The Generalised Kernel Covariance Measure

Luca Bergen BERGEN@LEIBNIZ-BIPS.DE
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany
Dino Sejdinovic DINO.SEJDINOVIC@ADELAIDE.EDU.AU
School of Mathematical Sciences and Australian Institute for Machine Learning
Adelaide University, Adelaide, Australia
Vanessa Didelez DIDELEZ@LEIBNIZ-BIPS.DE
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

We consider the problem of conditional independence (CI) testing and adopt a kernel-based approach. Kernel-based CI tests embed variables in reproducing kernel Hilbert spaces, regress their embeddings on the conditioning variables, and test the resulting residuals for marginal independence. This approach yields tests that are sensitive to a broad range of conditional dependencies. Existing methods, however, rely heavily on kernel ridge regression, which is computationally expensive when properly tuned and yields poorly calibrated tests when left untuned, which limits their practical usefulness. We propose the Generalised Kernel Covariance Measure (GKCM), a regression-model-agnostic kernel-based CI test that accommodates a broad class of regression estimators. Building on the Generalised Hilbertian Covariance Measure framework (Lundborg et al., 2022), we characterise conditions under which GKCM satisfies uniform asymptotic level guarantees. In simulations, GKCM paired with tree-based regression models frequently outperforms state-of-the-art CI tests across a diverse range of data-generating processes, achieving better type I error control and competitive or superior power.

Keywords: Conditional independence testing, kernel conditional independence test, causal discovery

1. Introduction

Conditional independence (CI) is a central notion in probability theory and statistics (Dawid, 1979). It admits several equivalent characterisations, cf. Constantinou and Dawid (2017) for details. We will use the following: for random variables X, Y, Z with values in $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$, and $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$, respectively, we say that X is conditionally independent of Y given Z, denoted by $X \perp\!\!\!\perp Y \mid Z$, if

$$\mathrm{Cov}(f(X), g(Y) \mid Z) = \mathbb{E}[f(X)g(Y) \mid Z] - \mathbb{E}[f(X) \mid Z]\mathbb{E}[g(Y) \mid Z] = 0 \tag{1}$$

almost surely for all bounded and measurable functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$. Conditional independence can be inferred from data via CI tests, i.e., statistical tests of the null hypothesis $X \perp\!\!\!\perp Y \mid Z$. Such tests play a key role in causal inference, for example in constraint-based causal discovery (Glymour et al., 2019) and invariant causal prediction (Peters et al., 2016; Heinze-Deml et al., 2018).

Different methods are used for CI testing in practice, even though they may formally target a null hypothesis that is larger than the null of conditional independence. Many of these fall into two groups, which we refer to as residual- and kernel-based tests. In the following we briefly review these methods before introducing GKCM.

1.1. Residual-based tests

Throughout this section, let X and Y be square-integrable and let $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^q$. The first group of methods, which we call residual-based tests, is based on testing whether the mean conditional covariance matrix

$$\mathbb{E}_Z[\mathrm{Cov}(X, Y \mid Z)] = \mathbb{E}[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z])^\intercal] = 0, \tag{2}$$

which is implied by CI. Well-known examples are tests of vanishing partial covariance under the assumption of joint normality, where the conditional expectations $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[Y \mid Z]$ are estimated by least-squares linear regression models (see, e.g., Anderson, 2003, Section 15.5). Another example is the Generalised Covariance Measure (GCM, Shah and Peters, 2020), which avoids overly restrictive assumptions like joint normality. Instead, it allows the conditional expectations to have complex nonlinear dependencies on Z, which may be estimated using flexible regression or machine-learning methods. Unlike most CI tests, GCM can achieve uniform rather than merely pointwise asymptotic level over subsets of the null where the prediction errors of the regression models vanish sufficiently fast.[1]

However, without strong parametric assumptions like joint normality, residual-based tests are limited by the fact that there exist distributions from the alternative satisfying $\mathbb{E}_Z[\mathrm{Cov}(X, Y \mid Z)] = 0$, which renders the conditional dependence undetectable to the tests. This limitation occurs because the condition (2) is strictly weaker than CI as characterised by (1) in two distinct respects. First, the tests do not consider the bounded functions of X and Y, but instead (implicitly) the linear functions, since

$$\mathbb{E}_Z[\mathrm{Cov}(X, Y \mid Z)] = 0 \iff \mathbb{E}_Z[\mathrm{Cov}(\mathrm{u}^\intercal X, \mathrm{v}^\intercal Y \mid Z)] = 0$$

for all $\mathrm{u} \in \mathbb{R}^p$ and $\mathrm{v} \in \mathbb{R}^q$. Second, the tests do not assess whether the conditional covariances vanish almost surely, but only whether they vanish in expectation over Z for all pairs of functions. Hence, even if the tests were to use sufficiently large function classes, they still could not detect conditional dependence under distributions where the mean conditional covariances vanish for all pairs of functions. We call methods targeting the mean conditional covariances mean-zero (as opposed to a.s.-zero) tests regardless of the function spaces considered. When considering all square-integrable functions of X and Y, mean-zero testing corresponds to testing for weak CI (Daudin, 1980).

