On the Hardness of Conditional Independence Testing In Practice

Zheng He
UBC
zhhe@cs.ubc.ca

Roman Pogodin
McGill and Mila
rmn.pogodin@gmail.com

Yazhe Li
Microsoft AI*
yazheli@outlook.com

Namrata Deka
CMU
ndeka@andrew.cmu.edu

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

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

Abstract

Tests of conditional independence (CI) underpin a number of important problems in machine learning and statistics, from causal discovery to evaluation of predictor fairness and out-of-distribution robustness. Shah and Peters (2020) showed that, contrary to the unconditional case, no universally finite-sample valid test can ever achieve nontrivial power. While informative, this result (based on "hiding" dependence) does not seem to explain the frequent practical failures observed with popular CI tests. We investigate the Kernel-based Conditional Independence (KCI) test – of which we show the Generalized Covariance Measure underlying many recent tests is *nearly* a special case – and identify the major factors underlying its practical behavior. We highlight the key role of errors in the conditional mean embedding estimate for the Type I error, while pointing out the importance of selecting an appropriate conditioning kernel (not recognized in previous work) as being necessary for good test power but also tending to inflate Type I error.

1 Introduction

Conditional independence (CI) testing is a fundamental task, required for almost any scientific hypothesis that "controls for" confounders; it is moreover a core subroutine in the standard PC algorithm for causal discovery and its many variants (Spirtes et al., 1993). Further recent major machine learning-specific applications include checking or enforcing the fairness of a predictor or representation with equalized odds (Hardt et al., 2016), and relatedly for a predictor's domain invariance, particularly in "anticausal" settings (e.g. Z. Wang and Veitch, 2022).

When the conditioning variable takes on a small number of discrete values, the problem is simple to reduce to that of unconditional independence testing, for which there are many good methods: for instance, many based on the Hilbert-Schmidt Independence Criterion (Gretton et al., 2005; Gretton et al., 2008). When the conditioning variable is continuous, however, the situation is much more challenging: when testing whether $A \perp\!\!\!\perp B \mid C$ based on samples for a continuously distributed C,[1] we will only observe one (A, B) pair for each value of C, and so we must make some form of assumption on the smoothness of the conditional distribution $(A, B) \mid C = c$ as a function of c. Shah and Peters (2020) proved that doing so in total generality is impossible. Their lower bound, however, is an adversarial construction of a particular distribution (discussed in Section 4) which does not seem especially informative as to the widespread failures of CI tests in practical settings. Since the importance of the task means that, despite its impossibility in general, we still want to pursue CI testing, we must consider particular types of tests used in practice and when, and why, they fail.

*Work done at Gatsby Unit, UCL.

[1]We will always use $A \perp\!\!\!\perp B \mid C$, since papers in this area use both $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

There are a few major categories of techniques. One is the Kernel-based Conditional Independence (KCI) technique introduced by K. Zhang et al. (2011). As a kernel method, this technique is applicable to data of any (potentially complex and structured) form. It has a reputation, however, of doing a poor job at controlling Type I error: that is, it falsely identifies conditional dependence too often (Shah and Peters, 2020; Pogodin et al., 2024). Recent extensions include CIRCE (Pogodin et al., 2023), which is useful as a regularizer for learning A but generally yields a much worse test, and SplitKCI (Pogodin et al., 2024), which helps reduce Type I error rates, but is far from solving the issue. KCI additionally requires a choice of as many as five kernels in order to operate; Pogodin et al. (2023) proposed a method to select two of those five, but it remains unclear how to set the other three.

A number of studies propose to test conditional independence by checking the covariance of residuals from regressions of A and B on C (e.g., H. Zhang et al., 2017; H. Zhang et al., 2018; Shah and Peters, 2020). We refer to this class of methods collectively as the Generalized Covariance Measure (GCM), following Shah and Peters (2020). While conceptually simple, GCM captures only linear covariance between residuals and averages the dependence over C, rather than evaluating the covariance conditional on specific values of C. Weighted GCM (Scheidegger et al., 2022) generalizes the GCM by applying weights based on C, allowing detection of a broader range of conditional dependencies. As we show in Section 3, the standard GCM corresponds to a special case of KCI with simple kernel choices, while Weighted GCM can be viewed as a more flexible, though still constrained, setting of the C kernel.

Having introduced measures of conditional independence, we revisit some theoretical work on the CI testing hardness in Section 4, where in particular we show that challenges in CI testing with kernel statistics arise specifically due to challenges in estimating the *conditional mean embedding*, a kernel embedding of the conditional distribution that underpins the majority of such tests (Song et al., 2009; Grünewälder et al., 2012; Klebanov et al., 2020; Park and Muandet, 2020; Li et al., 2024). In Section 5, we provide a clear demonstration that choosing an appropriate C kernel is vital to a sensitive KCI test – in contrast to an implicit claim by K. Zhang et al. (2011) and the approach taken by Pogodin et al. (2023) and Pogodin et al. (2024). Following related work in other settings (e.g. Jitkrittum et al., 2016; Liu et al., 2020; Xu et al., 2024), we suggest a method to select a C kernel which does help achieve more powerful tests. We observe, however, that this method can also make the problem of false rejection even more severe.

In Section 6, we investigate the problem of false rejections in KCI tests. We first analyze simple yet informative special cases, which allows analytical investigation of how regression errors in estimating conditional mean embeddings induce bias in the test statistic's moments. These insights motivate a more general theoretical analysis, where we derive formal bounds linking conditional mean estimation error to test validity. Together, the results clarify the root cause of false rejections and delineate the conditions under which KCI and GCM tests remain reliable.

2 Measuring Conditional Dependence

We first show how to measure conditional dependence with kernels. While the fundamental idea is due to K. Zhang et al. (2011), our framing is somewhat different in terms of the novel Theorem 2.2.

Conditional independence. We build on the characterization of Daudin (1980). To begin, we formalize the intuition that given C, A and B contain no additional information about one another:

Definition 2.1 (Daudin, 1980)**.** Random variables A and B are conditionally independent given C, denoted $A \perp\!\!\!\perp B \mid C$, if for all square-integrable functions $f \in L^2_{AC}$ and $g \in L^2_{BC}$,

$$\mathbb{E}[\, f(A,C)\; g(B,C) \mid C\,] = \mathbb{E}[\, f(A,C)|C\,]\;\; \mathbb{E}[\, g(B,C) \mid C\,] \quad \text{almost surely in } C.$$

This definition is equivalent to stating that the conditional joint distribution factorizes almost surely in C, $P_{A,B|C} = P_{A|C}\, P_{B|C}$, by considering functions f and g as indicators of events.

Building on this definition, we can derive the following equivalence for conditional independence:

Theorem 2.2. *Random variables A and B are conditionally independent given C if and only if*

$$\mathop{\mathbb{E}}_{C}\Big[\, w(C) \mathop{\mathbb{E}}_{AB|C}\big[\,(f(A) - \mathbb{E}[f(A) \mid C])\,(g(B) - \mathbb{E}[g(B) \mid C])\;\mid C\big]\,\Big] = 0, \tag{1}$$

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

This result, proved in Appendix A, extends the characterization of Daudin (1980) to a particularly interpretable form: does any residual dependence between A and B remains after accounting for C? The weighting function $w(C)$ allows emphasizing specific regions of the support of C. Under $A \perp\!\!\!\perp B \mid C$, the conditional covariances vanish C-almost surely; otherwise, there is some nonzero conditional covariance on a C-non-negligible region, which an appropriate $w(C)$ can capture.

Kernel spaces. Since it is infeasible to check all square-integrable functions for f, g, and w, we instead focus on a restricted yet sufficiently rich class of "smooth" functions. Specifically, we consider functions that lie in reproducing kernel Hilbert spaces (RKHSs), which enable characterization of conditional dependence via kernel mappings.

A reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\mathcal{A}$ is a particular space of functions $\mathcal{A} \to \mathbb{R}$; each RKHS is uniquely associated to a positive-definite kernel $k_A : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$. This kernel can itself be written as $k_A(a, a') = \langle \phi_A(a), \phi_A(a') \rangle_{\mathcal{H}_\mathcal{A}}$, where $\phi_A : \mathcal{A} \to \mathcal{H}_\mathcal{A}$ is known as a feature map. The defining *reproducing property* of an RKHS is that for all $f \in \mathcal{H}_\mathcal{A}$ and $a \in \mathcal{A}$, $f(a) = \langle f, \phi_A(a) \rangle_{\mathcal{H}_\mathcal{A}}$. We always assume that any RKHS we deal with is separable; this is guaranteed when k is continuous and the underlying space $\mathcal{A}$ is separable (Steinwart and Christmann, 2008, Lemma 4.33).

KCI operator. The following operator, introduced (in a different form) by K. Zhang et al. (2011), will help us characterize conditional dependence; we reframe it, following Theorem 2.2, to explicitly incorporate a conditional covariance structure.[2] We build this up in pieces.

First, the conditional mean embeddings $\mu_{A|C}(c) \coloneqq \mathbb{E}[\phi_A(A) \mid C = c] \in \mathcal{H}_\mathcal{A}$ and $\mu_{B|C}(c) \coloneqq \mathbb{E}[\phi_B(B) \mid C = c] \in \mathcal{H}_\mathcal{B}$ give RKHS representations of the conditional distributions of A and B given $C = c$: specifically, they satisfy the reproducing property $\langle \mu_{A|C}(c), f \rangle_{\mathcal{H}_\mathcal{A}} = \mathbb{E}[f(A) \mid C = c]$.

The conditional cross-covariance operator, $\mathfrak{C}_{AB|C}$, will capture the dependence structure between A and B with $\langle f, \mathfrak{C}_{AB|C}(c) g \rangle = \mathbb{E}_{AB|C}\left[\left(f(A) - \mathbb{E}[f(A) \mid C]\right) \left(g(B) - \mathbb{E}[g(B) \mid C]\right) \mid C = c \right]$:

$$\mathfrak{C}_{AB|C}(c) \coloneqq \mathop{\mathbb{E}}_{AB|C}\Big[\big(\phi_A(A) - \mu_{A|C}(c)\big) \otimes \big(\phi_B(B) - \mu_{B|C}(c)\big) \mid C = c \Big] \in \mathrm{HS}(\mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{A}). \quad (2)$$

Here $\mathrm{HS}(\mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{A})$ denotes the space of Hilbert–Schmidt operators from $\mathcal{H}_\mathcal{B}$ to $\mathcal{H}_\mathcal{A}$, and the outer product $\phi_A(a) \otimes \phi_B(b) \in \mathrm{HS}(\mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{A})$ is defined by $\big(\phi_A(a) \otimes \phi_B(b)\big) g = \langle \phi_B(b), g \rangle_{\mathcal{H}_\mathcal{B}} \phi_A(a)$ for any $g \in \mathcal{H}_\mathcal{B}$, analogous to the outer product of vectors in finite-dimensional spaces.

The KCI operator aggregates these conditional covariances with information about the context C:

$$\mathfrak{C}_{\mathrm{KCI}} \coloneqq \mathop{\mathbb{E}}_{C}\Big[\mathfrak{C}_{AB|C}(C) \otimes \phi_C(C) \Big] \in \mathrm{HS}(\mathcal{H}_\mathcal{C}, \mathrm{HS}(\mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{A})). \quad (3)$$

For any test functions $f \in \mathcal{H}_\mathcal{A}$, $g \in \mathcal{H}_\mathcal{B}$ and $w \in \mathcal{H}_\mathcal{C}$, the properties above give that

$$\langle f \otimes g, \mathfrak{C}_{\mathrm{KCI}}\, w \rangle_{\mathrm{HS}(\mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{A})} = \mathop{\mathbb{E}}_{C}\Big[w(C) \mathop{\mathbb{E}}_{AB|C}\left[\left(f(A) - \mathbb{E}[f(A) \mid C]\right) \left(g(B) - \mathbb{E}[g(B) \mid C]\right) \right] \Big].$$

If the KCI operator is itself zero, then the quantity above is zero for any choice of $f \in \mathcal{H}_\mathcal{A}$, $g \in \mathcal{H}_\mathcal{B}$, $w \in \mathcal{H}_\mathcal{C}$. If the KCI operator is nonzero, then there exist f, g, w for which it is nonzero, implying that $A \not\perp\!\!\!\perp B \mid C$. A natural measure of conditional dependence is then the magnitude of $\mathfrak{C}_{\mathrm{KCI}}$, as measured by its squared Hilbert-Schmidt norm:

$$\mathrm{KCI} \coloneqq \|\mathfrak{C}_{\mathrm{KCI}}\|^2_{\mathrm{HS}} = \mathop{\mathbb{E}}_{C, C'}\Big[k_C(C, C') \left\langle \mathfrak{C}_{AB|C}(C), \mathfrak{C}_{AB|C}(C') \right\rangle_{\mathrm{HS}(\mathcal{H}_\mathcal{B}, \mathcal{H}_\mathcal{A})} \Big]. \quad (4)$$

The Hilbert–Schmidt norm of an operator is zero if and only if the operator itself is the zero operator. If the RKHSs $\mathcal{H}_\mathcal{A}$, $\mathcal{H}_\mathcal{B}$ and $\mathcal{H}_\mathcal{C}$ are L^2-universal, meaning that they are dense in L^2, then $\mathrm{KCI} = 0$ if and only if $A \perp\!\!\!\perp B \mid C$. Many commonly used kernels, such as the Gaussian RBF kernel $k_A(a, a') = \exp\left(-\|a - a'\|^2 / (2\ell^2)\right)$, are L^2-universal (c.f. Sriperumbudur et al., 2011; Szabó and Sriperumbudur, 2018). Large values of KCI indicate strong evidence of conditional dependence, while values near zero suggest that any apparent dependence can be adequately explained by C.

[2] To obtain this formulation from theirs: first, following Pogodin et al. (2023), remove the C to C regression of the original version (also see Mastouri et al., 2021, Appendix B.9). Second, use a product kernel on (B, C); we are not aware of any uses that do *not* do this, and our framing of Theorem 2.2 makes the final product clearer.

3 Connecting KCI and GCM

Shah and Peters (2020) proposed a Generalized Covariance Measure, which has been the basis of many recent CI tests. For scalar A and B, GCM uses a studentized estimate of the average covariance between residuals, based on any regression method from C to A. Scheidegger et al. (2022) extend the approach to Weighted GCM, which adds a weighting function w; assuming perfect regressions, the population quantity becomes

$$\mathbb{E}\big[\, w(C)(A - \mathbb{E}[\, A \mid C\,])\,(B - \mathbb{E}[\, B \mid C\,])\,\big] \tag{5}$$

With $w(c) = 1$, this is the quantity estimated by GCM; an appropriate choice of w function increases the sensitivity to more types of dependence.

Consider KCI with scalar linear kernels $\phi_A(a) = a$ and $\phi_B(b) = b$. This makes the conditional mean embeddings $\mu_{A|C}(c) = \mathbb{E}[\phi_A(A) \mid C = c] = \mathbb{E}[A \mid C = c]$, and similarly $\mu_{B|C}(c) = \mathbb{E}[B \mid C = c]$. If we further pick the kernel $k_C(c, c') = w(c)w(c')$ so $\phi_C(c) = w(c)$, then (3) becomes identical to (5). The difference is that GCM estimates the value of that expectation (normalized by the standard deviation of the estimates), while the KCI operator estimates the absolute value. This relationship is analogous to that between classifier two-sample tests and maximum mean discrepancy-based tests (Liu et al., 2020, Section 4), and to that between variational mutual information-based independence tests and HSIC tests (Xu et al., 2024).

Consider instead $\mathcal{A} = \mathbb{R}^{d_A}$, $\mathcal{B} = \mathbb{R}^{d_B}$, with multivariate linear $k^{\mathrm{c}}_A(a, a')$, $k^{\mathrm{c}}_B(b, b')$ and the same $\phi_C = w$. The conditional cross-covariance (2) becomes the conditional cross-covariance matrix of shape $d_A \times d_B$, and the KCI operator (3) is the w-weighted average of that matrix. The multivariate (weighted) GCM again takes a studentized estimate of that matrix, and uses the maximum absolute value as its entry. KCI would instead use the Frobenius norm.

In this way, we can see that (weighted) GCM is almost a special case of KCI using simple kernels, further motivating our study of KCI in particular (especially with linear $k^{\mathrm{c}}_A(a, a')$, $k^{\mathrm{c}}_B(b, b')$). The advantage of the weighted over the unweighted statistic also foreshadows the importance of $k_C(c, c')$.

4 Revisiting the Theoretical Hardness of CI Testing

In null hypothesis significance testing, we wish to find a test procedure which rejects the null, i.e. claims that $A \not\perp\!\!\!\perp B \mid C$, with no more than α probability (say 0.05) when in fact the null hypothesis that $A \perp\!\!\!\perp B \mid C$ holds. Such rejections, also known as false positives, are called *Type I errors*. A test has (finite-sample) valid level if its Type I error rate is at most α, while it has (pointwise) asymptotically valid level if for any null distribution, the Type I error rate is asymptotically no more than α. Failing to reject the null when it does not hold is called a *Type II error*; the *power* of a test is the rate at which it does reject, i.e. one minus the Type II error rate for that distribution. Among valid tests, the best one is the one with the highest power. A test is *consistent against fixed alternatives* if for any distribution where the null does not hold, the power approaches 1 as $n \to \infty$.

Impossibility result. Shah and Peters (2020) showed that if a CI test has finite-sample valid level for all Lebesgue-continuous null distributions, then it has power no more than α for any Lebesgue-continuous alternative. This is in stark contrast to the unconditional case (or conditioning on a discrete variable), in which case there exist finite-sample valid, consistent tests (e.g. permutations based on HSIC; see Rindt et al., 2021).

Intuitively, when detecting unconditional dependence $A \perp\!\!\!\perp B$, dependence can be missed (causing a Type II error) but Type I error arises only from sampling variability. By contrast, for $A \perp\!\!\!\perp B \mid C$, it is possible either to miss actual dependence (Type II) or falsely detect dependence (Type I) because subtle conditional effects of C have been overlooked. For the latter case, consider generating $C, A', B' \sim \mathcal{N}(0, 1)$, extracting the thirtieth decimal place of C as $C_{30} \in \{0, 1, \ldots, 9\}$, and then taking $A = C_{30} + A'$, $B = C_{30} + B'$. Unless we know to look at the thirtieth decimal place of C, A and B will seem to be strongly dependent and C irrelevant; in fact, however, all information that A carries about B is present in C, so $A \perp\!\!\!\perp B \mid C$. Shah and Peters (2020) show that for all test procedures, for any case which is truly conditionally dependent, the test has such a "blind spot" which is conditionally independent but "looks the same" to the test.

