A|C}(c)\right) \otimes \left(\phi_B(B) - \mu_{B|C}(c)\right) \mid C = c\right]$$

gives $\langle f \otimes g, \mathfrak{C}_{AB|C}(c)\rangle = \operatorname{Cov}(f(A), g(B) \mid C = c)$.

The **KCI operator** summarizes over C:

$$\mathfrak{C}_{\mathrm{KCI}} \coloneqq \mathop{\mathbb{E}}_{C}\left[\mathfrak{C}_{AB|C}(C) \otimes \phi_C(C)\right]$$

which gives for any test functions f, g, w, $\langle f \otimes g, \mathfrak{C}_{\mathrm{KCI}}\, w\rangle_{\mathrm{HS}(\mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{A})} =$

$$\mathop{\mathbb{E}}_{C}\left[w(C) \mathop{\mathbb{E}}_{AB|C}\left[\left(f(A) - \mathbb{E}[f(A) \mid C]\right)\left(g(B) - \mathbb{E}[g(B) \mid C]\right)\right]\right].$$

So if $\mathcal{H}_\mathcal{A}, \mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{C}$ are L^2-universal, $\mathfrak{C}_{\mathrm{KCI}} = \mathbf{0}$ iff $A \perp\!\!\!\perp B \mid C$.
Common test statistic (Zhang et al. 2011) is based on $\mathrm{KCI} = \|\mathfrak{C}_{\mathrm{KCI}}\|^2_{\mathrm{HS}}$.

Unified Framework of KCI and GCM

GCM (Shah and Peters 2020) computes a studentized estimate of

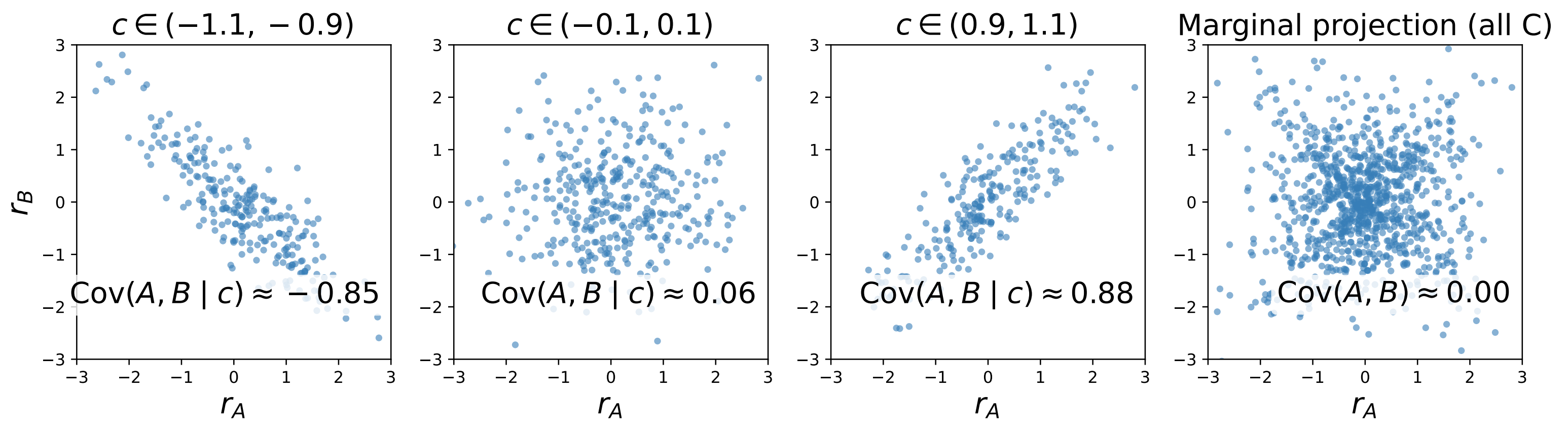
$$\mathbb{E}[w(C)(A - \mathbb{E}[A|C])(B - \mathbb{E}[B|C])],$$

with $w(C) = 1$. Weighted GCM (Scheidegger et al. 2022) chooses w to emphasize specific conditional structures.
KCI operator with linear kernels on A,B ($\phi_A(a) = a$) reduces to the same conditional covariance operator estimated by (w)GCM.