Motivated by these shortfalls Scheidegger et al. (2022) proposed the weighted GCM (wGCM) and Lundborg et al. (2024) the projected covariance measure (PCM). Both methods allow for the use of arbitrary regression models and uniform level guarantees (under the required assumptions), while being able to detect conditional dependence under larger subsets of the alternative. The wGCM tests whether $\mathrm{Cov}(X, Y \mid Z) = 0$ almost surely by weighting the regression residuals with a function of Z. For scalar X, PCM tests for conditional mean independence, i.e.,

$$\mathbb{E}[X \mid Y, Z] = \mathbb{E}[X \mid Z] \iff \mathbb{E}_Z[\mathrm{Cov}(X, h(Y, Z) \mid Z)] = 0$$

for all square-integrable functions h. It does so by replacing Y in (2) with a scalar-valued projection $f(Y, Z)$, which is regressed on Z. Testing for conditional mean independence is equivalent to a.s.-zero testing while additionally considering nonlinear functions of Y. Yet both tests only consider linear functions (at least) of X.

1. For the distinction between uniform and pointwise asymptotic level, see, e.g., Shah and Peters (2020, Section 1).

1.2. Kernel-based tests

The second group of methods, which we call kernel-based tests, combines residual-based testing with a large number of transformations by embedding the random variables in reproducing kernel Hilbert spaces as high-dimensional feature spaces. Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. A Hilbert space $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ is called a reproducing kernel Hilbert space (RKHS) if there exists a symmetric positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying $k(\cdot, x) \in \mathcal{F}$ for all $x \in \mathcal{X}$, and $\langle k(\cdot, x), f \rangle_{\mathcal{F}} = f(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$ (the *reproducing property*). When k satisfies these properties, it is called the reproducing kernel of $(\mathcal{F}, \langle \cdot, \cdot \rangle_{\mathcal{F}})$ (see, e.g., Berlinet and Thomas-Agnan, 2004). The reproducing kernel is used to map $\mathcal{X} \to \mathcal{F}$ via the canonical feature map $\phi : x \mapsto k(\cdot, x)$. Depending on the kernel and domain, each mapping or feature vector $\phi(x)$ may encode an infinite number of non-linear transformations of x. By the reproducing property the feature vectors satisfy

$$\langle \phi(x), \phi(x') \rangle_{\mathcal{F}} = k(x, x') \quad \forall x, x' \in \mathcal{X}, \tag{3}$$

which enables the implicit evaluation of their inner products via the kernel. Algorithms are typically designed only to rely on these inner products, which allows them to avoid evaluating the high-dimensional feature vectors (commonly known as the *kernel trick*).

Examples of kernel-based CI tests include the seminal Kernel CI Test (KCIT, Zhang et al., 2011), the Kernel Regression with Subsequent Independence Test (KRESIT, Zhang et al., 2017), the Randomized Conditional Independence Test (RCIT, Strobl et al., 2019), and the Randomized Correlation Test (RCoT, Strobl et al., 2019). In all four methods, the variables are embedded into RKHSs and the mean conditional covariance of the embeddings is estimated via kernel ridge regression (see Sections 1.3 and 4.1). The tests differ mainly in which variables are embedded and in the dimensions of the RKHSs used: whereas KRESIT and RCoT embed X and Y, KCIT and RCIT embed (X, Z) and Y, similar to PCM. Accordingly, the former are mean-zero and the latter are a.s.-zero tests. Furthermore, KCIT and KRESIT use Gaussian kernels which induce infinite-dimensional RKHSs, whereas RCIT and RCoT approximate these kernels using Random Fourier Features (Rahimi and Recht, 2007), which induce finite-dimensional RKHSs. Therefore RCIT and RCoT may be seen as faster, approximate versions of KCIT and KRESIT, respectively.

Using the kernel trick, kernel-based tests are able to consider a large number of transformations of X and Y simultaneously. This enables the tests to detect conditional dependencies under large subsets of the alternative. By using suitable kernels they can also accommodate different variable types and mixed data. However, even though there exists a wide range of RKHS-valued regression methods, existing kernel-based CI tests exclusively use kernel ridge regression. This is detrimental to their performance, since kernel ridge regression is sensitive to the choice of hyperparameters and tuning the hyperparameters is computationally costly. Thereby the tests become either computationally prohibitive or unreliable when hyperparameter tuning is bypassed. Furthermore, to the best of our knowledge, it has not been investigated whether asymptotic uniform type-I error guarantees can be shown to hold for kernel-based tests. This is subpar compared to the residual-based tests, where sufficient conditions have been stated without prior restriction of the regression methods to be used in testing (Shah and Peters, 2020; Scheidegger et al., 2022; Lundborg et al., 2024).

1.3. The key idea of GKCM

To resolve these issues, we propose the Generalised Kernel Covariance Measure (GKCM). Our method is similar to the existing kernel-based CI tests in procedure, especially to KRESIT, but

allows for the use of arbitrary Hilbert space-valued regression methods. Furthermore, we define GKCM as a Generalised Hilbertian Covariance Measure (GHCM, Lundborg et al., 2022), which allows us to state conditions for uniform asymptotic level guarantees. GHCM tests are a class of CI tests for Hilbert space-valued random variables introduced by Lundborg et al. (2022). The authors show that GHCM tests have uniform asymptotic type-I error control under assumptions similar to GCM; most importantly, the in-sample prediction errors of the regression methods need to vanish sufficiently fast uniformly over the subset of distributions. While the GHCM framework has originally been developed for CI testing in functional data, we use it instead on the RKHS embeddings to infer CI between X and Y. We show that, under mild assumptions and for many choices of kernels, X and Y are conditionally independent if and only if their embeddings are conditionally independent, which justifies our approach.