Interpretation with KCI. How do these issues manifest with KCI? We can show, in fact, that they arise *solely* because of the estimation of the conditional mean embedding.

In practice, conditional independence testing relies on empirical estimates constructed from finite samples. Given observations $\{(a_i, b_i, c_i)\}_{i=1}^n$, we first define the KCI statistic KCI_n as a U-statistic based on the true conditional mean embeddings $\mu_{A|C}$ and $\mu_{B|C}$:

$$\mathrm{KCI}_n = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h_{i,j} \qquad \text{where } h_{i,j} = (K_C)_{i,j} \, (K_A^{\mathrm{c}})_{i,j} \, (K_B^{\mathrm{c}})_{i,j}, \tag{6}$$

where $(K_C)_{i,j} = k_C(c_i, c_j)$ is the kernel matrix for C, $(K_A^{\mathrm{c}})_{i,j} = \langle \phi_A^{\mathrm{c}}(a_i, c_i), \phi_A^{\mathrm{c}}(a_j, c_j) \rangle_{\mathcal{H}_\mathcal{A}}$ with $\phi_A^{\mathrm{c}}(a_i, c_i) = \phi_A(a_i) - \mu_{A|C}(c_i)$ is the centered kernel matrix for A, and similarly $(K_B^{\mathrm{c}})_{i,j} = \langle \phi_B^{\mathrm{c}}(b_i, c_i), \phi_B^{\mathrm{c}}(b_j, c_j) \rangle_{\mathcal{H}_\mathcal{B}}$ with $\phi_B^{\mathrm{c}}(b_i, c_i) = \phi_B(b_i) - \mu_{B|C}(c_i)$ is that for B.

To run a KCI-based test, we require a *test threshold* t_n and reject the null whenever the KCI statistic exceeds t_n. This threshold t_n depends on the data size n, the choice of kernels, and the particular data distributions. K. Zhang et al. (2011) show that when $A \perp\!\!\!\perp B \mid C$, $n\mathrm{KCI}_n$ converges to a mixture of χ^2 variables,[3] so t_n could in principle be estimated by fitting the parameters of this limiting distribution. If we know the true $\mu_{A|C}$ and $\mu_{B|C}$, we can easily construct a finite-sample valid test with nontrivial power:

Proposition 4.1. *Suppose* $\sup_{a \in \mathcal{A}} k_A(a, a) \le \kappa_A$, $\sup_{b \in \mathcal{B}} k_B(b, b) \le \kappa_B$, $\sup_{c \in \mathcal{C}} k_C(c, c) \le \kappa_C$. *Then a test which rejects when* $\mathrm{KCI}_n > \tilde{t}_n \coloneqq 32 \kappa_A \kappa_B \kappa_C \sqrt{\frac{1}{n-1} \log \frac{1}{\alpha}}$ *has finite-sample level at most* α. *Moreover, if each kernel is* L^2*-universal, the test is consistent against fixed alternatives.*

The proof, given in Appendix B, is a simple consequence of Hoeffding's inequality for U-statistics. Although the resulting test is highly conservative – the correct threshold for the null distribution of KCI_n should be $\Theta(1/n)$ (K. Zhang et al., 2011, Theorem 3), much smaller than the chosen $\tilde{t}_n$ – the fact that it avoids the impossibility result of Shah and Peters (2020) indicates that the main challenge lies in estimating conditional mean embeddings.

Relationship to model-X. The recently popular "model-X" setting (Candes et al., 2018; Berrett et al., 2019; Grünwald et al., 2024) assumes that the conditional distribution of $A \mid C$ is known. This corresponds to perfect knowledge of $\mu_{A|C}$: for a characteristic (or a fortiori, L^2-universal) k_A, $\mu_{A|C}$ uniquely corresponds to $\mathrm{Law}(A \mid C)$. Given knowledge of both $A \mid C$ and $B \mid C$, the KCI-based test in Proposition 4.1 would be exactly valid; knowledge of only one is also sufficient using CIRCE rather than KCI (Pogodin et al., 2023; Pogodin et al., 2024). We discuss more aspects of the relationship to other CI tests in Appendix C.

5 Pitfalls of Kernel Choices for CI Testing in Practice

Since the true conditional mean embeddings are unknown, in practice we must use the empirical KCI statistic $\widehat{\mathrm{KCI}}_n$, which substitutes these embeddings with estimates $\widehat{\mu}_{A|C}$ and $\widehat{\mu}_{B|C}$. These embeddings are typically estimated via kernel ridge regression (Grünewälder et al., 2012; Li et al., 2024) with inputs c_i and labels $\phi_A(a_i)$ or $\phi_B(b_i)$. K. Zhang et al. (2011) used the same kernel for both regressions. Recognizing the need for high-quality regressions, Pogodin et al. (2023) proposed instead choosing separate kernels to minimize the leave-one-out validation error. Doing so introduces two new kernels, which we call $k_{C \to A}$ and $k_{C \to B}$. Pogodin et al. (2023) and Pogodin et al. (2024) then used k_C as either $k_{C \to A}$ or $k_{C \to B}$, implicitly assuming that a good kernel for this regression will also be a good kernel for measuring dependence.

We now demonstrate that the aforementioned choice for k_C – though computationally convenient – can be a very poor choice for measuring dependence in complex situations. For an intuitive example, consider an engineering problem involving high-dimensional vibration data: we wish to know if the behavior of part A is connected to that of part B given vibration data C. While predicting the

[3] Their Proposition 5 makes a stronger claim, that $\widehat{\mathrm{KCI}}_n$ does so under fixed-regularization ridge regression estimates for the conditional means; their argument (which was only sketched) appears to rely on a property that does not clearly always hold for this estimator, but does hold with the true $\mu_{A|C}$, $\mu_{B|C}$. Personal communication with the authors confirmed that they agree "there is a gap" between the published sketch and a true proof.

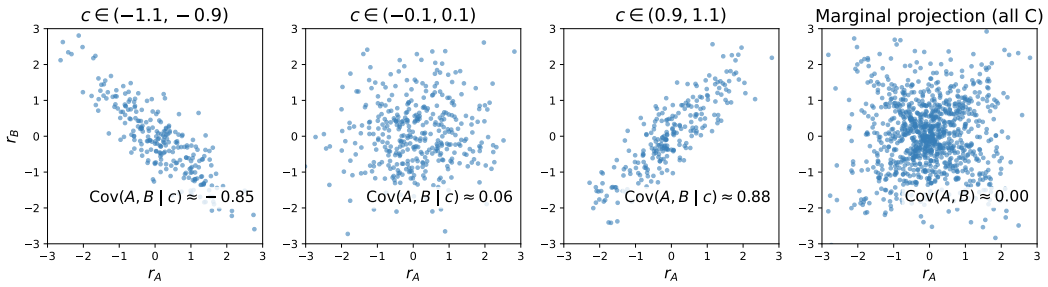


Figure 1: **Motivating example.** We simulate (r_A, r_B, C) following (7), where $\tau = 1$ and the residual correlation $\gamma(C) = \sin(C)$ introduces dependence that varies smoothly with C. The left three panels visualize samples of residuals for different slices of C, showing that $\mathrm{Cov}(A, B \mid C)$ changes substantially across C. The rightmost panel shows all of the residuals, for every value of C, where the averaged conditional covariance $\mathbb{E}_C[\mathrm{Cov}(A, B \mid C)]$ is zero. A kernel on C with an appropriately chosen lengthscale can instead focus on local regions where dependence is strong: if the lengthscale is too long, the conditional covariance will be "blurred out"; while if it is too short, there will be insufficient data to estimate the covariance.

behavior of either A or B depends on broad, long-term trends of C, the two parts may be coupled only by high-frequency sinusoidal resonances which require a substantially different kernel to efficiently detect. Using $k_{C\to A}$ or $k_{C\to B}$ then results in high Type-II error.

Motivated by this, consider a synthetic problem where A and B are determined as some functions of C plus noise factors which are zero mean, but potentially conditionally correlated given C:

$$C \sim \mathcal{N}(0,1), \quad A = f_A(C) + \tau\, r_A, \quad B = f_B(C) + \tau\, r_B,$$

where f_A, f_B are fixed functions, $\tau, \beta > 0$, and the additive residual terms (r_A, r_B) follow

$$(r_A, r_B) \mid C \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \gamma(C)\\ \gamma(C) & 1\end{bmatrix}\right), \quad \gamma(C) = \begin{cases}0 & \text{under } \mathfrak{H}_0\\ \sin(\beta C) & \text{under } \mathfrak{H}_1.\end{cases} \tag{7}$$

We use linear kernels for A and B, aligning closely to GCM, and a lengthscale-ℓ_C Gaussian kernel $k_C(C, C') = \exp\left(-\frac{(C-C')^2}{2\ell_C^2}\right)$, on C; GCM corresponds to $\ell_C = \infty$.

Figure 1 illustrates this setup under $\mathfrak{H}_1$: although the conditional covariance $\mathrm{Cov}(A, B \mid C) = \tau^2\, \mathbb{E}[r_A r_B \mid C]$ changes smoothly with C and alternates in sign, its expectation $\mathbb{E}_C[\mathrm{Cov}(A, B \mid C)]$ is nearly zero. As a result, methods like GCM—which effectively test this average—fail to detect dependence, highlighting the need for kernels on C that can localize to regions where the conditional covariance is nonzero.

The regressions which try to learn the functions f_A, f_B should use respective kernels $k_{C\to A}$ and $k_{C\to B}$ with lengthscales appropriate to those functions. On the other hand, the residuals' covariance varies on the lengthscale of γ, $1/\beta$, and so k_C's should be chosen to detect variations on that scale – which may be quite different than the appropriate scales for f_A and f_B.

In this setting, we can analytically evaluate the KCI, at least when using the true mean embeddings $\mu_{A|C}$ and $\mu_{B|C}$. (Details are given in Appendix F.1.) We first see, using properties of Gaussians, that

$$\mathrm{KCI} = \tau^4 \mathop{\mathbb{E}}_{C,C'}\left[k_C(C, C')\gamma(C)\gamma(C')\right] = \tau^4 \sqrt{\frac{\ell_C^2}{\ell_C^2 + 2}} \mathop{\mathbb{E}}_{(X,X')\sim\mathcal{N}_{\ell_C}}\left[\gamma(X)\gamma(X')\right] \tag{8}$$

for auxiliary variables $(X, X') \sim \mathcal{N}_{\ell_C} := \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 - \frac{1}{\ell_C^2+2} & \frac{1}{\ell_C^2+2}\\ \frac{1}{\ell_C^2+2} & 1 - \frac{1}{\ell_C^2+2}\end{bmatrix}\right)$. Under the null, we of course obtain $\mathrm{KCI} = 0$; under the alternative, we can use trigonometric identities to see

$$\mathrm{KCI} = \frac{1}{2}\tau^4 e^{-\beta^2} \sqrt{\frac{\ell_C^2}{\ell_C^2 + 2}} \left(e^{2\beta^2/(\ell_C^2+2)} - 1\right).$$

When $\ell_C \ll \sqrt{2}$, the square root term arising from $k_C(C, C')$ vanishes, giving zero KCI; for $\ell_C \gg \beta$, the other term coming from the covariance of γ vanishes, yielding the same problem. Consequently, for each β, the effective ℓ_C lies at an intermediate value that balances these effects (see Figure 2, left). GCM, with $\ell_C = \infty$, cannot detect dependence here at all.

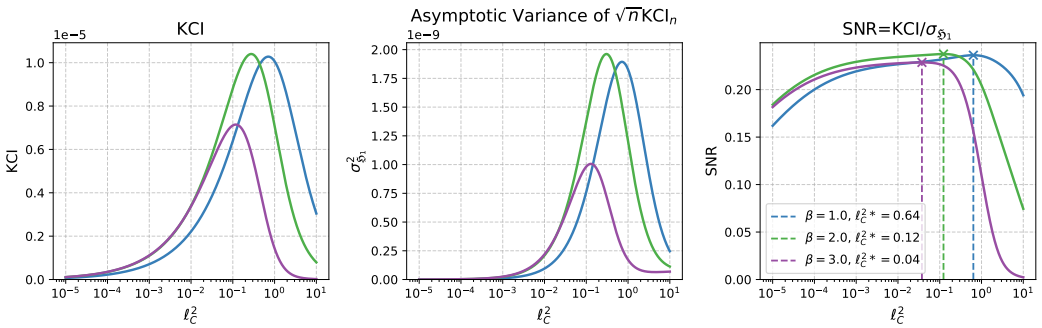


Figure 2: Effect of kernel lengthscale ℓ_C^2 on KCI, asymptotic variance $\sigma^2_{\mathfrak{H}_1}$, and approximate test power (SNR) for different conditional dependence lengthscale β, in the synthetic example (7) under the alternative. The optimal ${\ell_C^2}^*$ is selected by maximizing SNR. Different β values correspond to different ℓ_C^2 ranges yielding high approximate test power. Here we use noise scale $\tau = 0.1$.

Selecting a conditioning kernel. How can we choose the right C kernel for a given problem? One approach, following that taken in related settings (Jitkrittum et al., 2016; Liu et al., 2020; Xu et al., 2024), is to maximize the approximate power of the test, based on the following asymptotic result:

Proposition 5.1. *Under the alternative, there is a scalar $\hat{\sigma}^2_{\mathfrak{H}_1} \geq 0$ so that as $n \to \infty$,*

$$\sqrt{n}(\widehat{\mathrm{KCI}}_n - \widehat{\mathrm{KCI}}) \xrightarrow{d} \mathcal{N}(0, \hat{\sigma}^2_{\mathfrak{H}_1}). \tag{9}$$

As always, the hat here refers to the use of estimated mean embeddings, not to estimation of a quantity from samples; $\widehat{\mathrm{KCI}}$ and $\hat{\sigma}^2_{\mathfrak{H}_1}$ depend on the problem, the kernels, and the choice of $\widehat{\mu}_{A|C}$ and $\widehat{\mu}_{B|C}$, but not on n or any particular test sample. Under the alternative, we typically have $\hat{\sigma}^2_{\mathfrak{H}_1} > 0$, in which case the rejection probability is approximately

$$\Pr_{\mathfrak{H}_1}(\widehat{\mathrm{KCI}}_n > t_n) \sim \Phi\left(\frac{\sqrt{n}\,\widehat{\mathrm{KCI}}}{\hat{\sigma}_{\mathfrak{H}_1}} - \frac{\sqrt{n}\,t_n}{\hat{\sigma}_{\mathfrak{H}_1}}\right),$$

where $a \sim b$ means $\lim_{n\to\infty} a/b = 1$, Φ is the standard normal CDF, and t_n is any rejection threshold. We expect $t_n = \Theta(1/n)$, following the null distribution of KCI_n; the power is therefore dominated by the first term for reasonably large n, and the kernel yielding the most powerful test will approximately maximize the signal-to-noise ratio $\widehat{\mathrm{SNR}} = \widehat{\mathrm{KCI}}/\hat{\sigma}_{\mathfrak{H}_1}$.

We can estimate $\widehat{\mathrm{SNR}}$ by dividing $\widehat{\mathrm{KCI}}_n$ by the variance estimator given by Liu et al. (2020, Equation 5), and choose a kernel on a training split to maximize this estimate. (In independent work, W. Wang et al. (2025) used a similar scheme, but with a somewhat different estimator setup and with limited analysis; see Appendix C.1.) We can then use the selected kernel on a testing split; as long as the two splits are independent, this will not break the independence assumptions of the test procedure.

For a fixed $\widehat{\mu}_{A|C}$ and $\widehat{\mu}_{B|C}$, $\widehat{\mathrm{SNR}}_n$ in fact generalizes, identifying a good kernel:

Theorem 5.2 (Informal)**.** *Consider the U-statistic kernel $\hat{h}$ of $\widehat{\mathrm{KCI}}_n$; give it parameters ω, such as the parameters of k_C, in a finite-dimensional Banach space such that $\hat{h}$ is smooth with respect to those parameters. Then $\widehat{\mathrm{SNR}}_n$ converges uniformly to $\widehat{\mathrm{SNR}}$ over bounded sets of parameters with variance bounded away from zero; thus the maximizer of $\widehat{\mathrm{SNR}}_n$ approaches that of $\widehat{\mathrm{SNR}}$.*

This is a modification of the result of Liu et al. (2020, Theorem 6), since for fixed $\widehat{\mu}_{A|C}$, $\widehat{\mu}_{B|C}$ the U-statistic structure is very similar; a detailed statement and a proof are in Appendix D.

To evaluate whether maximizing the approximate test power is effective in practice, we compare the theoretical (approximate) power with the empirical power estimated from data. Figure 2 illustrates how the analytic results KCI, $\sigma^2_{\mathfrak{H}_1}$, and the corresponding SNR vary with the kernel lengthscale ℓ_C^2 in the synthetic example (7), where the optimal ${\ell_C^2}^*$ is obtained by maximizing the SNR. As shown in Figure 3 (right), the theoretical power curve (SNR vs.ℓ_C^2) closely tracks the empirical power curve

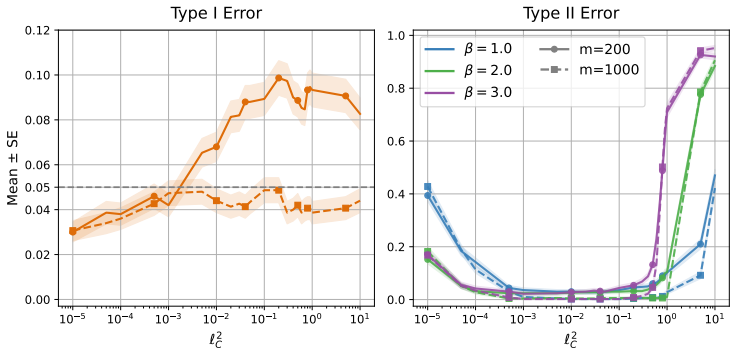


Figure 3: Means and standard error (over 500 runs) of Type-I/II errors on the synthetic (7) with $f_A = \cos, f_B = \exp, \tau = 0.1$ and different values of β, plotted against the kernel lengthscale ℓ_C^2. The training sample size is $m = 200$ (solid line with circles) or $m = 1000$ (dashed line with squares); the independent test set has size 200. The significance level is set at $\alpha = 0.05$. **Left**: When $m = 200$, varying ℓ_C^2 noticeably affects the Type-I error, for certain values of ℓ_C^2. In contrast, when $m = 1000$, the regressor is better trained, and the Type-I error remains well-controlled for all ℓ_C^2. **Right**: The empirical test power (1 - Type-II error) depends strongly on both β and ℓ_C^2, indicating the the importance of proper kernel selection for k_C.