Kernel Choice Tradeoff

$$A = f_A(C) + \tau r_A, \quad B = f_B(C) + \tau r_B, \quad (r_A, r_B) \mid C \sim N(0, \Sigma(C)),$$

with $\operatorname{Var}(r_A) = 1 = \operatorname{Var}(r_B)$, $\operatorname{Cov}(r_A, r_B) = \gamma(C) = \sin(\beta C)$



Take linear kernels for A and B, and a lengthscale-ℓ_C Gaussian kernel $k_C(C, C') = \exp\left(-(C - C')^2/(2\ell_C^2)\right)$ on C:

$$\mathrm{KCI} = \tau^4 \mathop{\mathbb{E}}_{C,C'}\left[k_C(C, C')\gamma(C)\gamma(C')\right]$$

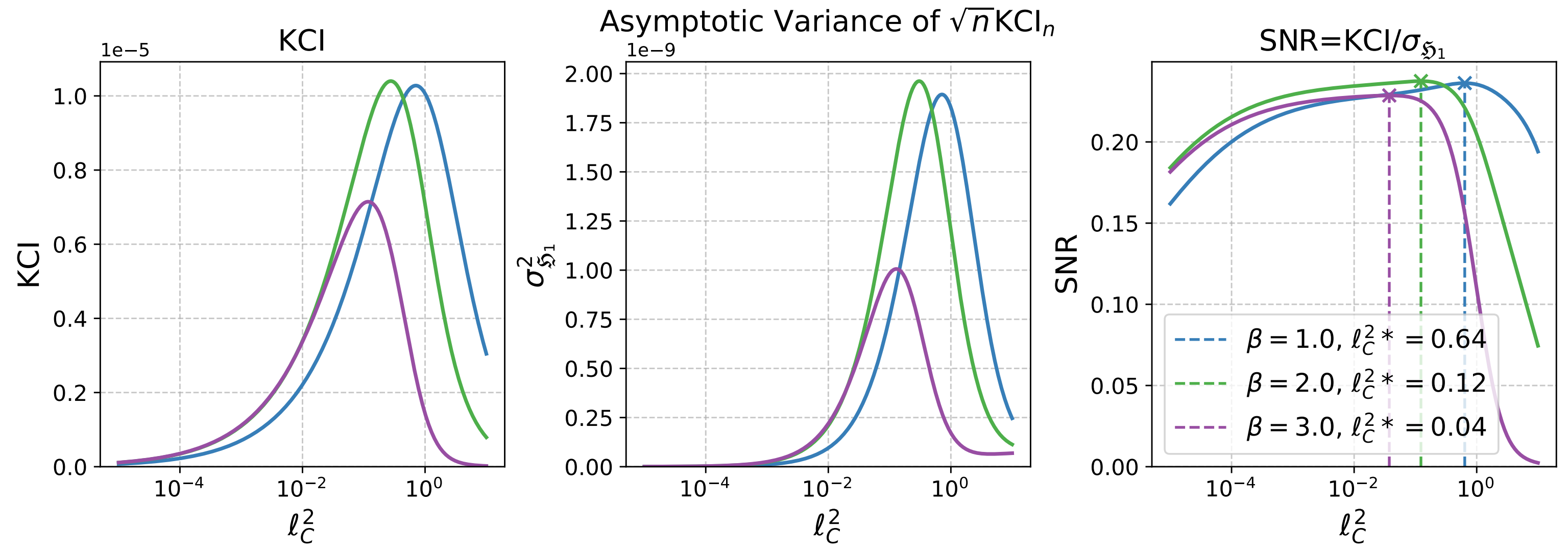
The kernel on C must resolve regions where dependence is strong:

- Too smooth: conditional structure blurred → statistic near 0.
- Too sharp: estimates become very noisy → unstable statistic.

GCM corresponds to $\ell_C = \infty$.
Conditioning kernel selection. Under the alternative, there is a scalar $\hat{\sigma}^2_{\mathfrak{H}_1} \geq 0$ so that as $n \to \infty$,

$$\sqrt{n}(\widehat{\mathrm{KCI}}_n - \widehat{\mathrm{KCI}}) \xrightarrow{d} \mathcal{N}(0, \hat{\sigma}^2_{\mathfrak{H}_1}).$$

Can roughly maximize test power by maximizing $\widehat{\mathrm{SNR}} = \widehat{\mathrm{KCI}}/\hat{\sigma}_{\mathfrak{H}_1}$.