For a first idea of GKCM, let $(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mathcal{F}})$ and $(\mathcal{G}, \langle\cdot,\cdot\rangle_{\mathcal{G}})$ denote RKHSs with reproducing kernels $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$, and canonical feature maps defined by $\phi(x) := k(\cdot, x)$ and $\varphi(y) := l(\cdot, y)$, respectively. GKCM targets the mean conditional covariance of $\phi(X)$ and $\varphi(Y)$ given Z defined as the operator

$$\mathbf{C}_{XY\cdot Z} := \mathbb{E}\big[\big(\phi(X) - \mathbb{E}[\phi(X) \mid Z]\big) \otimes \big(\varphi(Y) - \mathbb{E}[\varphi(Y) \mid Z]\big)\big],$$

where $\phi(X)$ and $\varphi(Y)$ as well as their conditional expectations are random variables taking values in $\mathcal{F}$ and $\mathcal{G}$, respectively, and for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, the rank-one operator $f \otimes g : \mathcal{G} \to \mathcal{F}$ is defined by $(f \otimes g)(g') := \langle g, g'\rangle_{\mathcal{G}} f$ analogously to the outer product $(\mathrm{u}\mathrm{v}^\intercal)\mathrm{v}' = (\mathrm{v}^\intercal\mathrm{v}')\mathrm{u}$ in Euclidean space (Pogodin et al., 2025; Park and Muandet, 2020). Since GKCM tests whether

$$\mathbf{C}_{XY\cdot Z} = 0 \iff \mathbb{E}_Z[\mathrm{Cov}(f(X), g(Y) \mid Z)] = 0 \quad \forall f \in \mathcal{F}, g \in \mathcal{G},$$

we expect it to detect conditional dependencies under large subsets of the alternative when using kernels with sufficiently rich RKHSs. In particular, if k and l are L^2-universal, i.e. $\mathcal{F}$ and $\mathcal{G}$ are dense in $L^2(\mathcal{X}, \mu)$ and $L^2(\mathcal{Y}, \nu)$ with respect to the L^2-norm for all Borel probability measures μ and ν (Sriperumbudur et al., 2011), the condition $\mathbf{C}_{XY\cdot Z} = 0$ is equivalent to weak CI. Table 1 compares GKCM to the other tests.

	GCM	**wGCM**	**PCM**	**KCIT**	**KRESIT**	**RCIT**	**RCoT**	**GKCM**
Regression methods	Any	Any	Any	KRR	KRR	KRR†	KRR†	Any
Type-I error control	Uniform	Uniform	Uniform	Pointw.	Pointw.	Pointw.	Pointw.	Uniform
Function spaces	Linear	Linear	Mixed	RKHS	RKHS	RKHS†	RKHS†	RKHS
Conditional covariances	Mean-z.	A.s.-zero	A.s.-zero	A.s.-zero	Mean-z.	A.s.-zero	Mean-z.	Mean-z.

† Approximated using Random Fourier Features

Table 1: Comparison of residual- and kernel-based CI tests

Like GCM, GKCM is a mean-zero test. In contrast, KCIT uses a joint embedding of (X, Z) to test for a.s.-zero conditional covariances. Under L^2-universal kernels, this is equivalent to testing the stronger null hypothesis of CI (Pogodin et al., 2025; He et al., 2025). Although the use of a joint embedding does not strengthen central assumptions required for asymptotic validity, it may inflate type-I error rates in finite samples (see Appendix B for details and the discussion in He et al., 2025). To better control finite-sample type-I error, we therefore refrain from using a joint embedding.

2. Background and assumptions

Let X, Y, Z be random variables on the measurable space $(\Omega, \mathcal{A})$ with values in measurable spaces $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$, and $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$, respectively, where $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are topological spaces and $\mathcal{B}_\mathcal{X}, \mathcal{B}_\mathcal{Y}, \mathcal{B}_\mathcal{Z}$ denote their Borel σ-algebras. The measurable space $(\Omega, \mathcal{A})$ is equipped with a family of probability measures $(\mathbb{P}_P)_{P\in\mathcal{P}}$ such that the joint distribution of (X, Y, Z) under $\mathbb{P}_P$ is P. We denote the null hypothesis by $\mathcal{P}_0 := \{P \in \mathcal{P} : X \perp\!\!\!\perp Y \mid Z\}$. For $\tilde{\mathcal{P}} \subseteq \mathcal{P}$ and a sequence of real-valued random variables $(V_n)_{n\in\mathbb{N}}$, we write $V_n = o_{\tilde{\mathcal{P}}}(1)$ to denote that for all $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{P\in\tilde{\mathcal{P}}} \mathbb{P}_P(|V_n| \geq \epsilon) = 0.$$

Furthermore, $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ denotes the set of Hilbert-Schmidt (HS) operators $\mathcal{G} \to \mathcal{F}$ (Hsing and Eubank, 2015, Ch. 4.4), which has a Hilbert space structure when equipped with the inner product $\langle \mathbf{A}, \mathbf{B}\rangle_{\mathrm{HS}} := \sum_{j\in J} \langle \mathbf{A}g_j, \mathbf{B}g_j\rangle_\mathcal{F}$, where $(g_j)_{j\in J}$ is an arbitrary orthonormal basis of $\mathcal{G}$. In particular, for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, we have $f \otimes g \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$.

Assumptions In the following, we assume that

A.1 $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are Polish spaces,

A.2 k and l are continuous,

A.3 k and l are bounded, i.e., $\kappa_X := \sup_{x\in\mathcal{X}} k(x, x) < \infty$ and $\kappa_Y := \sup_{y\in\mathcal{Y}} l(y, y) < \infty$,

A.4 ϕ and φ are injective.

These assumptions ensure that all population quantities from the previous section are well defined, that the regularity assumptions of the GHCM framework (Lundborg et al., 2022, Section 1.2) are satisfied, and that testing for $\phi(X) \perp\!\!\!\perp \varphi(Y) \mid Z$ is equivalent to testing for $X \perp\!\!\!\perp Y \mid Z$ since the two CI statements coincide, as we argue below.