((1 − Type II error) vs.ℓ_C^2), indicating that the selected ${\ell_C^2}^*$ remains effective in practice under the alternative hypothesis.

Although maximizing test power is effective under the alternative, it can substantially inflate Type-I error in conditional independence testing. In the unconditional independence settings of Liu et al. (2020) and Xu et al. (2024), the null threshold is determined by permutation, ensuring exact Type-I error control: any chosen kernel rejects at most at rate α. In our case, no such procedure is available; instead, we rely on asymptotic null approximations, which depend sensitively on kernel choice and regression quality, making null calibration delicate. Data splitting prevents overfitting to the points used to select ℓ_C^2, yet ℓ_C^2 can still overfit to the imperfect estimates $\widehat{\mu}_{A|C}$ and $\widehat{\mu}_{B|C}$. As shown in Figure 3 (left), Type-I error remains controlled with ample training data, but with limited training samples there exists a range of ℓ_C^2 values where it rises sharply. Power maximization tends to favor this region due to its higher $\widehat{\mathrm{SNR}}$. Hence, when ℓ_C^2 is chosen based on imperfect regressions, an inherent tension arises between Type-I error control and test power.

Relationship to wGCM selection scheme. One approach of Scheidegger et al. (2022) identifies a weight function by trying to predict the sign of the product of residuals; if that prediction works perfectly, then it changes GCM from measuring the average residual correlation to measuring the average absolute value of the residual correlation, which is potentially much more powerful. As discussed previously, this is essentially equivalent to choosing k_C, which they do by setting it to a ± 1 indicator of whether the residual signs are predicted to be the same or not. While the scheme works differently than ours, it has essentially the same trade-offs as other approaches for kernel selection.

6 Type-I Error Inflation with Regression Errors

As shown by Proposition 4.1 and reinforced by the previous section's example, the fundamental challenges in conditional independence testing stem from the estimation of conditional mean embeddings. To further illustrate this point, we examine the effect of regression errors by letting $\widehat{\mu}_{A|C} = \mu_{A|C} + \Delta_{A|C}$ and $\widehat{\mu}_{B|C} = \mu_{B|C} + \Delta_{B|C}$. Under the null hypothesis and the setting of problem (7), we can explicitly characterize $\widehat{\mathrm{KCI}}$ and its asymptotic variance in terms of $\Delta_{A|C}$ and $\Delta_{B|C}$. This allows us to quantify how regression errors distort the KCI statistic and its variance. More importantly, the discrepancy between the true limiting distribution and the null approximation can be analyzed in terms of $\widehat{\mathrm{KCI}}$ and $\mathrm{Var}(\widehat{\mathrm{KCI}})$, without requiring assumptions on the data distribution. This allows us to derive formal bounds that link conditional mean estimation error to test validity more generally.

Effect on moments. The following result, proved in Appendix E, is a more convenient form of textbook results about U-statistics (Serfling, 1980, Section 5.2.1) for kernel methods:

Theorem 6.1. *Let $h(X, X') = \langle \phi_h(X), \phi_h(X') \rangle$ with mean embedding $\mu_h = \mathbb{E}_X \phi_h(X)$ and the centered covariance operator $\mathfrak{C}_h = \mathbb{E}_X[\phi_h(X) \otimes \phi_h(X)] - \mu_h \otimes \mu_h$. Define $\nu_1 = \langle \mu_h, \mathfrak{C}_h \mu_h \rangle$ and $\nu_2 = \|\mathfrak{C}_h\|^2_{\mathrm{HS}}$. The corresponding U-statistic satisfies*

$$U_n = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j), \quad \mathbb{E}[U_n] = U = \|\mu_h\|^2, \quad \mathrm{Var}(U_n) = \frac{4}{n}\nu_1 + \frac{2}{n(n-1)}\nu_2.$$

The function h of (6), for KCI_n, has this form with $\phi_h(X) = \phi^{\mathrm{c}}_A(A, C) \otimes \phi^{\mathrm{c}}_B(B, C) \otimes \phi_C(C)$; moreover, $\mu_h = \mathbb{E}_X[\phi_h(X)] = \mathbb{E}_C \left[\mathbb{E}_{AB}[\phi^{\mathrm{c}}_A(A, C) \otimes \phi^{\mathrm{c}}_B(B, C) \mid C] \otimes \phi_C(C) \right]$ is exactly the definition (3) of $\mathfrak{C}_{\mathrm{KCI}}$. Thus, under the null where $\mu_h = \mathfrak{C}_{\mathrm{KCI}} = 0$, we have $\mathbb{E}\,\mathrm{KCI}_n = \mathrm{KCI} = 0$ and $\mathrm{Var}(\mathrm{KCI}_n) = \frac{2}{n(n-1)}\nu_2$.

$\widehat{\mathrm{KCI}}_n$ has the same decomposition, except $\mu_{\hat{h}} = \mathbb{E}_C \left[\mathbb{E}_{AB}[\, \widehat{\phi}^{\mathrm{c}}_A(A, C) \otimes \widehat{\phi}^{\mathrm{c}}_B(B, C) \mid C \,] \otimes \phi_C(C) \right]$ is now *not* zero if the error in $\widehat{\mu}_{A|C}$, $\widehat{\mu}_{B|C}$ is not exactly conditionally independent. As shown by Pogodin et al. (2024), with linear kernels k_A and k_B, under the null we have

$$\mathbb{E}\,\widehat{\mathrm{KCI}}_n = \widehat{\mathrm{KCI}} = \mathbb{E}\left[k_C(C, C') \, \langle \Delta_{A|C}(C), \Delta_{A|C}(C') \rangle_{\mathcal{H}_\mathcal{A}} \, \langle \Delta_{B|C}(C), \Delta_{B|C}(C') \rangle_{\mathcal{H}_\mathcal{B}} \right]. \quad (10)$$

As they note, we typically expect $\Delta_{A|C}$ and $\Delta_{B|C}$ to be relatively smooth functions of C; thus it is reasonable to expect that $\widehat{\mathrm{KCI}}$ can be nontrivial even though they were trained on independent datasets. Perhaps even more significantly, for fixed regression functions, it will generally be the case that $\nu_1 = \langle \mu_{\hat{h}}, \mathfrak{C}_{\hat{h}} \mu_{\hat{h}} \rangle > 0$. This implies that the standard deviation decreases as n shrinks only like $\Theta(1/\sqrt{n})$, rather than the $\Theta(1/n)$ obtained when $\Delta_{A|C}, \Delta_{B|C}$ are zero. (The exact expression for the variance of $\widehat{\mathrm{KCI}}_n$ in the synthetic example is given in Appendix F.2.)

Multi-dimensional C example. Though the analytical expression of $\widehat{\mathrm{KCI}}$ does not require assumptions on the data structure, so far we implicitly presumed that the same features of C are used both in the true/estimated conditional means and the conditional dependence $\gamma(C)$. We thus extend our analysis to cases where C is multi-dimensional, considering two scenarios: (1) using the same dimensions of C for both conditional means and dependence, and (2) using separate dimensions. This allows us to study how the information in C influences spurious dependence. See Appendix F.3 for the setup and Appendix H.1 for additional experiments.

Table 1: Comparison of Testing Results for Two Conditional Dependence Scenarios

Scenario	**Type I Error**	**Type II Error**
Scenario 1: Shared dimensions	0.21	0.0
Scenario 2: Separate dimensions	0.10	0.08

As observed in Table 1, Scenario 1 exhibits a notably higher Type I error (0.21) compared to Scenario 2 (0.10). This increase arises from regression errors leaking correlated noise into the test statistic when regressions $C \to A, C \to B$, and dependence share the same dimension. Consequently, Scenario 1 generates inflated KCI values, increasing false positives. In contrast, Scenario 2, with separate dimensions, shows lower Type I error but slightly higher Type II error, illustrating a trade-off driven by regression error correlations.

Real-world experiments. Appendix H.2 presents experiments on a real-world age-estimation task, following Y. Zhang et al. (2025), which investigate whether certain face regions are informative for age prediction. The results show imperfect regression can make the test more prone to indicating dependence, highlighting the practical impact of regression errors on CI testing.

Effect on null calibration. Standard methods for setting a test threshold for KCI do not incorporate regression error; rather, they rely on the asymptotic distribution of KCI_n. For instance, K. Zhang et al. (2011) estimate the parameters of a χ^2 mixture or a gamma approximation thereof, while Pogodin et al. (2024) suggest a wild bootstrap. In either case, the null threshold scales as $\Theta(1/n)$. However, if regression errors remain fixed while the number of test points grows, $\widehat{\mathrm{KCI}}_n = \Theta(1) + \mathcal{O}_p(1/\sqrt{n})$ will almost surely exceed the threshold, inflating Type-I error. This shows that regression errors must shrink as n grows, and motivates establishing the required decay rate.

Asymptotics. When $\nu_1 > 0$, $\sqrt{n}(U_n - U)$ converges to a normal distribution (Proposition 5.1); when $\nu_1 = 0$ but $\nu_2 > 0$, $n(U_n - U)$ converges in distribution to a weighted mixture of centered χ^2 variables (Serfling, 1980, Section 5.5). We can thus ask: under the null, how likely is a sample from KCI_n to exceed a test threshold set according to the limiting distribution of $n\mathrm{KCI}_n$?

Theorem 6.2. *Assume that $A \perp\!\!\!\perp B \mid C$. Let $Z_1 = \widehat{\mathrm{KCI}}_n$ and $Z_2 \sim \mathcal{N}\left(\widehat{\mathrm{KCI}}, \mathrm{Var}(\widehat{\mathrm{KCI}}_n)\right)$ be a normal variable moment-matched to Z_1. Let $q > 0$ and $\rho \in (0,1)$; define $T_1 = \sqrt{(1-\rho)/\rho}$ and $T_2 = \Phi^{-1}(1-\rho)$, where Φ is the standard normal CDF. Then the following holds for $i \in \{1,2\}$:*

$$\Pr\left(Z_i > \frac{q}{n}\right) \leq \rho \quad \text{if } q \geq n\widehat{\mathrm{KCI}} + T_i \sqrt{n^2 \,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}.$$

The proof is in Appendix G.1; the case for $\widehat{\mathrm{KCI}}_n$ is more precisely applicable, but using asymptotic normality gives better dependence on ρ. This theorem provides an upper bound on the probability that the inflated statistic $\widehat{\mathrm{KCI}}_n$ exceeds a nominal null threshold. Intuitively, the bound shows that if the regression bias induced$\widehat{\mathrm{KCI}}$ or the variance $\mathrm{Var}(\widehat{\mathrm{KCI}}_n)$ are non-negligible, the effective threshold q/n must grow proportionally to $\widehat{\mathrm{KCI}} + T_i\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}$ in order to maintain the level ρ.

The following result, stated with explicit constants and proved in Appendix G.2, instead bounds the approximation error between wild bootstrap and a moment-matched normal to $n\widehat{\mathrm{KCI}}_n$. The wild bootstrap generates surrogate samples $Y = \frac{1}{n}\sum_{i\neq j} \hat{h}_{ij}\varepsilon_i\varepsilon_j$, where the random multipliers ε_i are drawn independently and $\hat{h}_{ij}$ is defined analogously to (6) using estimated conditional mean embeddings. For given kernel matrix $\hat{h}_{ij}$, the wild bootstrap stastistic Y has zero mean and a variance that closely matches that of $n\mathrm{KCI}_n$ under the null, providing a practical approximation to the null distribution of the KCI statistic.

Theorem 6.3. *Assume $A \perp\!\!\!\perp B \mid C$, and let $Y = \frac{1}{n}\sum_{i,j=1}^{n} \hat{h}_{ij}\,\varepsilon_i\,\varepsilon_j$, where $\varepsilon_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$. Let $\widehat{H}$ be the matrix with entries $\hat{h}_{i,j}$; assume $\|\widehat{H}\|_F^2 > 2\|\widehat{H}\|_{\mathrm{op}}^2$. Let $Z_n \sim \mathcal{N}\Big(\widehat{\mathrm{KCI}},\ \mathrm{Var}(\widehat{\mathrm{KCI}}_n)\Big)$ be the moment-matched normal approximation to $\widehat{\mathrm{KCI}}_n$. Then, there exist constants $R_{i,n} \to R_i > 0$ for $i \in \{1,2,3\}$ (depending on the distribution of Y and $\widehat{\mathrm{KCI}}_n$) such that*

$$\sup_{x\in\mathbb{R}}\Big|\Pr(Y \mid \widehat{H} \leq x) - \Pr(n\,Z_n \leq x)\Big| \;\leq\; R_{1,n} + R_{2,n}\,\frac{\widehat{\mathrm{KCI}}}{\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}} + R_{3,n}\,\Big|\frac{\mathrm{Var}(Y \mid \widehat{H})}{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)} - 1\Big|.$$

Noting that $n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n) \sim 4n\nu_1 + 2\nu_2$. In Theorem 6.3 we would like $\widehat{\mathrm{KCI}} = o(1/n)$ and $\nu_1 = o(1/n)$. Under these conditions, the remainder terms R_2 and R_3 in Theorem 6.3 vanish, leaving only a constant R_1 gap between the wild bootstrap statistic Y and the normal approximation nZ_n; for the bound to be informative, R_1 should be small. Similarly, in Theorem 6.2, the asymptotic threshold behaves correctly if $4n\nu_1 + 2\nu_2 \to v \leq (q/T_i)^2$, which is most easily achieved when $\nu_1 = o(1/n)$ and $\nu_2 = \Theta(1)$. Although exact calibration would require more detailed knowledge of the distribution, the fact that both bounds demand $\widehat{\mathrm{KCI}}$ and ν_1 to scale as $o(1/n)$ strongly indicates that this is the correct scaling for controlling Type-I error.

7 Discussion

We provided a novel framing of the KCI test, one which helped us connect it closely to GCM-based tests. We explained how this category of tests interacts with the famed hardness result of Shah and Peters (2020), identifying regression error as the key difficulty, and showing bounds on the excess Type I error based on the amount of regression error. We showed that, contra the assumptions of most prior work, selecting a k_C kernel specifically for testing can be of vital importance in achieving test power, but that doing so can exacerbate Type I error.

While CI testing remains fundamentally difficult, our work makes a step towards understanding how this difficulty manifests in practice, and demonstrates paths towards addressing it. This underscores that users of GCM- or KCI-type tests must carefully consider how to mitigate spurious residual dependence under the null—something that sample splitting alone does *not* resolve.

Acknowledgments and Disclosure of Funding

The authors would like to thank Aaron Wei for productive discussions.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, the Canada CIFAR AI Chairs program, the Gatsby Charitable Foundation, Calcul Québec, the BC DRI Group, the Digital Research Alliance of Canada, and a Google research gift.

References

Thomas B. Berrett, Yi Wang, Rina Foygel Barber, and Richard J. Samworth (2019). "The Conditional Permutation Test for Independence While Controlling for Confounders." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.1, pp. 175–197. DOI: 10.1111/rssb.12340.

Michael J Buckley and Geoffrey K Eagleson (1988). "An approximation to the distribution of quadratic forms in normal random variables." *Australian Journal of Statistics* 30.1, pp. 150–159.

Emmanuel Candes, 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: Statistical Methodology* 80.3, pp. 551–577.

J. J. Daudin (1980). "Partial association measures and an application to qualitative regression." *Biometrika* 67.3, pp. 581–590.

Namrata Deka and Danica J. Sutherland (2023). "MMD-B-Fair: Learning Fair Representations with Statistical Testing." *AISTATS*. arXiv: 2211.07907.

Carl-Gustav Esseen (1945). "Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law."

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A Kernel Two-Sample Test." *Journal of Machine Learning Research* 13.25, pp. 723–773.

Arthur Gretton, Olivier Bousquet, Alex J. Smola, and Bernhard Schölkopf (2005). "Measuring Statistical Dependence with Hilbert-Schmidt Norms." *Algorithmic Learning Theory*.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola (2008). "A Kernel Statistical Test of Independence." *NeurIPS*.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil (2012). "Conditional mean embeddings as regressors." *ICML*.

Peter Grünwald, Alexander Henzi, and Tyron Lardy (2024). "Anytime-Valid Tests of Conditional Independence Under Model-X." *Journal of the American Statistical Association* 119.546, pp. 1554–1565.

László Györfi and Harro Walk (2012). "Strongly consistent nonparametric tests of conditional independence." *Statistics & Probability Letters* 82.6, pp. 1145–1150. ISSN: 0167-7152. DOI: 10.1016/j.spl.2012.02.023.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro (2016). "Equality of Opportunity in Supervised Learning." *NeurIPS*. arXiv: 1610.02413.

Wassily Hoeffding (1963). "On Sequences of Sums of Independent Random Vectors." *Journal of the American Statistical Association* 58.301, pp. 13–30.

J. P. Imhof (1961). "Computing the Distribution of Quadratic Forms in Normal Variables." *Biometrika* 48.3/4, pp. 419–426. DOI: 10.1093/biomet/48.3-4.419.

Wittawat Jitkrittum, Zoltan Szabo, Kacper Chwialkowski, and Arthur Gretton (2016). "Interpretable Distribution Features with Maximum Testing Power." *NeurIPS*. arXiv: 1605.06796.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman (2022). "Local permutation tests for conditional independence." *The Annals of Statistics* 50.6, pp. 3388–3414. arXiv: 2112.11666.

Ilja Klebanov, Ingmar Schuster, and Timothy John Sullivan (2020). "A rigorous theory of conditional mean embeddings." *SIAM Journal on Mathematics of Data Science* 2.3, pp. 583–606.

Anne Leucht and Michael H. Neumann (2013). "Dependent wild bootstrap for degenerate U- and V-statistics." *Journal of Multivariate Analysis* 117, pp. 257–280. DOI: 10.1016/j.jmva.2013.03.003.

Zhu Li, 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, pp. 1–51.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland (2020). "Learning Deep Kernels for Non-Parametric Two-Sample Tests." *ICML*. arXiv: 2002.09116.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt J. Kusner, Arthur Gretton, and Krikamol Muandet (2021). "Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction." *ICML*. arXiv: 2105.04544.

Robb J Muirhead (2009). *Aspects of multivariate statistical theory*. John Wiley & Sons.

Junhyung Park and Krikamol Muandet (2020). "A measure-theoretic approach to kernel conditional mean embeddings." *NeurIPS*.

Roman Pogodin, Namrata Deka, Yazhe Li, Danica J. Sutherland, Victor Veitch, and Arthur Gretton (2023). "Efficient Conditionally Invariant Representation Learning." *ICLR*. arXiv: 2212.08645.