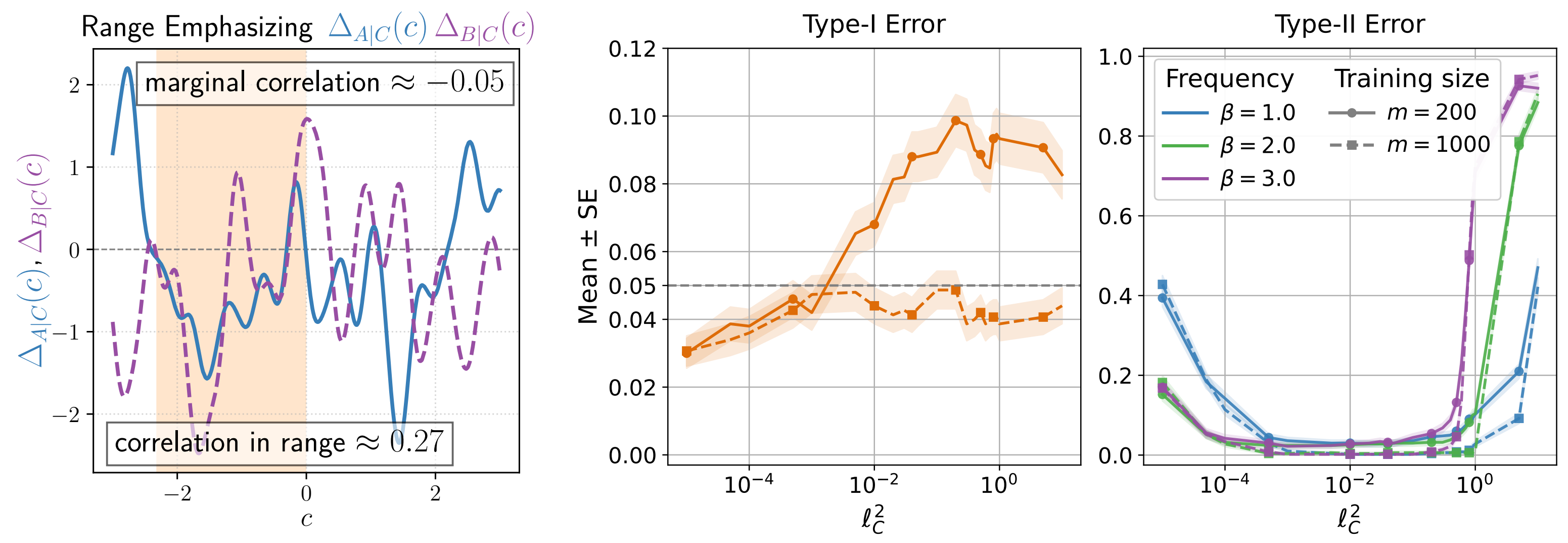


Type-I Error Inflation with Regression Errors

What if $A \perp\!\!\!\perp B \mid C$ but we have $\mathbb{E}\,\phi_A(A) \mid C$ wrong?
Let $\Delta_{A|C}(C) = \hat{\mu}_{A|C}(A) - \mu_{A|C}(C)$, same for $\Delta_{B|C}$.
When $A \perp\!\!\!\perp B \mid C$, the expected KCI with regression errors is

$$\widehat{\mathrm{KCI}} = \mathbb{E}\left[k_C(C, C') \langle \Delta_{A|C}(C), \Delta_{A|C}(C')\rangle_{\mathcal{H}_\mathcal{A}} \langle \Delta_{B|C}(C), \Delta_{B|C}(C')\rangle_{\mathcal{H}_\mathcal{B}}\right].$$

If we’re choosing a C kernel to look for dependence, we can usually find it in fixed, smooth $\Delta_{A|C}$, $\Delta_{B|C}$!



Why null calibration breaks: If $A \perp\!\!\!\perp B \mid C$, regressions fixed:

	KCI mean	KCI std. dev
Perfect regression	0	$\Theta(1/n)$
Incorrect regression	$\Theta(1)$	$\Theta(1/\sqrt{n})$

But the null approximations (χ^2 or Gamma from Zhang et al. 2012, wild bootstrap from Pogodin et al. 2024) *don’t know* what’s bias and what’s signal! Need regression’s seeming-dependence → 0 fast for calibration.

Practical Recommendations

- **Sample Splitting**: Use an independent training set (for conditional mean estimation) and a test set (for KCI statistic). Training size should be $\geq$, preferably $\gg$, test size depending on complexity.
- **Strong Regression:** Use flexible, low-bias models for regression.
- **Power Maximization**: Select k_C via SNR maximization on the training set.
- **Be really really careful**. Can still easily trick yourself.

Future directions

Will need to explicitly incorporate regression uncertainty in the null.
SplitKCI (Pogodin et al. 2024) is a step in this direction, but not enough!
Impossible in general…but so is regression, and we still do regression.

References

Shah and Peters. The hardness of conditional independence testing and the generalised covariance measure. Ann. Stat. 2020.
Daudin. Partial association measures and an application to qualitative regression. Biometrika 1980.
Zhang et al. Kernel-based Conditional Ind. Test and Application in Causal Discovery. UAI 2011.
Scheidegger, Hörrmann, and Bühlmann. The weighted generalised covariance measure. JMLR 2022.
Pogodin et al. Practical Kernel Tests of Conditional Independence. arXiv 2024.