Since $\mathcal{X}$ and $\mathcal{Y}$ are separable by assumption A.1, it follows from A.2 and Steinwart and Christmann (2008, Lemma 4.33) that $\mathcal{F}$ and $\mathcal{G}$ are separable. As $(\mathcal{F}, \langle\cdot,\cdot\rangle_\mathcal{F})$ and $(\mathcal{G}, \langle\cdot,\cdot\rangle_\mathcal{G})$ are complete, the separability of $\mathcal{F}$ and $\mathcal{G}$ implies that they are Polish when endowed with their norm topologies. Moreover, by A.2 and Steinwart and Christmann (2008, Lemma 4.29), ϕ and φ are continuous and therefore Borel measurable. Hence $\phi(X)$ and $\varphi(Y)$ are random variables on $(\Omega, \mathcal{A})$ taking values in the standard Borel spaces $(\mathcal{F}, \mathcal{B}_\mathcal{F})$ and $(\mathcal{G}, \mathcal{B}_\mathcal{G})$, respectively.

Since $\phi(X)$ and $\varphi(Y)$ are Borel measurable and $\mathcal{F}$ and $\mathcal{G}$ are separable, $\phi(X)$ and $\varphi(Y)$ are strongly measurable. By assumption A.3 $\mathbb{E}[\|\phi(X)\|_\mathcal{F}^2] = \mathbb{E}[k(X, X)] < \infty$ and $\mathbb{E}[\|\varphi(Y)\|_\mathcal{G}^2] = \mathbb{E}[l(Y, Y)] < \infty$ for all $P \in \mathcal{P}$. Hence $\phi(X)$ and $\varphi(Y)$ are square-integrable and therefore Bochner integrable (Cohn, 2013, Appendix E). Therefore the conditional expectations $\mathbb{E}[\phi(X) \mid Z]$ and $\mathbb{E}[\varphi(Y) \mid Z]$ (henceforth conditional mean embeddings) exist as $\sigma(Z)$-measurable random variables (Park and Muandet, 2020). By the Doob-Dynkin Lemma (Kallenberg, 2021, Lemma 1.14), there exist Borel-measurable functions $F_P : \mathcal{Z} \to \mathcal{F}$ and $G_P : \mathcal{Z} \to \mathcal{G}$ satisfying $\mathbb{E}[\phi(X) \mid Z] = F_P(Z)$ and $\mathbb{E}[\varphi(Y) \mid Z] = G_P(Z)$ almost surely for each $P \in \mathcal{P}$.

Assumptions A.1, A.2, and A.4 ensure that CI is preserved under ϕ and φ. By Constantinou and Dawid (2017, Proposition 2.3) $X \perp\!\!\!\perp Y \mid Z$ is equivalent to $\mathbb{E}[\mathbb{1}_{A\cap B} \mid Z] = \mathbb{E}[\mathbb{1}_A \mid Z]\mathbb{E}[\mathbb{1}_B \mid Z]$ almost surely for all $A \in \sigma(X)$ and $B \in \sigma(Y)$. Hence

$$X \perp\!\!\!\perp Y \mid Z \iff \phi(X) \perp\!\!\!\perp \varphi(Y) \mid Z$$

if $\sigma(\phi(X)) = \sigma(X)$ and $\sigma(\varphi(Y)) = \sigma(Y)$, which follows from Lemma 1 (proved in Appendix A).

Lemma 1 *Let $\mathcal{X}$ and $\mathcal{F}$ denote Polish spaces. For every random variable X taking values in the standard Borel space $(\mathcal{X}, \mathcal{B}_X)$ and every Borel-measurable and injective function $\phi : \mathcal{X} \to \mathcal{F}$, it holds that $\sigma(X) = \sigma(\phi(X))$.*

All of the assumptions are mild and can be shown to hold in many settings. Assumptions A.1 – A.4 hold, e.g., for $\mathbb{R}^d$ and open or closed subsets equipped with Gaussian kernels, or for finite sets equipped with the Dirac kernel.

3. GKCM: Definition and Properties

To formally define GKCM, let $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ be an i.i.d. sample of (X, Y, Z), and let $\hat{F}_n : \mathcal{Z} \to \mathcal{F}$, $\hat{G}_n : \mathcal{Z} \to \mathcal{G}$ denote regression functions trained in-sample. Define the centred residuals

$$\hat{\varepsilon}_i := \phi(X_i) - \hat{F}_n(Z_i) - \hat{\mu}_\varepsilon, \quad \hat{\xi}_i := \varphi(Y_i) - \hat{G}_n(Z_i) - \hat{\mu}_\xi,$$

$$\hat{\mu}_\varepsilon := \frac{1}{n}\sum_{i=1}^n (\phi(X_i) - \hat{F}_n(Z_i)), \quad \hat{\mu}_\xi := \frac{1}{n}\sum_{i=1}^n (\varphi(Y_i) - \hat{G}_n(Z_i)).$$

The resulting empirical mean conditional covariance operator is

$$\hat{\mathbf{C}}^{(n)}_{XY \cdot Z} := \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i \otimes \hat{\xi}_i,$$

and the test statistic of GKCM is defined as

$$T_n := n\|\hat{\mathbf{C}}^{(n)}_{XY \cdot Z}\|^2_{\mathrm{HS}}.$$

Note that analogous test statistics are used in KCIT and KRESIT as well as in the GHCM framework, which allows us to use properties established for the latter in the following.