Roman Pogodin, Antonin Schrab, Yazhe Li, Danica J. Sutherland, and Arthur Gretton (2024). *Practical Kernel Tests of Conditional Independence*. arXiv: 2402.13196.

Felipe Maia Polo, Yuekai Sun, and Moulinath Banerjee (2023). "Conditional independence testing under misspecified inductive biases." *NeurIPS*.

David Rindt, Dino Sejdinovic, and David Steinsaltz (2021). "Consistency of permutation tests of independence using distance covariance, HSIC and dHSIC." *Stat* 10.1, e364. DOI: 10.1002/sta4.364.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann (2022). "The weighted generalised covariance measure." *Journal of Machine Learning Research* 23.273.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai (2017). "Model-Powered Conditional Independence Test." *NeurIPS*. arXiv: 1709.06138.

Robert J. Serfling (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

Rajen Shah and Jonas Peters (2020). "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure." *Annals of Statistics* 48.3, pp. 1514–1538. DOI: 10.1214/19-AOS1857.

Le Song, Jonathan Huang, Alex J. Smola, and Kenji Fukumizu (2009). "Hilbert Space Embeddings of Conditional Distributions." *ICML*.

Peter Spirtes, Clark Glymour, and Richard Scheines (1993). *Causation, Prediction, and Search*. Springer New York. DOI: 10.1007/978-1-4612-2748-9.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet (2011). "Universality, Characteristic Kernels and RKHS Embedding of Measures." *Journal of Machine Learning Research* 12.70, pp. 2389–2410.

Ingo Steinwart and Andreas Christmann (2008). *Support Vector Machines*. Springer New York.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton (2017). "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy." *ICLR*. arXiv: 1611.04488.

Zoltán Szabó and Bharath K. Sriperumbudur (2018). "Characteristic and Universal Tensor Product Kernels." *Journal of Machine Learning Research* 18.233.

Wenjie Wang, Mingming Gong, Biwei Huang, James Bailey, Bo Han, Kun Zhang, and Feng Liu (2025). *Practical Kernel Learning for Kernel-based Conditional Independent Test*. URL: https://openreview.net/forum?id=GPcSYm89wK.

Zihao Wang and Victor Veitch (2022). *The Causal Structure of Domain Invariant Supervised Representation Learning*. arXiv: 2208.06987.

Nathaniel Xu, Feng Liu, and Danica J. Sutherland (2024). *Learning Deep Kernels for Non-Parametric Independence Testing*. arXiv: 2409.06890.

Hao Zhang, Shuigeng Zhou, and Jihong Guan (2018). "Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1.

Hao Zhang, Shuigeng Zhou, Kun Zhang, and Jihong Guan (2017). "Causal discovery using regression-based conditional independence tests." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1.

Jin-Ting Zhang (2005). "Approximate and asymptotic distributions of chi-squared–type mixtures with applications." *Journal of the American Statistical Association* 100.469, pp. 273–285.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf (2011). "Kernel-based Conditional Independence Test and Application in Causal Discovery." *UAI*.

Yi Zhang, Linjun Huang, Yun Yang, and Xiaofeng Shao (2025). "Testing Conditional Mean Independence Using Generative Neural Networks." *ICML*.

Zhifei Zhang, Yang Song, and Hairong Qi (2017). "Age progression/regression by conditional adversarial autoencoder." *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818.

A Conditional independence decomposition

We recall and prove Theorem 2.2, which extends results of Daudin (1980).

Theorem 2.2. *Random variables A and B are conditionally independent given C if and only if*

$$\mathop{\mathbb{E}}_{C}\Big[\, w(C) \mathop{\mathbb{E}}_{AB|C}\big[\,(f(A)-\mathbb{E}[f(A)\mid C])\,(g(B)-\mathbb{E}[g(B)\mid C])\;\mid C\big]\,\Big]=0, \tag{11}$$

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

Proof. (i) Let A and B be conditionally independent given C. Let $\tilde{f}\in L^2_{AC}$ and $\tilde{g}\in L^2_{BC}$. Then by Definition 2.1, almost surely in C it holds that

$$\mathbb{E}\big[\,\tilde{f}(A,C)\,\tilde{g}(B,C)\mid C\,\big]=\mathbb{E}\big[\,\tilde{f}(A,C)\mid C\,\big]\;\mathbb{E}\big[\,\tilde{g}(B,C)\mid C\,\big],$$

which is equivalent to the statement that almost surely in C,

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

Since this expectation is almost surely zero, it holds for any $w\in L^2_C$ that

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

Given any $f\in L^2_A$ and any $g\in L^2_B$, we can choose $\tilde{f}(\cdot,c)=f$ and $\tilde{g}(\cdot,c)=g$ to simply ignore the second argument. These functions satisfy $\tilde{f}\in L^2_{AC}$ and $\tilde{g}\in L^2_{BC}$. Then, as desired,

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

(ii) Suppose (11) holds for all functions $\tilde{f}\in L^2_A$, $\tilde{g}\in L^2_B$, and $\tilde{w}\in L^2_C$. Let P_C denote the marginal distribution of C, and let $P_{A|C}$, $P_{B|C}$, and $P_{AB|C}$ denote the conditional distributions of A, B, and (A,B) given C, respectively. Let $\mathcal{Q}$ be a Borel subset of the image set of C. Pick $w^*=\mathbb{1}_{\mathcal{Q}}\in L^2_C$, where $\mathbb{1}_{\mathcal{Q}}$ is the indicator function of $\mathcal{Q}$. Substituting this choice into equation (11) yields

$$\int_{\mathcal{Q}}\mathop{\mathbb{E}}_{AB|C}\big[\,(\tilde{f}(A)-\mathbb{E}[\,\tilde{f}(A)\mid C\,])\,(\tilde{g}(B)-\mathbb{E}[\,\tilde{g}(B)\mid C])\mid C\,\big]dP_C=0,$$

Since this holds for all Borel sets $\mathcal{Q}$, it follows that the integrand must vanish almost surely with respect to P_C. That is, for P_C-almost every value of C, (11) implies that

$$\mathbb{E}\big[\,\tilde{f}(A)\,\tilde{g}(B)\mid C=c\,\big]=\mathbb{E}\big[\,\tilde{f}(A)\mid C=c\,\big]\;\mathbb{E}\big[\,\tilde{g}(B)\mid C=c\,\big].$$

Given any $f\in L^2_{AC}$ and any $g\in L^2_{BC}$, for each $C=c$ in its domain, $f(\cdot,c)\in L^2_A$ and $g(\cdot,c)\in L^2_B$ for almost every c. Thus, for any $f\in L^2_{AC}$ and any $g\in L^2_{BC}$, we have for almost every c,

$$\mathbb{E}\big[\,f(A,c)\,g(B,c)\mid C=c\,\big]=\mathbb{E}\big[\,f(A,c)\mid C=c\,\big]\;\mathbb{E}\big[\,g(B,c)\mid C=c\,\big],$$

which is precisely Definition 2.1. This completes the proof. □

B Finite-sample valid test with exact mean embeddings

We recall and prove Proposition 4.1.

Proposition 4.1. *Suppose* $\sup_{a\in\mathcal{A}}k_A(a,a)\le\kappa_A$, $\sup_{b\in\mathcal{B}}k_B(b,b)\le\kappa_B$, $\sup_{c\in\mathcal{C}}k_C(c,c)\le\kappa_C$. *Then a test which rejects when* $\mathrm{KCI}_n>\tilde{t}_n:=32\kappa_A\kappa_B\kappa_C\sqrt{\frac{1}{n-1}\log\frac{1}{\alpha}}$ *has finite-sample level at most* α. *Moreover, if each kernel is L^2-universal, the test is consistent against fixed alternatives.*

Proof. KCI_n is a U-statistic with kernel

$$k_C(c, c') \langle \phi_A(a) - \mu_{A|C}(c), \phi_A(a') - \mu_{A|C}(c') \rangle_{\mathcal{H}_\mathcal{A}} \langle \phi_B(b) - \mu_{A|C}(c), \phi_A(b') - \mu_{A|C}(c') \rangle_{\mathcal{H}_\mathcal{B}}.$$

We have that

$$\|\phi_A(a)\| = \sqrt{\langle \phi_A(a), \phi_A(a) \rangle} = \sqrt{k_A(a, a)} \leq \sqrt{\kappa_A}$$

and by Jensen's inequality

$$\|\mu_{A|C}(c)\| = \big\|\mathbb{E}[\phi_A(A) \mid C = c]\big\| \leq \mathbb{E}[\|\phi_A(A)\| \mid C = c] \leq \sqrt{\kappa_A},$$

so that $\|\phi_A(a) - \mu_{A|C}(c)\| \leq 2\sqrt{\kappa_A}$. Hence, by Cauchy-Schwarz,

$$|\langle \phi_A(a) - \mu_{A|C}(c), \phi_A(a') - \mu_{A|C}(c') \rangle_{\mathcal{H}_\mathcal{A}}| \leq 4\kappa_A.$$

Similarly, $|\langle \phi_B(b) - \mu_{B|C}(c), \phi_B(b') - \mu_{B|C}(c') \rangle_{\mathcal{H}_\mathcal{B}}| \leq 4\kappa_B$. Thus the kernel of the U-statistic KCI_n has absolute value at most $16\kappa_A\kappa_B\kappa_C$. Hoeffding (1963)'s inequality for U-statistics (c.f. Serfling, 1980, Section 5.6.1, Theorem A) thus shows that when $\mathrm{KCI} = 0$,

$$\Pr\left(\mathrm{KCI}_n \geq t_n\right) \leq \exp\left(-\frac{2\lfloor n/2 \rfloor\, t_n^2}{4 \cdot \left(16\kappa_A\kappa_B\kappa_C\right)^2}\right) \leq \exp\left(-\frac{(n-1)\, t_n^2}{\left(32\kappa_A\kappa_B\kappa_C\right)^2}\right) = \alpha,$$

showing finite-sample validity of the test.

On the other hand, when $A \not\perp\!\!\!\perp B \mid C$, since each kernel is L^2-universal, we know that $\mathrm{KCI} > 0$. Thus a symmetric application of Hoeffding's inequality tells us that once n is large enough that $t_n < \mathrm{KCI}/2$, we have that

$$\begin{aligned}\Pr\left(\mathrm{KCI}_n < t_n\right) &= \Pr\left(\mathrm{KCI} - \mathrm{KCI}_n > \mathrm{KCI} - t_n\right) \\ &\leq \exp\left(-(n-1)\left(\frac{\mathrm{KCI} - t_n}{32\kappa_A\kappa_B\sqrt{\kappa_C}}\right)^2\right) \leq \exp\left(-(n-1)\left(\frac{\mathrm{KCI}/2}{32\kappa_A\kappa_B\sqrt{\kappa_C}}\right)^2\right) \to 0,\end{aligned}$$

and hence for any fixed alternative, the probability of a Type-II error goes to zero. □

C Relationship to other testing methods

Relationship to other CI tests. One major category of conditional independence tests are based on variations of approximate permutation, i.e. that samples with similar C values have similar A and B distributions, which can be exploited either by "swapping" samples with nearby C values (e.g. Sen et al., 2017; Berrett et al., 2019; Kim et al., 2022) or by producing bins of C values and assuming the distribution is constant within (e.g. Györfi and Walk, 2012). While this approach might seem fundamentally different than the regression or conditional mean embedding approaches, we emphasize that it is not. For instance, Kim et al. (2022) assume that the Hellinger or Rényi distance between $A \mid C = c$ and $A \mid C = c'$ is at most a constant times $\|c - c'\|$, and the same for B; similar assumptions underlie all methods of this type. This smoothness justifies using the distribution $A \mid C = c'$ to estimate $A \mid C = c$ for some similar value of c'. Bearing in mind the one-to-one correspondence between mean embeddings and distributions, this assumption is essentially equivalent to using a nearest-neighbor type estimator for $\widehat{\mu}_{A|C}, \widehat{\mu}_{B|C}$.

Another recent CI test is the Rao-Blackwellized Predictor Test, RBPT (Polo et al., 2023). This method is regression-based, but based on comparing predictors of $B \mid A, C$ to an averaged predictor of $B \mid C$. This structure makes it harder to compare to the KCI-type tests directly, but we note that it relies on a good estimate of $A \mid C$ and hence is essentially in the model-X framework. Like most tests in this area, it suffers from severe bias problems, as discussed by Pogodin et al. (2024).

Smoothness of distributions. In the model-X setting where the conditional distribution is only approximately known, Berrett et al. (2019, Section 5) bound the worst-case inflation of the Type I error for two common model-X tests by at most the average conditional total variation distance between the true distribution and the approximation. Generic distribution modeling methods are likely to succeed in this sense only if the distribution changes slowly in total variation. Similarly, the bound of Kim et al. (2022) in a permutation case assumes that the distribution changes slowly in Hellinger distance; note that the total variation distance is upper-bounded by a constant times the

Hellinger distance, and so slow Hellinger change is a (slightly) stronger assumption than slow total variation change.

By contrast, while the precise conditions for effective conditional mean embedding estimation are complex (Li et al., 2024), we can roughly expect them to work when the mean embedding changes smoothly as a function of C; that is, the maximum mean discrepancy, MMD (Gretton et al., 2012), between $A \mid C = c$ and $A \mid C = c'$ changes slowly as a function of c. It is easy to see (e.g. Xu et al., 2024) that for bounded kernels, the MMD is a lower bound on the total variation. Thus, in the settings where the bounds of Berrett et al. (2019) and Kim et al. (2022) are applicable, we can *roughly* expect that $\mu_{A|C}$, $\mu_{B|C}$ should also be learnable. The reverse, however, is not true: the total variation is a much stronger distance than the MMD, i.e. there are many cases where the MMD is quite small and the total variation is very large. For instance, the Gaussian-kernel MMD between two nearby point masses will be small, while for total variation and Hellinger it will be maximal.

C.1 KCI-Power method

W. Wang et al. (2025), in unpublished parallel work, proposed choosing the parameters of kernels for a KCI test based on signal-to-noise ratio maximization, in a scheme very similar to that discussed in Section 5. They suggested doing it only for $k_C(C, C')$ and not for the mean embeddings, as did we; in a few application settings, they obtained marginal improvements to the test power. They said they did not update the A or B kernels because "due to the presence of conditional expectation bias, we empirically found that updating these parameters does not yield residual matrices with higher expected test power." In our view, the failure is exactly for the reasons identified in Section 5. While this can in fact be termed "conditional expectation bias," we think that our theoretical results and framing of exactly how this problem arises adds significant understanding over their paper.

It is also worth noting that W. Wang et al. also provided a decomposition of $\mathfrak{C}_{\mathrm{KCI}}$ identical to ours, although they originated from the previous definition as in our Footnote 2, rather than our novel first-principles derivation.

D Generalization bound for SNR

For a formal version of Theorem 5.2, we generalize the proof of Liu et al. (2020, Theorem 6) to other second-order U-statistics.

Given a set of samples $X_1, \ldots, X_n$ and a function h, define

$$U_n \coloneqq \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} h(X_i, X_j) \qquad \to U \coloneqq \mathbb{E}\, h(X, X')$$

$$\sigma^2_{\mathfrak{H}_1,n} \coloneqq \frac{4}{n} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{j=1}^{n} h(X_i, X_j) \right)^2 - 4 \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j) \right)^2 \to \sigma^2_{\mathfrak{H}_1} \coloneqq 4 \mathop{\mathbb{E}}_{X} \left[\mathop{\mathrm{Var}}_{X'} [h(X, X') \mid X] \right]$$

$$\mathrm{SNR}_{n,\lambda} \coloneqq U_n / \sqrt{\sigma^2_{\mathfrak{H}_1,n} + \lambda} \qquad \to \mathrm{SNR}_\lambda \coloneqq U / \sqrt{\sigma^2_{\mathfrak{H}_1} + \lambda}$$

and let $\mathrm{SNR} \coloneqq \mathrm{SNR}_0$.

Here U_n is the usual second-order U-statistic; we assume, without loss of generality, that $h(x, x') = h(x', x)$ for all x, x'. We know from Section 5.2.1 of Serfling (1980) (also see Theorem 6.1) that $\mathrm{Var}(U_n) = 4\nu_1/n + \mathcal{O}(1/n^2)$. The estimator $\nu_{1,n}$ follows the biased estimator used by Liu et al. (2020); while Sutherland et al. (2017) used an unbiased variance estimator, the biased estimator is much simpler and also performs better in this setting (Deka and Sutherland, 2023).

Note that using $X = (A, B, C)$ and $h((A, B, C), (A', B', C'))$ given by

$$k_C(C, C')\, \langle \phi_A(A) - \widehat{\mu}_{A|C}(C), \phi_A(A') - \widehat{\mu}_{A|C}(C') \rangle_{\mathcal{H}_\mathcal{A}}\, \langle \phi_B(B) - \widehat{\mu}_{B|C}(C), \phi_B(B') - \widehat{\mu}_{B|C}(C') \rangle_{\mathcal{H}_\mathcal{B}},$$

we obtain that $U = \widehat{\mathrm{KCI}}$, $\sigma^2_{\mathfrak{H}_1}$ is $\hat{\sigma}^2_{\mathfrak{H}_1}$, and SNR is $\widehat{\mathrm{SNR}}$.

Theorem D.1. *Let $h_\omega : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a set of functions for each $\omega \in \Omega$ such that:*

(i) The h_ω are uniformly bounded: $\sup_{\omega \in \Omega} \sup_{x, x' \in \mathcal{X}} |h_\omega(x, x')| \le \rho$ *for some* $1 \le \rho < \infty$.

(ii) Ω *is a subset of some* D*-dimensional Banach space, and* $\sup_{\omega\in\Omega}\|\omega\| \le R$.

(iii) The functions are Lipschitz in their parameterization: there is some $L < \infty$ *such that for all* $x, x' \in \mathcal{X}$ *and* $\omega, \omega' \in \Omega$, $|h_\omega(x,x') - h_{\omega'}(x,x')| \le L\|\omega - \omega'\|$.

Use $U_n^{(\omega)}$ *and similar to denote the quantities defined above with the function* h_ω. *Let* $\bar{\Omega}_s \subseteq \Omega$ *be a set of parameters for which* $\sigma_{\mathfrak{H}_1}^{(\omega)} \ge s$. *Take* $\lambda = \ell n^{-1/3}$. *Then, with probability at least* $1-\delta$,

$$\sup_{\omega\in\bar{\Omega}_s} |\mathrm{SNR}_{n,\lambda}^{(\omega)} - \mathrm{SNR}^{(\omega)}| \le \frac{\rho}{s^2 n^{1/3}} \left[\frac{\ell}{2s} + \left[\frac{448\rho}{\sqrt{\ell}} + \frac{2s}{n^{1/6}}\right]\left[L + \sqrt{2\log\frac{4}{\delta} + 2D\log(4R\sqrt{n})}\right] + \frac{72\rho^3}{\sqrt{\ell n}}\right].$$

Thus, treating ρ *and* ℓ *as constants, we have that*

$$\sup_{\omega\in\bar{\Omega}_s} |\mathrm{SNR}_{n,\lambda}^{(\omega)} - \mathrm{SNR}^{(\omega)}| = \mathcal{O}\left(\frac{1}{s^2 n^{1/3}}\left[\frac{1}{s} + \left(1 + \frac{s}{n^{1/6}}\right)\left[L + \sqrt{D\log(Rn) + \log\frac{1}{\delta}}\right]\right]\right).$$

This further implies that if $\mathrm{SNR}^{(\omega)}$ *has a unique maximizer* $\omega^* \in \bar{\Omega}_s$, *the sequence of empirical minimizers of* $\mathrm{SNR}_{n,\lambda=\ell n^{-\frac{1}{3}}}^{(\omega)}$ *converges in probability to* ω^*.

The assumptions in Theorem D.1 agree with those of Liu et al. (2020). Their Appendix A.4's bounds on L directly apply to the h of $\widehat{\mathrm{KCI}}$ if we only consider changing k_C, as we do in our experiments. These techniques could be readily adapted to changing other parameters, whether k_A or k_B (if the regressions are also updated appropriately) or parameters inside $\widehat{\mu}_{A|C}$ and $\widehat{\mu}_{B|C}$. We emphasize, however, that doing so only increases $\widehat{\mathrm{SNR}}$; any of these operations could increase the probability of rejecting the null under the alternative, but they will *also* increase the probability of rejecting the null under the null, further inflating Type I error.

Proof. Let $\sigma^2_{\mathfrak{H}_1,n,\lambda} = \sigma^2_{\mathfrak{H}_1,n} + \lambda$ and $\sigma^2_{\mathfrak{H}_1,\lambda} = \sigma^2_{\mathfrak{H}_1} + \lambda$. We begin with the decomposition

$$\begin{aligned}
\sup_{\omega\in\bar{\Omega}_s} |\mathrm{SNR}_{n,\lambda}^{(\omega)} - \mathrm{SNR}^{(\omega)}| &= \sup_{\omega\in\bar{\Omega}_s}\left|\frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda}} - \frac{U^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1}}\right| \\
&\le \sup_{\omega\in\bar{\Omega}_s}\left|\frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda}} - \frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda}}\right| + \sup_{\omega\in\bar{\Omega}_s}\left|\frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda}} - \frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1}}\right| + \sup_{\omega\in\bar{\Omega}_s}\left|\frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1}} - \frac{U^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1}}\right|.
\end{aligned}$$

Now notice that $|U_n^\omega| \le \rho$, $\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda} \ge \sqrt{s^2+\lambda} \ge s$, and $\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda} \ge \sqrt{\lambda}$. Hence the first term is

$$\begin{aligned}
\sup_{\omega\in\bar{\Omega}_s}\left|\frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda}} - \frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda}}\right| &= \sup_{\omega\in\bar{\Omega}_s}|U_n^{(\omega)}|\,\frac{1}{\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda}}\,\frac{1}{\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda}}\,\frac{|(\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda})^2 - (\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda})^2|}{(\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda})^2 + (\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda})^2} \\
&\le \frac{\rho}{\sqrt{\lambda}\sqrt{s^2+\lambda}(\sqrt{s^2+\lambda}+\sqrt{\lambda})}\sup_{\omega\in\bar{\Omega}_s}|(\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda})^2 - (\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda})^2| \\
&\le \frac{\rho}{s^2\sqrt{\lambda}}\sup_{\omega\in\bar{\Omega}_s}|(\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda})^2 - (\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda})^2|,
\end{aligned}$$

the second is

$$\sup_{\omega\in\bar{\Omega}_s}\left|\frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda}} - \frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1}}\right| = \sup_{\omega\in\bar{\Omega}_s}|U_n^{(\omega)}|\,\frac{1}{\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda}}\frac{1}{\sigma^{(\omega)}_{\mathfrak{H}_1}}\left|\frac{(\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda})^2 - (\sigma^{(\omega)}_{\mathfrak{H}_1})^2}{\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda} + \sigma^{(\omega)}_{\mathfrak{H}_1}}\right| \le \frac{\rho\lambda}{2s^3},$$

and the third is

$$\sup_{\omega\in\bar{\Omega}_s}\left|\frac{U_n^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1}} - \frac{U^{(\omega)}}{\sigma^{(\omega)}_{\mathfrak{H}_1}}\right| = \sup_{\omega\in\bar{\Omega}_s}\frac{1}{\sigma^{(\omega)}_{\mathfrak{H}_1}}\left|U_n^{(\omega)} - U^{(\omega)}\right| \le \frac{1}{s}\sup_{\omega\in\bar{\Omega}_s}\left|U_n^{(\omega)} - U^{(\omega)}\right|.$$

Thus we have reduced to needing uniform convergence of U_n and $\sigma^2_{\mathfrak{H}_1,n,\lambda}$.

Propositions 15 and 16 of Liu et al. (2020) show this, up to replacing their ν with our $\rho/4$, their R_Ω with our R, and their L_k with our $L/4$; this can be seen by inspecting the proofs. The results become

$$\Pr\left(\sup_{\omega\in\Omega}\left|U_n^{(\omega)} - U^{(\omega)}\right| \le \frac{2}{\sqrt{n}}\left[\rho\sqrt{2\log\frac{2}{\delta} + 2D\log(4R\sqrt{n})} + L\right]\right) \ge 1-\delta$$

$$\Pr\left(\sup_{\omega\in\Omega}\left|(\sigma^{(\omega)}_{\mathfrak{H}_1,n,\lambda})^2 - (\sigma^{(\omega)}_{\mathfrak{H}_1,\lambda})^2\right| \le \frac{64}{\sqrt{n}}\left[7\sqrt{2\log\frac{2}{\delta} + 2D\log(4R\sqrt{n})} + \frac{9\rho^2}{8\sqrt{n}} + \frac{1}{2}L\rho\right]\right) \ge 1-\delta.$$

Combining the results, it holds with probability at least $1-\delta$ that the worst-case error $\sup_{\omega\in\bar{\Omega}_s}|\mathrm{SNR}^{(\omega)}_{n,\lambda} - \mathrm{SNR}^{(\omega)}|$ is at most

$$\frac{\rho\lambda}{2s^3} + \left[\frac{2\rho}{s\sqrt{n}} + \frac{448\rho}{s^2\sqrt{\lambda n}}\right]\sqrt{2\log\frac{4}{\delta} + 2D\log(4R\sqrt{n})} + \left[\frac{2}{s\sqrt{n}} + \frac{32\rho^2}{s^2\sqrt{\lambda n}}\right]L + \frac{72\rho^3}{s^2 n\sqrt{\lambda}}.$$

Plugging in $\lambda = \ell n^{-\frac{1}{3}}$ yields

$$\frac{\rho\ell}{2s^3 n^{\frac{1}{3}}} + \left[\frac{2\rho}{s\sqrt{n}} + \frac{448\rho}{\sqrt{\ell}s^2 n^{\frac{1}{3}}}\right]\sqrt{2\log\frac{4}{\delta} + 2D\log(4R\sqrt{n})} + \left[\frac{2}{s\sqrt{n}} + \frac{32\rho^2}{\sqrt{\ell}s^2 n^{\frac{1}{3}}}\right]L + \frac{72\rho^3}{\sqrt{\ell}s^2 n^{\frac{5}{6}}}.$$

We can use our assumption $\rho \ge 1$ and that $448 > 32$ to get a slightly looser but simpler upper bound of

$$\frac{\rho\ell}{2s^3 n^{1/3}} + \left[\frac{2\rho}{s\sqrt{n}} + \frac{448\rho^2}{\sqrt{\ell}s^2 n^{1/3}}\right]\left[L + \sqrt{2\log\frac{4}{\delta} + 2D\log(4R\sqrt{n})}\right] + \frac{72\rho^3}{\sqrt{\ell}s^2 n^{5/6}},$$

which reduces to the result in the theorem statement.

The final result is a standard consequence of the prior statement, as in Corollary 12 of Liu et al. (2020). □

E U-Statistic moments for Hilbert space kernels

Theorem 6.1. *Let $h(X, X') = \langle\phi_h(X), \phi_h(X')\rangle$ with mean embedding $\mu_h = \mathbb{E}_X\,\phi_h(X)$ and the centered covariance operator $\mathfrak{C}_h = \mathbb{E}_X[\phi_h(X)\otimes\phi_h(X)] - \mu_h\otimes\mu_h$. Define $\nu_1 = \langle\mu_h, \mathfrak{C}_h\mu_h\rangle$ and $\nu_2 = \|\mathfrak{C}_h\|^2_{\mathrm{HS}}$. The corresponding U-statistic satisfies*

$$U_n = \frac{1}{n(n-1)}\sum_{1\le i\ne j\le n} h(X_i, X_j), \quad \mathbb{E}[U_n] = U = \|\mu_h\|^2, \quad \mathrm{Var}(U_n) = \frac{4}{n}\nu_1 + \frac{2}{n(n-1)}\nu_2.$$

Proof. U_n is the definition of a second-order U-statistic. We have that

$$\begin{aligned}\mathbb{E}\,U_n &= \mathop{\mathbb{E}}_{X,X'} h(X, X')\\ &= \mathop{\mathbb{E}}_{X,X'}\langle\phi_h(X), \phi_h(X')\rangle\\ &= \langle\mathop{\mathbb{E}}_{X}\phi_h(X), \mathop{\mathbb{E}}_{X'}\phi_h(X)\rangle = \langle\mu_h, \mu_h\rangle = \|\mu_h\|^2\end{aligned}$$

when μ_h exists in the Bochner sense, $\mathbb{E}\|\phi_h(X)\| < \infty$.

For the variance, it is a standard result that (e.g. Serfling, 1980, Section 5.2.1):

$$\mathrm{Var}(U_n) = \frac{4(n-2)}{n(n-1)}\mathop{\mathrm{Var}}_{X}\Big[\mathop{\mathbb{E}}_{X'|X}[\,h(X, X')\mid X\,]\Big] + \frac{2}{n(n-1)}\mathrm{Var}\Big[h(X, X')\Big].$$

Using the law of total variance,

$$\mathrm{Var}\left[h(X, X')\right] = \mathop{\mathrm{Var}}_{X}\Big[\mathop{\mathbb{E}}_{X'|X}[\,h(X, X')\mid X\,]\Big] + \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathrm{Var}}_{X'|X}[\,h(X, X')\mid X\,]\Big]$$

and so

$$\mathrm{Var}(U_n) = \frac{4n-6}{n(n-1)}\mathop{\mathrm{Var}}_{X}\Big[\mathop{\mathbb{E}}_{X'|X}[\,h(X, X')\mid X\,]\Big] + \frac{2}{n(n-1)}\mathop{\mathbb{E}}_{X}\Big[\mathop{\mathrm{Var}}_{X'|X}[\,h(X, X')\mid X\,]\Big].$$

We can now compute that

$$\begin{aligned}
\mathop{\mathbb{E}}_{X'|X}[h(X,X') \mid X] &= \langle \phi_h(X), \mu_h \rangle \\
\mathop{\mathrm{Var}}_{X}\Big[\mathop{\mathbb{E}}_{X'|X}[h(X,X') \mid X]\Big] &= \mathop{\mathbb{E}}_{X}\Big[\big(\mathop{\mathbb{E}}_{X'|X}[h(X,X') \mid X]\big)^2\Big] - \Big(\mathop{\mathbb{E}}_{X}\mathop{\mathbb{E}}_{X'|X}[h(X,X') \mid X]\Big)^2 \\
&= \mathop{\mathbb{E}}_{X}\langle \phi_h(X), \mu_h \rangle^2 - \langle \mu_h, \mu_h \rangle^2 \\
&= \mathop{\mathbb{E}}_{X}\langle \mu_h, \phi_h(X) \rangle \langle \phi_h(X), \mu_h \rangle - \langle \mu_h, \mu_h \rangle \langle \mu_h, \mu_h \rangle \\
&= \mathop{\mathbb{E}}_{X}\big\langle \mu_h, \big(\phi_h(X) \otimes \phi_h(X) - \mu_h \otimes \mu_h\big)\mu_h \big\rangle \\
&= \Big\langle \mu_h, \mathop{\mathbb{E}}_{X}\big[\phi_h(X) \otimes \phi_h(X) - \mu_h \otimes \mu_h\big]\mu_h \Big\rangle \\
&= \langle \mu_h, \mathfrak{C}_h \mu_h \rangle = \nu_1.
\end{aligned}$$

The remaining term is given by

$$\begin{aligned}
\mathop{\mathrm{Var}}_{X'|X}\big[h(X,X') \mid X\big] &= \mathop{\mathbb{E}}_{X'|X}\big[h(X,X')^2 \mid X\big] - \Big(\mathop{\mathbb{E}}_{X'|X}\big[h(X,X') \mid X\big]\Big)^2 \\
&= \mathop{\mathbb{E}}_{X'|X}\langle \phi_h(X), \phi_h(X') \rangle \langle \phi_h(X'), \phi_h(X) \rangle - \Big(\mathop{\mathbb{E}}_{X'|X}\langle \phi_h(X), \phi_h(X') \rangle\Big)^2 \\
&= \mathop{\mathbb{E}}_{X'|X}\langle \phi_h(X), \phi_h(X') \rangle \langle \phi_h(X'), \phi_h(X) \rangle - \langle \phi_h(X), \mu_h \rangle \langle \phi_h(X), \mu_h \rangle \\
&= \Big\langle \phi_h(X), \big(\mathop{\mathbb{E}}_{X'|X} \phi_h(X') \otimes \phi_h(X') - \mu_h \otimes \mu_h\big)\phi_h(X) \Big\rangle \\
&= \langle \phi_h(X), \mathfrak{C}_h \phi_h(X) \rangle \\
\mathop{\mathbb{E}}_{X}\mathop{\mathrm{Var}}_{X'|X}\big[h(X,X') \mid X\big] &= \mathop{\mathbb{E}}_{X}\langle \phi_h(X), \mathfrak{C}_h \phi_h(X) \rangle \\
&= \mathop{\mathbb{E}}_{X}\langle \phi_h(X) \otimes \phi_h(X), \mathfrak{C}_h \rangle_{\mathrm{HS}} \\
&= \Big\langle \mathop{\mathbb{E}}_{X} \phi_h(X) \otimes \phi_h(X), \mathfrak{C}_h \Big\rangle_{\mathrm{HS}} \\
&= \langle \mathfrak{C}_h + \mu_h \otimes \mu_h, \mathfrak{C}_h \rangle_{\mathrm{HS}} = \nu_1 + \nu_2.
\end{aligned}$$

Combining, we find that

$$\mathrm{Var}(U_n) = \frac{4n-6}{n(n-1)}\nu_1 + \frac{2}{n(n-1)}(\nu_1 + \nu_2) = \frac{4}{n}\nu_1 + \frac{2}{n(n-1)}\nu_2. \qquad \square$$

F Analytical example

F.1 With correct regressions

KCI as an expectation under a bivariate Gaussian. Under the assumption of linear kernels $\phi_A(a) = a$ and $\phi_B(b) = b$, the conditional cross-covariance operator with correct regressions can be written as:

$$\begin{aligned}
\mathfrak{C}_{AB|C} &= \mathop{\mathbb{E}}_{AB|C}\big[(A - \mu_{A|C}(C))(B - \mu_{B|C}(C)) \mid C\big] \\
&= \mathop{\mathbb{E}}_{AB|C}\big[(A - f_A(C))\,(B - f_B(C)) \mid C\,\big] \\
&= \tau^2 E_{AB|C}\big[r_A\, r_B \mid C\,\big] \\
&= \tau^2 \gamma(C)
\end{aligned}$$

Since $C, C' \sim \mathcal{N}(0,1)$ independently, and $k_C(c,c') = \exp\left(-\frac{(c-c')^2}{2\ell_C^2}\right)$, then the KCI statistic becomes:

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

Using the fact that C, C' are independent standard Gaussians, this becomes:

$$\begin{aligned}\mathrm{KCI} &= \tau^4 \iint \frac{1}{2\pi} \exp\left(-\frac{c^2 + c'^2}{2}\right) \exp\left(-\frac{(c-c')^2}{2\ell_C^2}\right) \gamma(c)\gamma(c')dcdc' \\ &= \tau^4 \iint \frac{1}{2\pi} \exp\left(-\frac{(\ell_C^2+1)c^2 - 2cc' + (\ell_C^2+1)c'^2}{2\ell_C^2}\right) \gamma(c)\gamma(c')dcdc',\end{aligned}$$

Define the vector $\mathbf{x} = \begin{bmatrix} c \\ c' \end{bmatrix}$, and write the integrand as a bivariate Gaussian density with covariance matrix $\mathbf{\Sigma}$. That is,

$$\mathrm{KCI} = \tau^4 \sqrt{\frac{\ell_C^2}{\ell_C^2 + 2}} \int_{\mathbb{R}^2} \phi_{\mathbf{\Sigma}}(c, c')\, \gamma(c)\, \gamma(c')\, dcdc',$$

where $\phi_{\mathbf{\Sigma}}$ denotes the bivariate normal density with zero mean and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \frac{\ell_C^2+1}{\ell_C^2+2} & \frac{1}{\ell_C^2+2} \\ \frac{1}{\ell_C^2+2} & \frac{\ell_C^2+1}{\ell_C^2+2} \end{bmatrix}, \quad \det(\mathbf{\Sigma}) = \frac{\ell_C^2}{\ell_C^2+2}.$$

We may thus express:

$$\mathrm{KCI} = \tau^4 \sqrt{\frac{\ell_C^2}{\ell_C^2 + 2}} \mathop{\mathbb{E}}_{(X,X')\sim\mathcal{N}_{\ell_C}} \big[\gamma(X)\gamma(X')\big],$$

with auxiliary variables $(X, X') \sim \mathcal{N}_{\ell_C} := \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} 1 - \frac{1}{\ell_C^2+2} & \frac{1}{\ell_C^2+2} \\ \frac{1}{\ell_C^2+2} & 1 - \frac{1}{\ell_C^2+2} \end{bmatrix}\right)$.