Lundborg et al. (2022) show that over subsets of the null hypothesis where the in-sample prediction errors of the regression models vanish sufficiently fast and additional moment conditions are met, the scaled operator $\sqrt{n}\hat{\mathbf{C}}^{(n)}_{XY \cdot Z}$ converges uniformly in distribution to a mean-zero Gaussian random element (Da Prato and Zabczyk, 2014, Section 2.3) with unknown covariance operator $\mathbf{C}_P$. As a consequence, the asymptotic null distribution of the quadratic form $T_n = \|\sqrt{n}\hat{\mathbf{C}}^{(n)}_{XY \cdot Z}\|^2_{\mathrm{HS}}$ is characterised by the eigenvalues of $\mathbf{C}_P$. These are consistently estimated by the non-zero eigenvalues $(\lambda_i)_{i=1}^d$ of the matrix

$$\mathrm{T} := \frac{1}{(n-1)}\mathrm{HRH},$$

where $\mathrm{R} \in \mathbb{R}^{n \times n}$ with $\mathrm{R}_{ij} := \langle \hat{\varepsilon}_i, \hat{\varepsilon}_j \rangle_{\mathcal{F}} \langle \hat{\xi}_i, \hat{\xi}_j \rangle_{\mathcal{G}}$, and $\mathrm{H} := \mathrm{I}_n - n^{-1}\mathrm{J}_n$. The null distribution of T_n is then approximated by the generalised chi-square distribution of $\sum_{i=1}^d \lambda_i V_i^2$, where V_i are i.i.d. standard normal random variables, and the level-α test function is defined as

$$\tau_n := \mathbb{1}\{T_n \geq q_\alpha\},$$

where q_α denotes the $1-\alpha$ quantile of $\sum_{i=1}^d \lambda_i V_i^2$. In practice the test statistic can be computed as $T_n = n^{-1}\sum_{i,j=1}^n \mathrm{R}_{ij}$ and the null distribution can be approximated using miscellaneous methods (see, e.g., Bodenham and Adams, 2016).

Lundborg et al. (2022, Theorems 2 and 3) state conditions for uniform convergence and for the resulting uniform level guarantees of GHCM tests. For completeness, these are reproduced below in Theorem 2. In order to state the condition and results we define the population residuals

$$\varepsilon_P := \phi(X) - \mathbb{E}[\phi(X) \mid Z], \quad \xi_P := \varphi(Y) - \mathbb{E}[\varphi(Y) \mid Z],$$

the conditional variances

$$u_P(z) := \mathbb{E}[\|\varepsilon_P\|_{\mathcal{F}}^2 \mid Z = z], \quad v_P(z) := \mathbb{E}[\|\xi_P\|_{\mathcal{G}}^2 \mid Z = z],$$

and the (weighted) in-sample mean square prediction errors

$$\mathcal{E}_n^F := \frac{1}{n}\sum_{i=1}^n \|F_P(Z_i) - \hat{F}_n(Z_i)\|_{\mathcal{F}}^2, \quad \mathcal{E}_n^G := \frac{1}{n}\sum_{i=1}^n \|G_P(Z_i) - \hat{G}_n(Z_i)\|_{\mathcal{G}}^2,$$

$$\tilde{\mathcal{E}}_n^F := \frac{1}{n}\sum_{i=1}^n \|F_P(Z_i) - \hat{F}_n(Z_i)\|_{\mathcal{F}}^2 v_P(Z_i), \quad \tilde{\mathcal{E}}_n^G := \frac{1}{n}\sum_{i=1}^n \|G_P(Z_i) - \hat{G}_n(Z_i)\|_{\mathcal{G}}^2 u_P(Z_i).$$

With the above arguments, we have the following result.

Theorem 2 *Let $\tilde{\mathcal{P}}_0 \subset \mathcal{P}_0$ such that*

B.1 $n\mathcal{E}_n^F\mathcal{E}_n^G = o_{\tilde{\mathcal{P}}_0}(1)$,

B.2 $\tilde{\mathcal{E}}_n^F = o_{\tilde{\mathcal{P}}_0}(1)$ *and* $\tilde{\mathcal{E}}_n^G = o_{\tilde{\mathcal{P}}_0}(1)$,

B.3 $\inf_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}[\|\varepsilon_P\|_{\mathcal{F}}^2\|\xi_P\|_{\mathcal{G}}^2] > 0$ *and* $\sup_{P\in\tilde{\mathcal{P}}_0} \mathbb{E}[\|\varepsilon_P\|_{\mathcal{F}}^{2+\eta}\|\xi_P\|_{\mathcal{G}}^{2+\eta}] < \infty$ *for some* $\eta > 0$,

B.4 for some orthonormal bases $(f_i)_{i\in I}$ and $(g_j)_{j\in J}$ of $\mathcal{F}$ and $\mathcal{G}$, it holds that

$$\lim_{K\to\infty} \sup_{P\in\tilde{\mathcal{P}}_0} \sum_{(i,j):i+j\geq K} \mathbb{E}[\langle f_i, \varepsilon_P\rangle_{\mathcal{F}}^2 \langle g_j, \xi_P\rangle_{\mathcal{G}}^2] = 0,$$

B.5 $\inf_{P\in\tilde{\mathcal{P}}_0} \|\mathbf{C}_P\|_{\mathrm{op}} > 0$.