Exact expression for KCI We can analytically compute both the population KCI value and its variance to generate the theoretical curve shown in Figure 2. Under the alternative hypothesis, suppose the conditional dependence takes the form $\gamma(X) = \sin(\beta X)$. Then the KCI statistic becomes:

$$\begin{aligned}\mathrm{KCI} &= \tau^4 \sqrt{\frac{\ell_C^2}{\ell_C^2 + 2}} \mathop{\mathbb{E}}_{(X,X')\sim\mathcal{N}_{\ell_C}} \big[\sin(\beta X)\sin(\beta X')\big] \\ &= \frac{\tau^4}{2} \sqrt{\frac{\ell_C^2}{\ell_C^2 + 2}} \mathop{\mathbb{E}}_{(X,X')\sim\mathcal{N}_{\ell_C}} \big[\cos(\beta(X - X')) - \cos(\beta(X + X'))\big],\end{aligned}$$

Now note that $X - X'$ and $X + X'$ are linear functions of a jointly Gaussian vector and hence are Gaussian themselves. Since $(X, X') \sim \mathcal{N}(0, \mathbf{\Sigma})$, the random variables $Z_1 = X - X'$, $Z_2 = X + X'$ are zero-mean and have variances:

$$\mathrm{Var}(Z_1) = 2\left(1 - \frac{2}{\ell_C^2 + 2}\right), \qquad \mathrm{Var}(Z_2) = 2.$$

We now compute the expectations using the identity for the cosine of a Gaussian: $\mathbb{E}[\cos(\beta Z)] = \exp\left(-\frac{1}{2}\beta^2 \mathrm{Var}(Z)\right)$. Thus,

$$\begin{aligned}\mathbb{E}_{X,X'}\left[\cos\left(\beta(X - X')\right)\right] &= \exp\left(-\beta^2\left(1 - \frac{2}{\ell_C^2 + 2}\right)\right), \\ \mathbb{E}_{X,X'}\left[\cos\left(\beta(X + X')\right)\right] &= \exp\left(-\beta^2\right).\end{aligned}$$

Substituting into the expression for KCI, we obtain:

$$\begin{aligned}\mathrm{KCI} &= \frac{\tau^4}{2} \sqrt{\frac{\sigma_C}{\sigma_C + 2}} \left(\exp\left(-\beta^2\left(1 - \frac{2}{\sigma_C + 2}\right)\right) - \exp\left(-\beta^2\right)\right) \\ &= \frac{\tau^4}{2} \exp(-\beta^2) \sqrt{\frac{\sigma_C}{\sigma_C + 2}} \left(\exp\left(\frac{2\beta^2}{\sigma_C + 2}\right) - 1\right).\end{aligned}$$

Exact expression of variance of KCI_n. The variance of the U-statistic KCI_n can be decomposed into three components, as described in Appendix E. We now provide exact expressions for each term under the alternative hypothesis.

$$
\begin{aligned}
v_{\mathrm{c}} &\coloneqq \mathop{\mathbb{E}}_{ABC}\Big[\mathop{\mathbb{E}}_{A'B'C'}\left[h(ABC, A'B'C') \mid ABC\right]^2\Big] \\
&= \mathop{\mathbb{E}}_{ABC}\Big[\mathop{\mathbb{E}}_{A'B'C'}\left[\tau^4 k_C(C,C') r_A r_B r_A{}' r_B{}' \mid ABC\right]^2\Big] \\
&= \tau^8 \mathop{\mathbb{E}}_{C}\left[\mathop{\mathbb{E}}_{AB}[r_A{}^2 r_B{}^2 \mid C] \mathop{\mathbb{E}}_{C'}\left[k_C(C,C') \mathop{\mathbb{E}}_{A'B'|C'}[r_A{}' r_B{}' \mid C']\right]^2\right] \\
&= \tau^8 \mathop{\mathbb{E}}_{C}\Big[\big(1+2\gamma^2(\beta C)\big) \mathop{\mathbb{E}}_{C'}\left[k_C(C,C')\gamma(\beta C')\right]^2\Big] \\
&= \frac{\tau^8 \ell_C^2}{\sqrt{(\ell_C^2+1)(\ell_C^2+3)}} \mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,\frac{\ell_C^2+1}{\ell_C^2+3})}\left[\big(1+2\gamma^2(\beta X)\big) \mathop{\mathbb{E}}_{X'\sim\mathcal{N}(\frac{X}{\ell_C^2+1},\frac{\ell_C^2}{\ell_C^2+1})}[\gamma(\beta X')]^2\right] \\
&= \frac{\tau^8 \ell_C^2 \exp(-\frac{\beta^2\ell_C^2}{\ell_C^2+1})}{\sqrt{(\ell_C^2+1)(\ell_C^2+3)}}\Bigg(1 - \exp\Big(-\frac{2\beta^2}{(\ell_C^2+1)(\ell_C^2+3)}\Big) - \frac{1}{2}\exp\Big(-\frac{2\beta^2(\ell_C^2+1)}{\ell_C^2+3}\Big) \\
&\quad + \frac{1}{4}\exp\Big(-\frac{2\beta^2(\ell_C^2+2)^2}{(\ell_C^2+1)(\ell_C^2+3)}\Big) + \frac{1}{4}\exp\Big(-\frac{2\beta^2\ell_C^4}{(\ell_C^2+1)(\ell_C^2+3)}\Big)\Bigg).
\end{aligned}
$$

Besides,

$$
v_{\mathrm{m}} \coloneqq \mathop{\mathbb{E}}_{ABC,A'B'C'}\left[h(ABC, A'B'C')\right]^2 = \mathrm{KCI}^2.
$$

Also,

$$
\begin{aligned}
v_{\mathrm{s}} &\coloneqq \mathbb{E}\left[h^2(ABC, A'B'C')\right] \\
&= \tau^8 \mathbb{E}\left[k_C{}^2(C,C') r_A{}^2 r_B{}^2 r_A{}'^2 r_B{}'^2\right] \\
&= \tau^8 \mathop{\mathbb{E}}_{C,C'}\left[k_C{}^2(C,C') \mathop{\mathbb{E}}_{A,B}[r_A{}^2 r_B{}^2 \mid C] \mathop{\mathbb{E}}_{A',B'}[r_A{}'^2 r_B{}'^2 \mid C']\right] \\
&= \tau^8 \mathop{\mathbb{E}}_{C,C'}\left[\exp\Big(\frac{(C-C')^2}{\ell_C^2}\Big)(1+2\gamma^2(C))(1+2\gamma^2(C'))\right] \\
&= \tau^8 \sqrt{\frac{\ell_C^2}{\ell_C^2+4}}\Bigg(4 - 2\exp\Big(-\frac{2\beta^2(\ell_C^2+2)}{\ell_C^2+4}\Big) + \exp\Big(-\frac{2\beta^2\ell_C^2}{\ell_C^2+2}\Big) \\
&\quad \cdot\Big(-2\exp\Big(\frac{-8\beta^2}{(\ell_C^2+2)(\ell_C^2+4)}\Big) + \frac{1}{2}\exp\Big(\frac{-2\beta^2(\ell_C^2+4)}{\ell_C^2+2}\Big) + \frac{1}{2}\exp\Big(\frac{-2\beta^2\ell_C^4}{(\ell_C^2+2)(\ell_C^2+4)}\Big)\Big)\Bigg).
\end{aligned}
$$

Therefore, the variance can be obtained by combining those three terms together:

$$
\mathrm{Var}(\mathrm{KCI}_n) = \frac{(4n-8)v_{\mathrm{c}} - (4n-6)v_{\mathrm{m}} + 2v_{\mathrm{s}}}{n(n-1)}.
$$

F.2 With regression errors

Suppose the conditional mean embeddings have errors, that $\widehat{\mu}_{A|C} = \mu_{A|C} + \Delta_{A|C}$ and $\widehat{\mu}_{B|C} = \mu_{B|C} + \Delta_{B|C}$, where $\Delta_{A|C}$ and $\Delta_{B|C}$ denote the respective regression errors. Then the conditional cross-covariance operator becomes:

$$
\begin{aligned}
\widehat{\mathfrak{C}}_{AB|C}(C) &= \mathop{\mathbb{E}}_{AB|C}\left[(A - \widehat{\mu}_{A|C}(C))(B - \widehat{\mu}_{B|C}(C)) \mid C\right] \\
&= \mathop{\mathbb{E}}_{AB|C}\left[(A - f_A(C) - \Delta_{A|C}(C))\,(B - f_B(C) - \Delta_{B|C}(C)) \mid C\right] \\
&= E_{AB|C}\left[(\tau r_A - \Delta_{A|C}(C))\,(\tau r_B - \Delta_{B|C}(C)) \mid C\right] \\
&= \tau^2\gamma(C) + \Delta_{A|C}(C)\Delta_{B|C}(C).
\end{aligned}
$$

The final equality follows from the assumption that the regression estimates are obtained using an independent training set, and are thus independent of the test-time noise in r_A and r_B. Consequently, the cross terms involving $r_A\Delta_{B|C}(C)$ and $r_B\Delta_{A|C}(C)$ have zero conditional expectation.

Thus, under the null hypothesis ($A \perp\!\!\!\perp B \mid C$), the KCI with noisy conditional means becomes

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

Under the null hypothesis, the variance of the U-statistic $\widehat{\mathrm{KCI}}_n$ includes the following three components:

$$\begin{aligned}
v_{\mathrm{c}} &= \mathop{\mathbb{E}}_{ABC}\left[\mathop{\mathbb{E}}_{A'B'C'}\left[k_C(C,C')(\tau r_A - \Delta_{A|C}(C))(\tau r_B - \Delta_{B|C}(C))(\tau {r_A}' - \Delta_{A|C}(C'))(\tau {r_B}' - \Delta_{B|C}(C'))\right]^2\right] \\
&= \mathop{\mathbb{E}}_{ABC}\left[(\tau r_A - \Delta_{A|C}(C))^2(\tau r_B - \Delta_{B|C}(C))^2 \mathop{\mathbb{E}}_{C'}\left[k_C(C,C')\Delta_{A|C}(C')\Delta_{B|C}(C')\right]^2\right] \\
&= \mathop{\mathbb{E}}_{C}\left[(\tau^2 + {\Delta_{A|C}}^2(C))(\tau^2 + {\Delta_{B|C}}^2(C)) \mathop{\mathbb{E}}_{C'}\left[k_C(C,C')\Delta_{A|C}(C')\Delta_{B|C}(C')\right]^2\right].
\end{aligned}$$

$$v_{\mathrm{m}} = \mathop{\mathbb{E}}_{C}\left[k_C(C,C')\Delta_{A|C}(C)\Delta_{B|C}(C)\Delta_{A|C}(C')\Delta_{B|C}(C')\right]^2.$$

$$\begin{aligned}
v_{\mathrm{s}} &= \mathbb{E}\left[{k_C}^2(C,C')(\tau r_A - \Delta_{A|C}(C))^2\,(\tau r_B - \Delta_{B|C}(C))^2(\tau {r_A}' - \Delta_{A|C}(C'))^2\,(\tau {r_B}' - \Delta_{B|C}(C'))^2\right] \\
&= \mathbb{E}\left[{k_C}^2(C,C')(\tau^2 + {\Delta_{A|C}}^2(C))(\tau^2 + {\Delta_{B|C}}^2(C))(\tau^2 + {\Delta_{A|C}}^2(C'))(\tau^2 + {\Delta_{B|C}}^2(C'))\right].
\end{aligned}$$

F.3 Complex conditional dependence scenario (3-dimensional C)

We now extend our motivating example to a more complex setting by considering a three-dimensional conditioning variable $C = (C_1, C_2, C_3)$. Specifically, we define random variables:

$$C \sim \mathcal{N}(0, I_3), \quad A = f_A({\boldsymbol{e}_A}^\top C) + \tau\, r_A, \quad B = f_B({\boldsymbol{e}_B}^\top C) + \tau\, r_B,$$

where $\boldsymbol{e}_A$ and $\boldsymbol{e}_B$ are indicator vectors selecting specific dimensions of (one entry equals 1, others 0), determining which dimension influences each variable. The additive noise terms (r_A, r_B) are conditionally dependent via on $\gamma({\boldsymbol{e}_C}^\top C)$, as previously defined.

We employ a generalized Gaussian kernel for C with dimension-specific bandwidths ℓ_{Ci}^2, as commonly implemented in libraries such as `sklearn`: $k_C(C,C') = \exp\left(-\sum_{i=1}^3 \frac{(C_i - C_i')^2}{2\ell_{Ci}^2}\right)$. We assume regression models are trained specifically on the relevant dimensions (as selected by $\boldsymbol{e}_A$ and $\boldsymbol{e}_B$), effectively ignoring irrelevant or noisy dimensions. Thus, $\Delta_{A|C}({\boldsymbol{e}_A}^\top C) = \widehat{\mu}_{A|C}({\boldsymbol{e}_A}^\top C) - \mu_{A|C}({\boldsymbol{e}_A}^\top C)$ and $\Delta_{B|C}({\boldsymbol{e}_B}^\top C) = \widehat{\mu}_{B|C}({\boldsymbol{e}_B}^\top C) - \mu_{B|C}({\boldsymbol{e}_B}^\top C)$ depend only on the dimensions directly influencing A and B. The noisy KCI statistic $\widehat{\mathrm{KCI}}$, using linear kernels for A and B, becomes:

$$\begin{aligned}
&\widehat{\mathrm{KCI}} \\
&= \mathbb{E}\left[k_C(C,C')\big(\tau^2\gamma({\boldsymbol{e}_C}^\top C) + \xi({\boldsymbol{e}_A}^\top C, {\boldsymbol{e}_B}^\top C)\big)\big(\tau^2\gamma({\boldsymbol{e}_C}^\top C') + \xi({\boldsymbol{e}_A}^\top C', {\boldsymbol{e}_B}^\top C')\big)\right] \\
&= \left(\prod_{i=1}^3 \sqrt{\frac{\ell_{Ci}^2}{\ell_{Ci}^2 + 2}}\right)\mathbb{E}\left[\tau^4\gamma(X_C)\gamma(X_C') + \xi(X_A, X_B)\gamma(X_C') + \xi(X_A', X_B')\gamma(X_C) + \xi(X_A, X_B)\xi(X_A', X_B')\right]
\end{aligned}$$

where $X_C = {\boldsymbol{e}_C}^\top X$, $\xi(X_A, X_B) = \Delta_{A|C}({\boldsymbol{e}_A}^\top X)\Delta_{B|C}({\boldsymbol{e}_B}^\top X)$. Similar to Appendix F.1, X is an auxiliary variable, and for $i = 1, 2, 3$, we have $X_i, X_i' \sim \mathcal{N}_{\ell_{Ci}}$.

Kernel bandwidth selection and regression errors critically influence test performance. We discuss two illustrative scenarios:

Scenario 1: Shared-dimension dependence ($\boldsymbol{e}_A = \boldsymbol{e}_B = \boldsymbol{e}_C$): . When A, B, and their conditional dependence all rely on the same coordinate (e.g., C_1). The KCI is

$$\widehat{\mathrm{KCI}} = \left(\prod_{i=1}^3 \sqrt{\frac{\ell_{Ci}^2}{\ell_{Ci}^2 + 2}}\right)\mathbb{E}\left[\tau^4\gamma(X_1)\gamma(X_1') + \xi(X_1, X_1)\gamma(X_1') + \xi(X_1', X_1')\gamma(X_1) + \xi(X_1, X_1)\xi(X_1', X_1')\right].$$

Specifically, under the null hypothesis, the regression error will leak "dependence" into the test,

$$\widehat{\mathrm{KCI}} = \left(\prod_{i=1}^{3}\sqrt{\frac{\ell_{Ci}^2}{\ell_{Ci}^2+2}}\right)\mathbb{E}\left[\xi(X_1,X_1)\xi(X_1',X_1')\right].$$

Explicitly, this is:

$$\widehat{\mathrm{KCI}} = \left(\prod_{i=1}^{3}\sqrt{\frac{\ell_{Ci}^2}{\ell_{Ci}^2+2}}\right)\mathbb{E}\left[\Delta_{A|C}(X_1)\Delta_{A|C}(X_1')\ \Delta_{B|C}(X_1)\Delta_{B|C}(X_1')\right].$$

Scenario 2: Independent-dimension dependence (distinct e_A, e_B, e_C) . When A, B and their conditional dependence each utilize distinct coordinates (e.g., C_1, C_2, C_3 respectively), because of the independence between (X_A, X_B) and X_C, the KCI becomes:

$$\widehat{\mathrm{KCI}} = \left(\prod_{i=1}^{3}\sqrt{\frac{\ell_{Ci}^2}{\ell_{Ci}^2+2}}\right)\mathbb{E}\left[\tau^4\gamma(X_3)\gamma(X_3') + \xi(X_1,X_2)\xi(X_1',X_2').\right].$$

where we can further decompose the noise

$$\mathbb{E}[\xi(X_2,X_3)\xi(X_2',X_3')] = \mathbb{E}\left[\Delta_{A|C}(X_2)\Delta_{A|C}(X_2')\right]\ \mathbb{E}\left[\Delta_{B|C}(X_3)\Delta_{B|C}(X_3')\right]$$

Under the null hypothesis, the KCI becomes

$$\widehat{\mathrm{KCI}} = \left(\prod_{i=1}^{3}\sqrt{\frac{\ell_{Ci}^2}{\ell_{Ci}^2+2}}\right)\mathbb{E}\left[\Delta_{A|C}(X_2)\Delta_{A|C}(X_2')\right]\ \mathbb{E}\left[\Delta_{B|C}(X_3)\Delta_{B|C}(X_3')\right]$$

Experiments. Table 1 summarizes empirical results from synthetic experiments. Both scenarios used $f_A = \cos$, $f_B = \exp$, noise scale $\tau = 0.1$, dependence frequency $\beta = 2$, with 200 training points for regression and 200 test points, averaged over 100 runs. The significance level is set at $\alpha = 0.05$. We used Gaussian kernels for all kernels. Kernel ridge regression with leave-one-out validation was used to select kernels for $k_{C\to A}$ and $k_{C\to B}$, while kernel bandwidth for condition kernel k_C was selected by maximizing test power.