Then, for each $\alpha \in (0,1)$, the level-α GKCM test τ_n satisfies

$$\lim_{n\to\infty} \sup_{P\in\tilde{\mathcal{P}}_0} |\mathbb{P}_P(\tau_n = 1) - \alpha| = 0.$$

Theorem 2 implies that the central challenge in kernel-based CI testing consists in estimating the conditional mean embeddings. This conclusion has recently been stated independently for kernel-based CI tests based on U-statistics by He et al. (2025). Assumption B.1 specifies the (product) error rates required to hold for the regression models uniformly over $\tilde{\mathcal{P}}_0$. The assumption is satisfied, e.g., if $\sqrt{n}\mathcal{E}_n^F = o_{\tilde{\mathcal{P}}_0}(1)$ and $\sqrt{n}\mathcal{E}_n^G = o_{\tilde{\mathcal{P}}_0}(1)$, yet one model may converge slower if the other is able to compensate. Since the mean square prediction errors are defined in-sample, the regression models are not required to extrapolate well out-of-sample. By the discussion in Lundborg et al. (2022, Section 4.2), assumption B.2 is satisfied if $\mathcal{E}_n^F = o_{\tilde{\mathcal{P}}_0}(1)$ and $\mathcal{E}_n^G = o_{\tilde{\mathcal{P}}_0}(1)$, since $u_P(Z_i) \leq 4\kappa_X$ and $v_P(Z_i) \leq 4\kappa_Y$ almost surely.

4. RKHS-valued regression

Since the central challenge in kernel-based CI testing is to specify suitable regression models, we review the main approaches to this RKHS-valued or output-kernel regression problem. In particular, we focus on aspects most relevant in practice, namely modelling assumptions, robustness to model misspecification, tuning requirements, and computational cost.

4.1. Kernel ridge regression

A prominent subgroup consists of identity-decomposable input–output kernel regression (IOKR) methods, which are parametrised by an input kernel $m : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ (Brouard et al., 2016). In these methods, the feature map $\psi(z) := m(\cdot, z) \in \mathcal{H}$ is used as a high-dimensional feature expansion of the covariates. The regression coefficients are HS operators $\hat{\mathbf{C}} : \mathcal{H} \to \mathcal{F}$ learned via regularised empirical risk minimisation, yielding regression models of the form $\hat{F}_n(z) = \hat{\mathbf{C}}\psi(z)$. Examples include kernel ridge regression (Grünewälder et al., 2012; Li et al., 2022), robust regression methods (Laforgue et al., 2020), and kernel principal component regression (Meunier et al., 2024).

The most common IOKR method and the standard method in kernel-based CI testing is kernel ridge regression (KRR). For a fixed input kernel m and regularisation parameter $\lambda > 0$, the KRR model is defined by $\hat{F}_n(z) = \hat{\mathbf{C}}_\lambda \psi(z)$ with coefficients

$$\hat{\mathbf{C}}_\lambda := \operatorname*{arg\,min}_{\mathbf{C} \in \mathrm{HS}(\mathcal{H},\mathcal{F})} \frac{1}{n} \sum_{i=1}^{n} \|\phi(x_i) - \mathbf{C}\psi(z_i)\|_{\mathcal{F}}^2 + \lambda \|\mathbf{C}\|_{\mathrm{HS}}^2.$$

The model admits the closed-form expression

$$\hat{F}_n(z) = \frac{1}{n} \sum_{i=1}^{n} w_i(z)\phi(x_i), \quad \mathrm{w}(z) := (\mathrm{M} + n\lambda \mathrm{I}_n)^{-1} \mathrm{m}(z) \in \mathbb{R}^n, \tag{4}$$

where $\mathrm{M} \in \mathbb{R}^{n \times n}$ is a Gram matrix with $\mathrm{M}_{ij} := m(z_i, z_j)$ and $\mathrm{m}(z) := [m(z_1, z), \ldots, m(z_n, z)]^\intercal$.

The performance of KRR is codetermined by the choice of input kernel. For covariates taking values in $\mathcal{Z} \subseteq \mathbb{R}^d$, the established kernel-based CI tests often use Gaussian tensor product kernels (Szabó and Sriperumbudur, 2018)

$$m(z, z') = \prod_{j=1}^{d} m_j(z_j, z_j'), \quad m_j(z_j, z_j') = \exp\left(-\frac{(z_j - z_j')^2}{2\sigma_j^2}\right),$$

which are parametrised by lengthscales $\sigma_1, \ldots, \sigma_d > 0$ (other parametric kernel families may be considered in addition). The lengthscales and λ are treated as hyperparameters and can be tuned jointly, e.g. using surrogate likelihood maximisation (Zhang et al., 2011) or leave-one-out cross-validation (Pogodin et al., 2025). While λ controls the norm of the coefficients, the lengthscales control the effective size of the RKHS, since the RKHSs associated with Gaussian tensor-product kernels are nested in the lengthscales (Zhang and Zhao, 2011, Propositions 3.5 and 5.2).

The input kernel matters because the population learning rate of KRR depends on the eigenvalue decay of the kernel integral operator, on an embedding property of the associated RKHS, and on the smoothness of the true regression function F_P relative to the hypothesis space (Li et al., 2024). Heuristically, enlarging the hypothesis space may improve approximation, whereas shrinking it may

improve estimation. Whether a more restrictive hypothesis space is beneficial depends on whether it still contains sufficiently good approximations to the true regression function. In practice, however, this is typically unknown and cannot be assessed directly.

At the same time, data-driven model selection is computationally demanding, since fitting a candidate model has computational cost growing cubically in n. Even for a fixed kernel family, choosing a separate lengthscale for each covariate plus λ requires tuning $d+1$ hyperparameters. In many settings, this makes exhaustive tuning infeasible, so implementations often rely on heuristic choices instead. Since KRR is typically sensitive to the choice of tuning parameters, the covariate feature map often becomes a bottleneck in kernel-based CI testing, either computationally through tuning costs or statistically through suboptimal regression performance.