G Type-I bound proofs

G.1 Moment-matched normal against a threshold

Theorem 6.2. *Assume that $A \perp\!\!\!\perp B \mid C$. Let $Z_1 = \widehat{\mathrm{KCI}}_n$ and $Z_2 \sim \mathcal{N}\left(\widehat{\mathrm{KCI}}, \mathrm{Var}(\widehat{\mathrm{KCI}}_n)\right)$ be a normal variable moment-matched to Z_1. Let $q > 0$ and $\rho \in (0,1)$; define $T_1 = \sqrt{(1-\rho)/\rho}$ and $T_2 = \Phi^{-1}(1-\rho)$, where Φ is the standard normal CDF. Then the following holds for $i \in \{1,2\}$:*

$$\Pr\left(Z_i > \frac{q}{n}\right) \le \rho \quad \text{if } q \ge n\widehat{\mathrm{KCI}} + T_i\sqrt{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}.$$

Proof. Notice that, when either bound is satisfied, we have that

$$\left(\frac{q}{n} - \widehat{\mathrm{KCI}}\right)/\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)} \ge T_i.$$

The result for $\widehat{\mathrm{KCI}}_n$ follows by Cantelli's inequality, which slightly improves the better-known Chebyshev inequality for one-sided bounds; it says that for any random variable X,

$$\Pr(X \ge \mathbb{E}[X] + \lambda) \le \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \lambda^2}$$

and so, equivalently,

$$\Pr\left(\frac{X - \mathbb{E}\,X}{\sqrt{\mathrm{Var}(X)}} \ge t\right) \le \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + t^2\,\mathrm{Var}(X)} = \frac{1}{1+t^2}.$$

Plugging in T_1 yields that

$$\Pr\left(\widehat{\mathrm{KCI}}_n \geq \frac{q}{n}\right) \leq \Pr\left(\frac{\widehat{\mathrm{KCI}}_n - \widehat{\mathrm{KCI}}}{\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}} > \sqrt{\frac{1-\rho}{\rho}}\right) \leq \frac{1}{1+\frac{1-\rho}{\rho}} = \rho,$$

as desired. The bound for Z_2 is similar:

$$\Pr\left(Z_2 \geq \frac{q}{n}\right) \leq \Pr\left(\frac{Z_2 - \mathbb{E}\, Z_2}{\sqrt{\mathrm{Var}(Z_2)}} \geq \Phi^{-1}(1-\rho)\right) = 1 - \Phi\left(\Phi^{-1}(1-\rho)\right) = \rho. \qquad \square$$

G.2 Alignment to wild bootstrap

We provide a bound on the distance between two null distributions used in testing:

1. Wild bootstrap distribution given the test dataset.
2. Normal approximation to the test statistic $n\widehat{\mathrm{KCI}}_n$ when regression errors are present.

Setup. Let $\widehat{H} \in \mathbb{R}^{n\times n}$ be the kernel matrix with noisy regression under the null hypothesis, with entries $\widehat{H}_{ij} = \hat{h}_{ij}$. We define a random variable

$$Y \;:=\; \frac{1}{n} \sum_{1\leq i\neq j\leq n} \hat{h}_{ij}\, \varepsilon_i\, \varepsilon_j,$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(0,1)$ variables. It is known from the results of Imhof (1961) that the wild bootstrap distribution of $Y \mid \widehat{H}$ is the same as

$$(Y \mid \widehat{H}) \equiv Q \;:=\; \sum_{r=1}^{n} \lambda_r\, (X_r^2 - 1),$$

where $X_r \sim \mathcal{N}(0,1)$ i.i.d., and $\{\lambda_r\}_{r=1}^n$ are the eigenvalues of $\widehat{H}/n$. This centered form (X_r^2-1) ensures that $\mathbb{E}[Q] = 0$. The variance is $\mathrm{Var}(Q) = 2\sum_{r=1}^n \lambda_r^2 = \frac{2}{n^2}\,\mathrm{tr}(\widehat{H}^2)$. And the third central moment of Q is $8\sum_{r=1}^n \lambda_r^3 = \frac{8}{n^3}\mathrm{tr}(\widehat{H}^3)$ (see Buckley and Eagleson, 1988). Moreover, in the limit $n\to\infty$, Q and $n\mathrm{KCI}_n$ under a "perfect regression" null converge to the same distribution (see Leucht and Neumann 2013, Theorem 2.1 and Pogodin et al. 2024, Theorem 4).

When regression errors are present, the errors include a small but nonzero leading variance term, and thus the null distribution of $\widehat{\mathrm{KCI}}_n$ becomes slightly *non-degenerate*. By a suitable central limit theorem argument (analogous to Theorem 5.1), $\widehat{\mathrm{KCI}}_n$ is approximately normal for large n:

$$\frac{\left(\widehat{\mathrm{KCI}}_n - \widehat{\mathrm{KCI}}\right)}{\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}} \;\xrightarrow{d}\; \mathcal{N}(0,1).$$

Recall that for a second-order U-statistic with kernel $\hat{h}_{ij} = \hat{h}(X_i, X_j)$, a standard formula (see Serfling (1980)) gives

$$\begin{aligned}\mathrm{Var}(\widehat{\mathrm{KCI}}_n) \;&=\; \frac{4(n-2)}{n(n-1)}\, \mathop{\mathrm{Var}}_i\big[\mathop{\mathbb{E}}_j(\hat{h}_{ij})\big] \;+\; \frac{2}{n(n-1)}\, \mathrm{Var}\big[\hat{h}_{ij}\big] \\ &=\; \frac{4(n-2)}{n(n-1)}\, \mathop{\mathrm{Var}}_i\big[\mathop{\mathbb{E}}_j(\hat{h}_{ij})\big] \;+\; \frac{2}{n(n-1)}\, \mathbb{E}\big[\hat{h}_{ij}^2\big] \;-\; \frac{2}{n(n-1)}\, \big(\mathbb{E}[\hat{h}_{ij}]\big)^2.\end{aligned}$$

Meanwhile, for the wild-bootstrap statistic, we have

$$\mathrm{Var}(Y \mid \widehat{H}) = \mathrm{Var}(Q) \;=\; 2\sum_{r=1}^{n} \lambda_r^2 \;=\; \frac{2}{n^2}\,\mathrm{tr}(\widehat{H}^2) \;\longrightarrow\; 2\,\mathbb{E}\big[\hat{h}_{ij}^2\big] \quad \text{as } n\to\infty.$$

If the test uses correct regressions, under the null, $\mathbb{E}[h_{ij}] \;=\; \mathrm{Var}_i\big[\mathbb{E}_j(h_{ij})\big] \;=\; 0$, then $\mathrm{Var}(Q)$ converges exactly to $n^2\ \mathrm{Var}(\widehat{\mathrm{KCI}}_n)$. If with regression errors, note in $n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)$ that the factor

$4\,n\ \mathrm{Var}_i[\mathbb{E}_j(\hat{h}_{ij})]$ can remain substantial if $\mathrm{Var}_i[\mathbb{E}_j(\hat{h}_{ij})]$ does not vanish. This $\mathrm{Var}_i[\mathbb{E}_j(\hat{h}_{ij})]$ term can contribute a larger leading order when multiplied by n. Hence, if $\mathrm{Var}_i[\mathbb{E}_j(\hat{h}_{ij})]$ is non-negligible, for large n, $n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)$ can be bigger than $\mathrm{Var}(Q)$.

In practice, we use wild bootstrap to sample from the distribution of $Y \mid \widehat{H}$ under the noisy-regression null to determine a test threshold. Meanwhile, the actual test statistic $n\,\widehat{\mathrm{KCI}}_n$ can be approximated by a normal variable

$$S \equiv n\,Z_n \sim \mathcal{N}\left(n\widehat{\mathrm{KCI}}, n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)\right).$$

Hence, we want to quantify the distance between the distribution of $Y \mid \widehat{H}$ and the distribution of nZ_n. Concretely, we measure

$$\sup_{x\in\mathbb{R}}\Big|\Pr\big(Y \mid \widehat{H} \le x\big)\ -\ \Pr\big(n\,Z_n \le x\big)\Big|. \tag{12}$$

A small supremum indicates that $Y \mid \widehat{H}$ (wild bootstrap) and nZ_n (normal approximation under regression error) produce nearly identical thresholds, while a large value implies a more significant discrepancy between the two distributions.

Theorem 6.3. *Assume $A \perp\!\!\!\perp B \mid C$, and let $Y = \frac{1}{n}\sum_{i,j=1}^{n}\hat{h}_{ij}\,\varepsilon_i\,\varepsilon_j$, where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Let $\widehat{H}$ be the matrix with entries $\hat{h}_{i,j}$; assume $\|\widehat{H}\|_F^2 > 2\|\widehat{H}\|_{\text{op}}^2$. Let $Z_n \sim \mathcal{N}\Big(\widehat{\mathrm{KCI}},\ \mathrm{Var}(\widehat{\mathrm{KCI}}_n)\Big)$ be the moment-matched normal approximation to $\widehat{\mathrm{KCI}}_n$. Then, there exist constants $R_{i,n} \to R_i > 0$ for $i \in \{1,2,3\}$ (depending on the distribution of Y and $\widehat{\mathrm{KCI}}_n$) such that*

$$\sup_{x\in\mathbb{R}}\Big|\Pr(Y \mid \widehat{H} \le x) - \Pr(n\,Z_n \le x)\Big| \ \le\ R_{1,n} + R_{2,n}\,\frac{\widehat{\mathrm{KCI}}}{\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}} + R_{3,n}\,\Big|\frac{\mathrm{Var}(Y \mid \widehat{H})}{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)} - 1\Big|.$$

Proof. The overarching goal is to bound $\sup_{x\in\mathbb{R}}\big|\Pr(Q \le x)\ -\ \Pr(S \le x)\big|$, where Q is a centered weighted sum of chi-squared variables (which, as noted above, is exactly the distribution of $Y \mid H$), and $S \equiv nZ_n$ is a normal approximation to a U-statistic-based test statistic. The classical approach (Buckley and Eagleson, 1988; J.-T. Zhang, 2005) utilizes characteristic functions ($\psi(\cdot)$) and the Fourier inversion formula to control the Kolmogorov distance between distributions.

Let T be a generic random variable with characteristic function $\psi_T(t) = \mathbb{E}[e^{itT}]$. If $\log\big(\psi_T(t)\big)$ admits the power series expansion

$$\log(\psi_T(t))\ =\ \sum_{l=1}^{\infty}\mathcal{K}_l(T)\ \frac{(it)^l}{l!},$$

then the constants $\mathcal{K}_\ell(T)$ for $\ell = 1, 2, \dots$) are the cumulants of T (Muirhead, 2009, Sec. 2.4). In particular: $\mathcal{K}_1(T) = \mathbb{E}[T]$ is the mean, $\mathcal{K}_2(T) = \mathrm{Var}(T)$ is the variance, $\mathcal{K}_3(T) = \mathbb{E}[(T - \mathbb{E}[T])^3]$ is related to skewness.

Recall that $Q\ =\ \sum_{r=1}^{n}\lambda_r\left(Z_r^2 - 1\right)$, where $Z_r \sim \mathcal{N}(0,1)$ i.i.d., and $\{\lambda_r\}$ are positive (eigenvalues of H/n). By construction,

$$\mathcal{K}_1(Q) = 0, \quad \mathcal{K}_2(Q) = 2\sum_{r=1}^{n}\lambda_r^2, \quad \mathcal{K}_3(Q) = 8\sum_{r=1}^{n}\lambda_r^3, \quad \mathcal{K}_l(Q) = 2^{l-1}\,(l-1)!\,\sum_{r=1}^{n}\lambda_r^l \quad (l \ge 3).$$

Define the normalized version Q^* by

$$Q^*\ =\ \frac{Q - \mathbb{E}[Q]}{\sqrt{\mathrm{Var}(Q)}}\ =\ \frac{Q}{\sqrt{\mathcal{K}_2(Q)}}.$$

Hence, $\mathcal{K}_1(Q^*) = 0$, $\mathcal{K}_2(Q^*) = 1$, $\mathcal{K}_3(Q^*) = 8\sum_{r=1}^{n}\lambda_r^3/\mathcal{K}_2^{3/2}(Q)$, and for $l \ge 3$,

$$\mathcal{K}_\ell(Q^*)\ =\ \frac{\mathcal{K}_\ell(Q)}{(\mathcal{K}_2(Q))^{\ell/2}}.$$

For ease of comparison with Q, define

$$S^* \;=\; \frac{S}{\sqrt{\mathcal{K}_2(Q)}} \;\sim\; \mathcal{N}\Big(\frac{n\,\widehat{\mathrm{KCI}}}{\sqrt{\mathcal{K}_2(Q)}},\; \frac{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)}\Big).$$

The distance $\sup_x |\Pr(Q \le x) - \Pr(S \le x)|$ is equivalent to comparing Q^* and S^*: $\sup_{x\in\mathbb{R}}\big|\Pr(Q^* \le x) - \Pr(S^* \le x)\big|$.

By results of Esseen (1945, page 33), we have

$$\sup_{x\in\mathbb{R}}\big|\Pr(Q^* \le x) - \Pr(S^* \le x)\big| \;\le\; \frac{1}{2\pi}\int_{-\infty}^{+\infty}\left|\frac{\psi_{Q^*}(t) - \psi_{S^*}(t)}{t}\right| dt,$$

where ψ_{Q^*} and ψ_{S^*} are the characteristic functions of Q^* and S^*, respectively.

$$\psi_{Q^*}(t) = \prod_{r=1}^{n}\exp\left(-it\frac{\lambda_r}{\mathcal{K}_2^{1/2}(Q)}\right)\cdot\left(1 - it\frac{2\lambda_r}{\mathcal{K}_2^{1/2}(Q)}\right)^{-1/2}$$
$$\psi_{S^*}(t) = \exp\left(it\frac{n\widehat{\mathrm{KCI}}}{\mathcal{K}_2^{1/2}(Q)} - t^2\frac{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\mathcal{K}_2(Q)}\right).$$

To handle the integral

$$\int_{-\infty}^{\infty}\left|\frac{\psi_{Q^*}(t) - \psi_{S^*}(t)}{t}\right| dt,$$

it is standard to split the domain at $|t| = A$ for some positive A. Define

$$I_1 \;=\; \int_{|t|\le A}\left|\frac{\psi_{Q^*}(t) - \psi_{S^*}(t)}{t}\right| dt, \quad I_2 \;=\; \int_{|t|>A}\left|\frac{\psi_{S^*}(t)}{t}\right| dt, \quad I_3 \;=\; \int_{|t|>A}\left|\frac{\psi_{Q^*}(t)}{t}\right| dt.$$

Then,

$$\int_{-\infty}^{\infty}\left|\frac{\psi_{Q^*}(t) - \psi_{S^*}(t)}{t}\right| dt \;\le\; I_1 + I_2 + I_3.$$

Optimizing over A balances these different regions. This is a classical technique in Fourier-based proofs of Berry–Esseen-type inequalities.

Bounding I_1. We decompose I_1 based on the characteristic function ratio. Define

$$r(t) \;\coloneqq\; \log\big(\psi_{Q^*}(t)\big) \;-\; \log\big(\psi_{S^*}(t)\big).$$

Then

$$\begin{aligned} I_1 &= \int_{|t|\le A}|\psi_{S^*}(t)|\left|\frac{\psi_{Q^*}(t)/\psi_{S^*}(t) - 1}{t}\right| dt \\ &= \int_{|t|\le A}|\psi_{S^*}(t)|\left|\frac{\exp(r(t)) - 1}{t}\right| dt \\ &\le \int_{|t|\le A}|\psi_{S^*}(t)|\frac{|r(t)|\exp(|r(t)|)}{|t|}dt, \end{aligned}$$

where the last step comes from the inequality $|\exp(z) - 1| \le |z|\exp(|z|)$.

We use the following expansion bound for real θ, which be easily verified using the mean-value theorem (see also Buckley and Eagleson, 1988; J.-T. Zhang, 2005):

$$\left|\log(1 + i\theta) - \left\{i\theta + \frac{\theta^2}{2}\right\}\right| \le |\theta|^3/3. \tag{13}$$

Concretely,

$$r(t) = \Big(i\,t\frac{\sum_{r=1}^{n}\lambda_r}{\mathcal{K}_2(Q)^{1/2}}\Big) + \frac{1}{2}\sum_{r=1}^{n}\log\Big(1 - i\,t\,\frac{2\,\lambda_r}{\mathcal{K}_2(Q)^{1/2}}\Big) \;+\; \Big(i\,t\,\frac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2^{1/2}(Q)} \;-\; t^2\,\frac{n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big).$$

By bounding each $\log(\cdot)$ via the expansion (13), we obtain

$$|r(t)| \;\le\; \frac{1}{6}\,\frac{\sum_{r=1}^{n}|2\,t\,\lambda_r|^3}{\mathcal{K}_2^{3/2}(Q)} \;+\; \Big|i\,t\,\frac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2^{1/2}(Q)}\Big| \;+\; \Big|t^2\,\Big(\frac{\sum_{r=1}^{n}\lambda_r^2}{\mathcal{K}_2(Q)} \;-\; \frac{n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big)\Big|.$$

Recognizing $\mathcal{K}_3(Q^*) = 8\,\sum_r \lambda_r^3/(\mathcal{K}_2(Q))^{3/2}$ and $\sum_r \lambda_r^2 = \frac{1}{2}\,\mathcal{K}_2(Q)$, we rewrite:

$$|r(t)| \;\le\; \frac{1}{6}\,|t|^3\,\mathcal{K}_3(Q^*) \;+\; |t|\,\frac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2^{1/2}(Q)} \;+\; \frac{t^2}{2}\,\Big|\frac{n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)} - 1\Big|.$$

Hence, for $|t| \le A$,

$$I_1 \;\le\; \exp\big(|r(A)|\big)\int_{|t|\le A}\exp\Big(-\tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big)\Bigg(\frac{1}{6}\,t^2\,\mathcal{K}_3(Q^*) + \frac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2(Q)^{1/2}} \;+\; \frac{|t|}{2}\,\Big|\frac{n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)} - 1\Big|\Bigg)dt.$$

This splits naturally into three integrals:

$$\begin{aligned} I_1 \;\le\; & \frac{1}{6}\,\exp\big(|r(A)|\big)\,\mathcal{K}_3(Q^*)\int_{|t|\le A} t^2\,\exp\Big(-\tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big)\,dt \\ & +\; \frac{1}{2}\,\exp\big(|r(A)|\big)\,\Big|\frac{n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)} - 1\Big|\int_{|t|\le A}|t|\,\exp\Big(-\tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big)\,dt \\ & +\; \exp\big(|r(A)|\big)\,\frac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2(Q)^{1/2}}\int_{|t|\le A}\exp\Big(-\tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big)\,dt. \end{aligned}$$

In each term, the integral is bounded by Gaussian-like tail and one can get explicit numerical constants. The factor $\exp(|r(A)|)$ can then be controlled by choosing A suitably in combination with the expansions for $|r(t)|$.

Boudning I_2. Recall

$$I_2 \;=\; \int_{|t|>A}\Big|\frac{\psi_{S^*}(t)}{t}\Big|\,dt,$$

where

$$\psi_{S^*}(t) = \exp\Big(i\,t\,\tfrac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2(Q)^{1/2}} - \tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big).$$

Since $\big|\psi_{S^*}(t)\big| = \exp\big(-\tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\big)$, we have

$$I_2 = \int_{|t|>A}\frac{1}{|t|}\,\exp\Big(-\tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big)\,dt.$$

Next, use the fact that $|t| \ge A$ implies $\frac{1}{|t|} \le \frac{t^2}{A^3}$ for $|t| > 0$. Hence,

$$I_2 \;\le\; \int_{|t|>A}\frac{t^2}{A^3}\,\exp\Big(-\tfrac{t^2\,n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{2\,\mathcal{K}_2(Q)}\Big)\,dt.$$

Combining I_1 and I_2. We select A' such that $A'^{-3} \ge \frac{\exp\left(|r(A')|\right)\,\mathcal{K}_3(Q^*)}{6}$, ensuring we can combine the first term of I_1 and I_2. Let

$$\eta \;=\; \frac{n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)}$$

for shorthand, we derive

$$\begin{aligned} I_1 + I_2 \;\le\; & \frac{2}{A'^3}\int_0^\infty t^2\,\exp\Big(-\tfrac{\eta\,t^2}{2}\Big)\,dt \;+\; \frac{6}{A'^3\,\mathcal{K}_3(Q^*)}\Big|\frac{n^2\;\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)} - 1\Big|\int_0^\infty t\,\exp\Big(-\tfrac{\eta\,t^2}{2}\Big)\,dt \\ & +\; \frac{12}{A'^3\,\mathcal{K}_3(Q^*)}\,\frac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2(Q)^{1/2}}\int_0^\infty\exp\Big(-\tfrac{\eta\,t^2}{2}\Big)\,dt. \end{aligned}$$

Then, from the Gaussian integral identity

$$\int_0^\infty t^k \, \exp\!\big(-\tfrac{\eta t^2}{2}\big)\, dt \;=\; \text{(finite constants scaling with } \eta^{-\frac{k+1}{2}}\text{)},$$

evaluate the integrals yields the following bound:

$$I_1 + I_2 \;\le\; \frac{6\sqrt{2\pi}}{A'^3\,\mathcal{K}_3(Q^*)}\,\frac{n\,\widehat{\mathrm{KCI}}}{\mathcal{K}_2^{1/2}(Q)}\eta^{-1/2} \;+\; \frac{6}{A'^3\,\mathcal{K}_3(Q^*)}\left|\frac{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)} - 1\right|\eta^{-1} \;+\; \frac{\sqrt{2\pi}}{A'^3}\eta^{-3/2}.$$

Substitute the value of η, then

$$\begin{aligned} I_1 + I_2 \;\le\; & \frac{6\sqrt{2\pi}}{A'^3\,\mathcal{K}_3(Q^*)}\,\frac{\widehat{\mathrm{KCI}}}{\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}} \;+\; \frac{6}{A'^3\,\mathcal{K}_3(Q^*)}\left|\frac{\mathcal{K}_2(Q)}{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)} - 1\right| \\ & +\; \frac{\sqrt{2\pi}}{A'^3}\left(\frac{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}{\mathcal{K}_2(Q)}\right)^{-3/2}. \end{aligned}$$

Bounding I_3. Recall

$$I_3 = \int_{|t|>A}\left|\tfrac{\psi_{Q^*}(t)}{t}\right| dt.$$

We apply a refined argument from Buckley and Eagleson (1988) and J.-T. Zhang (2005) regarding $|\psi_{Q^*}(t)|$. Assuming the Frobenius norm of the kernel matrix dominates the operator norm, $\|\widehat{H}\|_F^2 > 2\|\widehat{H}\|_{\mathrm{op}}^2$, which implies $\delta \;=\; \max_{1\le j\le n}\frac{\lambda_j^2}{\sum_{r=1}^n \lambda_r^2} \;<\; \frac{1}{2}$, we then obtain

$$\big|\psi_{Q^*}(t)\big| \;\le\; 2^{1/4}\,\big(1-2\,\delta\big)^{-\frac{1}{2}}\,(2\,t^2)^{-\frac{1}{2}}.$$

It follows that

$$I_3 \;=\; \int_{|t|>A'}\left|\frac{\psi_{Q^*}(t)}{t}\right| dt \;\le\; 2^{3/4}(1-2\,\delta)^{-1/2}\int_{t\ge A'} t^{-2}\,dt \;=\; 2^{3/4}(1-2\,\delta)^{-1/2}A'^{-1}.$$

Final Assembly. Combining the pieces, we have

$$\int_{-\infty}^{\infty}\left|\frac{\psi_{Q^*}(t)-\psi_{S^*}(t)}{t}\right| dt \;\le\; (I_1+I_2+I_3)$$

which in turn implies the Kolmogorov bound

$$\sup_{x\in\mathbb{R}}\big|\Pr\big(Q^*\le x\big) - \Pr\big(S^*\le x\big)\big| \;\le\; \frac{1}{2\pi}\big(I_1+I_2+I_3\big).$$

Putting everything together, the distributional distance satisfies:

$$\begin{aligned} & \sup_{x\in\mathbb{R}}\Big|\Pr(Q\le x) \;-\; \Pr(S\le x)\Big| \\ \le\; & \frac{3\sqrt{2}}{\sqrt{\pi}A'^3\,\mathcal{K}_3(Q^*)}\,\frac{\widehat{\mathrm{KCI}}}{\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}} \;+\; \frac{3}{\pi A'^3\,\mathcal{K}_3(Q^*)}\left|\frac{\mathcal{K}_2(Q)}{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)} - 1\right| \\ & +\; \frac{1}{2^{1/4}\,\pi\,\sqrt{(1-2\,\delta)}A'} \;+\; \frac{1}{\sqrt{2\pi}A'^3}\left(\frac{\mathcal{K}_2(Q)}{n^2\,\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}\right)^{3/2}. \end{aligned}$$

Here A' is the splitting parameter and $\delta = \max_j\big\{\lambda_j^2/\sum_r\lambda_r^2\big\}$ measures spectral concentration of $\widehat{H}/n$. This first term quantifies the normalized mean shift $\frac{\widehat{\mathrm{KCI}}}{\sqrt{\mathrm{Var}(\widehat{\mathrm{KCI}}_n)}}$ caused by regression errors. Under perfect regressions, this deviation is 0. However, increased regression errors magnify this term. The second term captures discrepancies in the variance between the wild bootstrap distribution

and the Gaussian approximation. Under perfect regressions, we have equality $n^2 \operatorname{Var}(\widehat{\mathrm{KCI}}_n) = \operatorname{Var}(Q) = \mathcal{K}_2(Q)$. However, regression errors cause inflation in $\operatorname{Var}(\widehat{\mathrm{KCI}}_n)$, resulting in deviations and increasing distributional mismatch. Thus, the wild bootstrap becomes misspecified under noisy regression, as it fails to accurately reflect the mean and variance of the statistic. The third and fourth terms combined, represent intrinsic limitations and higher-order adjustments. The intrinsic mismatch term (third term) does not directly depend on regression error. The fourth term rapidly becomes negligible as sample size n increases.

Consequently, noting $\operatorname{Var}(Q) = \mathcal{K}_2(Q)$, the constants in the bound satisfy

$$R_{1,n} = \frac{1}{2^{1/4}\,\pi\,\sqrt{1-2\delta}\,A'} + \frac{1}{\sqrt{2\pi}\,A'^3}\left(\frac{\operatorname{Var}(Q)}{n^2\,\operatorname{Var}(\widehat{\mathrm{KCI}}_n)}\right)^{3/2},$$

$$R_{2,n} = \frac{3\sqrt{2}}{\sqrt{\pi}\,A'^3\,\mathcal{K}_3(Q^*)}, \quad R_{3,n} = \frac{3}{\pi\,A'^3\,\mathcal{K}_3(Q^*)}.$$

With these expressions established, the proof is complete. □

H Experimental Results

H.1 Synthetic Data

1D Synthetic Test Case We compare standard KCI, KCI with power-maximizing kernel selection, and GCM, using linear kernels for A and B in all methods, on problem (7). Figure 4 shows that GCM, while maintaining low Type I error, fails to detect conditional dependence in this setting. Standard KCI exhibits high Type II error, whereas power-maximized KCI achieves low Type II error. For both KCI variants, Type I error decreases as the training size increases.

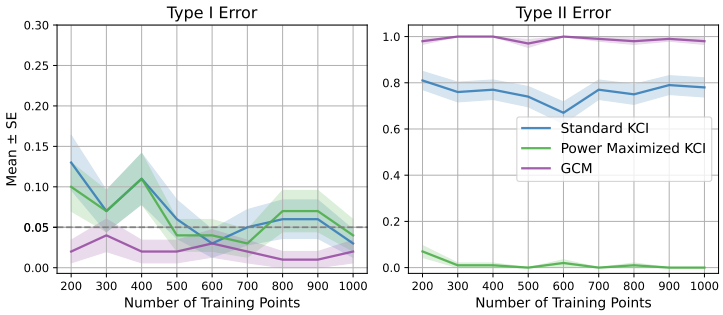
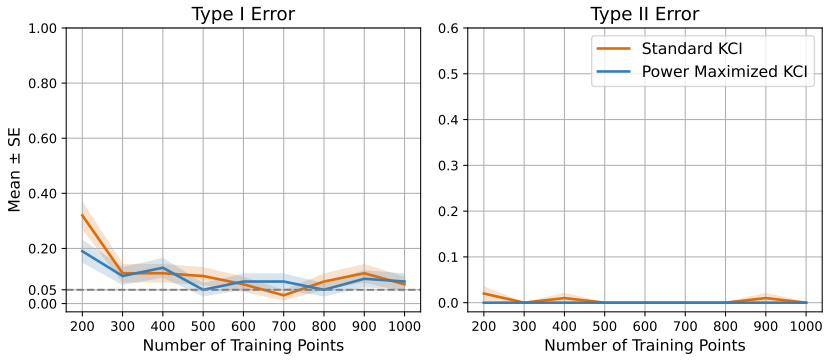


Figure 4: Mean and standard error (over 100 runs) of errors on the synthetic (7) with $f_A = \cos$, $f_B = \exp$, $\tau = 0.1$ and $\beta = 3$ across training sizes; the independent test set has size 200. The significance level is set at $\alpha = 0.05$

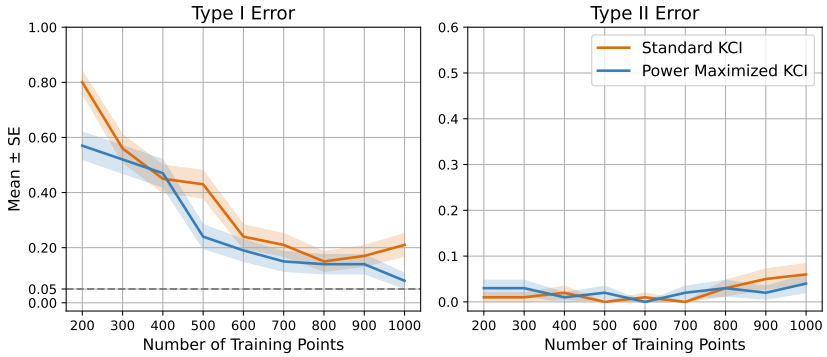
3D Synthetic Test Case We work on the problems introduced in Appendix F.3, using Gaussian kernels for A and B with $\beta = 2$. We compare standard KCI with fixed bandwidths and KCI with power-maximized kernel selection across training sizes 200–1000, with a fixed test size of 200. All kernels are Gaussian, and regression and power maximization selects per-dimension bandwidths for the C kernel. We compare results obtained with regressors trained for 500 epochs and those obtained with early stopping. Experiments are repeated 100 times, and we report mean and standard error.

We conduct experiments on two scenarios. Scenario 1: shared-dimension dependence, where A and B depend on the same coordinate of C_1 (see Figure 5). Scenario 2: separate-dimension dependence, where A depends on C_1, B on C_2, correlation on C_3 (see Figure 6).

Results show that poorly-trained conditional mean embeddings inflate Type I error, particularly in Scenario 1 where shared dimensions induce strongly correlated regression errors. Interestingly, power maximization further amplifies Type I error in Scenario 2 when CMEs are undertrained, but in Scenario 1, it slightly reduces the error by focusing on the relevant coordinate C_1 and ignoring irrelevant ones.



(a) Sufficient Training.



(b) Early Stopping.

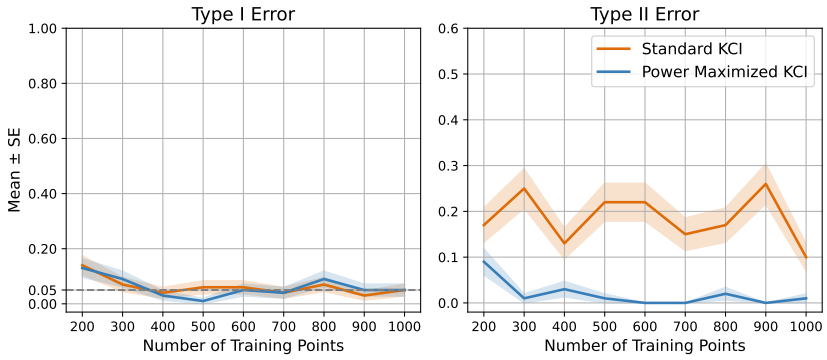
Figure 5: Shared-dimension dependence. Means and standard errors (over 100 runs) of Type I and Type II errors on the 3D synthetic case with $f_A = \cos$, $f_B = \exp$, and $\tau = 0.1$. All kernels are Gaussian, and the significance level is set at $\alpha = 0.05$.

H.2 real data

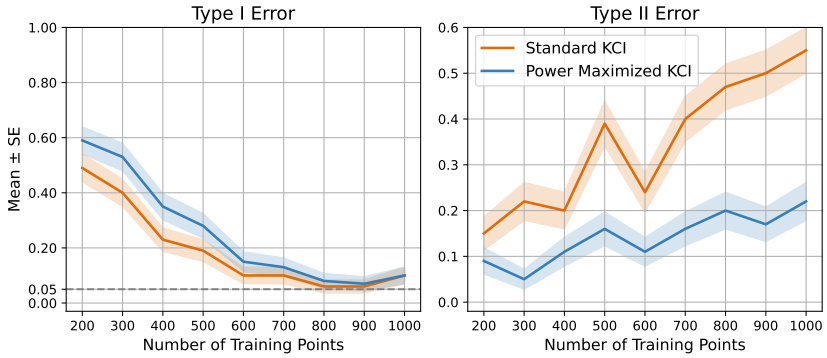
We conducted experiments on the UTKFace dataset (Z. Zhang et al., 2017), following the setup of Y. Zhang et al. (2025). Although not described as such in their paper, this test is effectively a KCI test. In particular, we used the cropped and aligned UTKFace dataset to test whether age (A) depends on the full face image (B) when conditioned on the same image with a specific region masked out (C). The null hypothesis is $\mathbb{E}[A \mid C] = \mathbb{E}[A \mid B, C]$, corresponding to a linear kernel on A, where the conditional mean embedding can be estimated as a regressor from C to A via neural networks.

The dataset is split into ten subsets, each with its own training and test partition. We reran their code and report both the resulting p-values for the conditional independence tests and the mean absolute error (MAE) of the age regressors (ImageNet-pretrained) used in testing. For comparison, we also retrained the same network from scratch (random initialization) and report its MAE and corresponding test results. The results are shown in Figure 7.

In Y. Zhang et al. (2025), p-values remain above 5% when a facial region is masked, suggesting that the region is not critical for age estimation. However, when the same network is trained from random initialization, the validation loss increases–indicating a less accurate conditional mean embedding–and the resulting p-values drop consistently across all regions. This shows that test outcomes are highly sensitive to regressor quality: imperfect conditional mean estimation makes the test more prone to signal dependence.



(a) Sufficient Training.



(b) Early Stopping.

Figure 6: Separate-dimension dependence. Means and standard errors (over 100 runs) of Type I and Type II errors on the 3D synthetic case with $f_A = \cos$, $f_B = \exp$, and $\tau = 0.1$. All kernels are Gaussian, and the significance level is set at $\alpha = 0.05$.

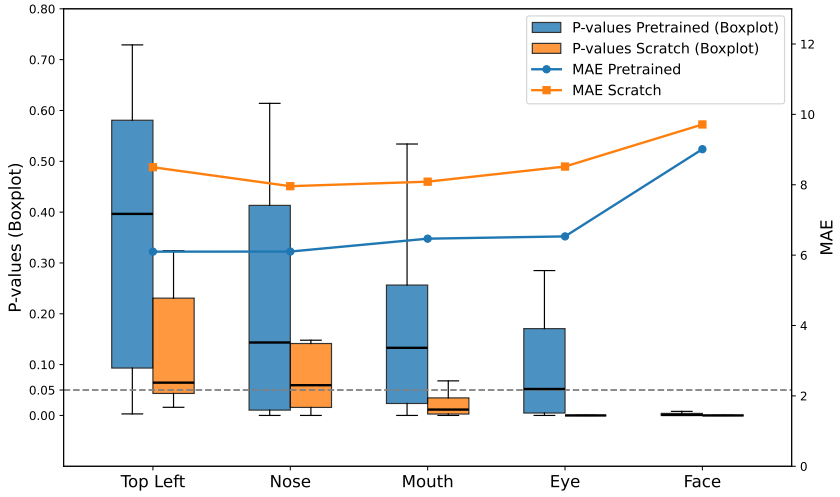


Figure 7: Box plots of p-values (left y-axis) and test MAE (right y-axis) across different facial regions in the age estimation task. "Pretrained" refers to using an ImageNet-pretrained age regressor, while "Scratch" indicates training the same model from random initialization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made cover what we did in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Throughout, especially e.g. following Theorems 6.2 and 6.3 and in the supplement.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Either in the theorem statement or the supplement.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: While we have few experiments, they are clearly described and code will be provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: In the supplement.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the supplement and the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiments are limited, but we have error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The amount of computation used was very limited by modern standards.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No relevant issues.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This is a foundational theoretical paper and does not really propose a new algorithm.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The only existing asset is code we built on, which we cite and specify in the source.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our code (in the supplement) is documented.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Not used in such ways.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.