4.2. Random Forests

The above practical issues motivate the use of alternative regression methods that do not rely on input kernels. Examples include RKHS-valued extensions of random forests (Geurts et al., 2006; Cevid et al., 2022), neural networks (Shimizu et al., 2024; El Ahmad et al., 2024) or gradient-boosted learners (Geurts et al., 2007).

In the following, we will focus on random forests. Compared to KRR, they avoid the need to specify an input kernel and typically require much less hyperparameter tuning. This is well known in the scalar-valued setting and appears to hold for the RKHS-valued case as well. In our experiments, random forests often achieved strong performance without any parameter tuning (see Section 5). Finally, their representation of the covariates is more transparent: while KRR represents the covariates through a kernel feature map, regression trees represent them through split-induced regions of $\mathcal{Z}$.

In analogy to scalar-valued random forests, RKHS-valued random forests are bagged ensembles of binary RKHS-valued regression trees. Each tree is parametrised by a set of node-wise split rules of the covariates, which recursively partition $\mathcal{Z}$ into disjoint regions. For any $\mathcal{I} \subseteq \{1, ..., n\}$, denote $\Phi_{\mathcal{I}} := \{\phi(x_i)\}_{i\in\mathcal{I}}$. The splits are chosen to maximise the reduction in the variance of the embeddings,

$$\mathrm{Var}(\Phi_{\mathcal{S}}) - \frac{|\mathcal{S}_l|}{|\mathcal{S}|}\mathrm{Var}(\Phi_{\mathcal{S}_l}) - \frac{|\mathcal{S}_r|}{|\mathcal{S}|}\mathrm{Var}(\Phi_{\mathcal{S}_r}),$$

where $\mathcal{S}$ is the index set of the subsample at the current split node and $\mathcal{S}_l$ and $\mathcal{S}_r$ are the corresponding left and right child nodes. In Output Kernel Random Forests (Geurts et al., 2006) the variance is defined as

$$\mathrm{Var}(\Phi_{\mathcal{S}}) := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \|\phi(x_i) - \hat{\mu}_{\mathcal{S}}\|^2_{\mathcal{F}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} k(x_i, x_i) - \frac{1}{|\mathcal{S}|^2}\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}} k(x_i, x_j), \tag{5}$$

where $\hat{\mu}_{\mathcal{S}} := |\mathcal{S}|^{-1}\sum_{i\in\mathcal{S}}\phi(x_i)$. This criterion is evaluated by the kernel trick, while Distributional Random Forests (Cevid et al., 2022) use a related criterion, in which the feature vectors in (5) are approximated by Random Fourier Features to accelerate learning by avoiding the double sum in the kernel expression. Given m trained regression trees, the random forest model can be represented by

$$\hat{F}_n(z) = \sum_{i=1}^{n} \tilde{w}_i(z)\phi(x_i), \quad \tilde{w}_i(z) := \frac{1}{m}\sum_{j=1}^{m}\frac{\mathbb{1}\{z_i \in \mathcal{L}_j(z)\}}{|\mathcal{L}_j(z)|} \in [0, 1],$$

where $\mathcal{L}_j(z) \subseteq \{z_1, \ldots, z_n\}$ denotes the subset of observations falling into the same terminal node or leaf as z in the jth tree.

5. Simulation study

To illustrate the finite-sample performance of GKCM regarding level and power, we conduct a simulation study where we compare our method to GCM, wGCM and PCM, and to KCIT, RCIT and RCoT. The code is publicly available at GitHub.[2] Since KRESIT is analogous to GKCM when using KRR (modulo computation of the p-value), we expect their performance to be very similar. Therefore we exclude KRESIT from the comparison.

Methods and parameters To investigate the impact of different regression methods in kernel-based testing, we include GKCM using RKHS-valued random forests (i.e., Distributional Random Forests from the `R`-package `drf`; Michel and Cevid, 2021) and using KRR. We refer to them by GKCM RF and GKCM KRR, respectively. Denoting the number of conditioning variables by p, we set `num.trees` $= p \times 100$, `mtry` $= p$, and `min.node.size` $= 5$ to ensure that the random forests are sufficiently flexible. The hyperparameter values for KRR are described below.

For GCM, wGCM and PCM, we use the `R`-package `comets` (Kook, 2025) with random forests auto-tuned on `max.depth` $\times$ `mtry` with 500 trees for all regressions. This corresponds to using general-purpose, flexible regression models (as opposed to domain-informed regression models), which we consider to be a realistic use case for the residual-based tests in practice.

RCIT, RCoT and KCIT are self-implemented for comparability. For all kernel ridge regressions (including GKCM KRR) we use a Gaussian input kernel on Z, set the lengthscale via the median heuristic (Garreau et al., 2018), and set $\lambda = 10^{-3}/n$. For KCIT, we use a U-statistic and the wild bootstrap as described in He et al. (2025, Appendix H.1), yet without using sample-splitting.

Data-generating processes We consider seven scenarios, four under the null and three under the alternative, with continuous X, Y, Z and varying sample sizes ($n \in \{500, 1000, 1500, 2000\}$). In each setting we sample i.i.d. errors $(\varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_2(0, \mathrm{I}_2)$ and conditioning variables $Z = (Z_1, \ldots, Z_7) \sim \mathcal{N}_7(0, \mathrm{I}_7)$. The null settings include linear main effects, conditional means and variances with complex functional dependencies, post-nonlinear models (Zhang and Hyvärinen, 2009), and a strong correlation between X and Y. The results are shown in Figure 1.

Null 1: $X = 0.4Z_1 + 0.5Z_2 + 0.6Z_3 - 0.7Z_4 + Z_7 + \varepsilon_X,$
$Y = 0.6Z_1 - 0.2Z_2 + 0.3Z_4 + 0.9Z_5 - 0.5Z_6 + \varepsilon_Y$

Null 2: $X = 0.5Z_1 - 0.9Z_2 + 0.4Z_3^2 + Z_4 Z_5 \varepsilon_X,$
$Y = -0.8Z_4 + Z_5^2 + \exp(Z_6) + \sin(2\pi Z_7)\varepsilon_Y$

Null 3: $X = \tanh(0.5Z_1 - 0.9Z_2 + Z_3 + \varepsilon_X),$
$Y = \exp(-0.8Z_4 Z_5 + 0.6Z_6 Z_7 + \varepsilon_Y)$

Null 4: $X = \sin(2\pi Z_1) + 0.1\varepsilon_X,$
$Y = \sin(2\pi Z_1) + \varepsilon_Y$

The alternative settings include a small linear effect of X on Y, a non-linear effect interacting with a covariate, and a non-linear effect on the variance. Settings 2 and 3 satisfy $\mathbb{E}_Z[\mathrm{Cov}(X, Y \mid$

2. https://github.com/lucabergen/GKCM

$Z)] = 0$ and setting 3 satisfies $\mathrm{Cov}(X, Y \mid Z) = 0$, which implies that the conditional dependencies in these two groups cannot be identified by the GCM and wGCM, respectively. The results are shown in Figure 2.

Alt. 1: $X = 0.7Z_1 + Z_2 + \varepsilon_X$,
$Y = 0.4Z_3 - 0.2Z_4 - 0.1X + \varepsilon_Y$

Alt. 2: $X = \sin(Z_1) + \varepsilon_X$,
$Y = \tanh(Z_2) + 0.4X^2 Z_3 + \varepsilon_Y$

Alt. 3: $X = 0.2Z_2^3 + \tanh(Z_4) + \varepsilon_X$,
$Y = \sin(\pi Z_1) - 0.4Z_2^2 + \cos(0.2\pi X)\varepsilon_Y$

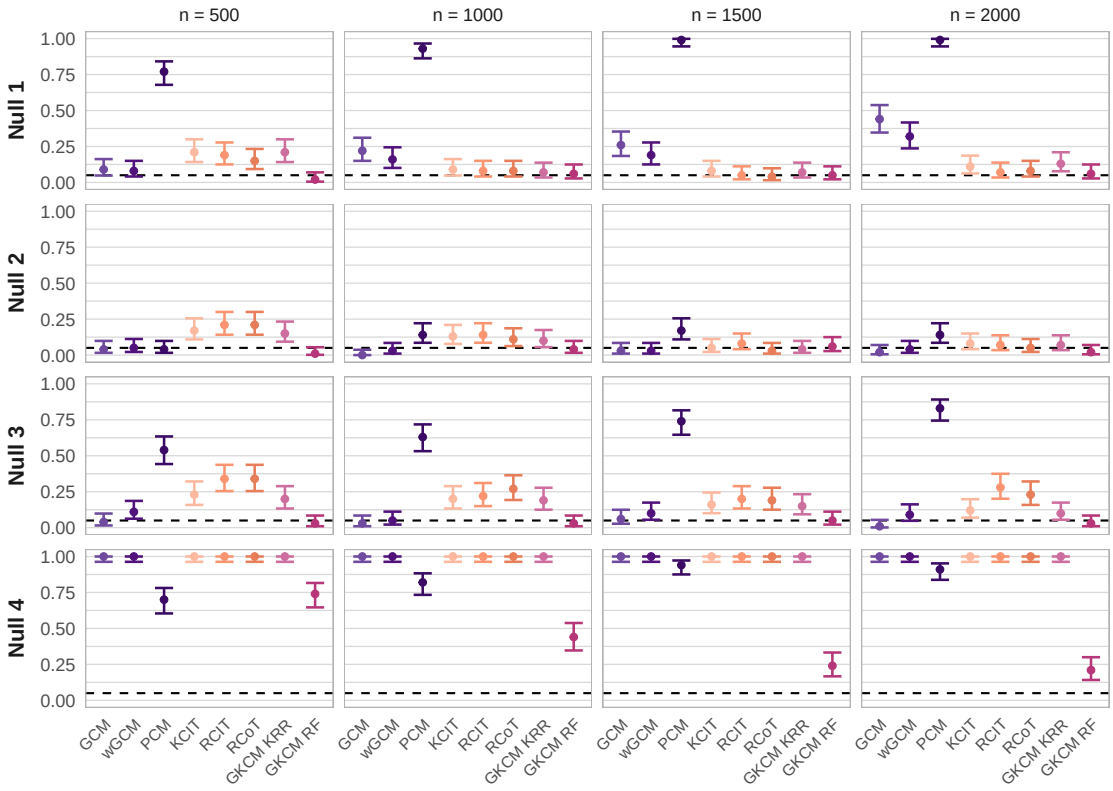

Figure 1: Rejection rates in the null settings with rejection threshold $p < 0.05$ (100 iterations). Error bars indicate 95% Wilson confidence intervals and dashed lines the nominal level.

Results In the null settings 1 – 3 GKCM RF has approximately nominal type-I error rates at each sample size, while all other tests have significantly inflated type-I error rates in multiple scenarios or sample sizes. In the challenging null setting 4 by Lundborg et al. (2024) GKCM RF approaches the nominal error rates with increasing sample size, while most other methods consistently reject the null hypothesis even at the highest sample size.

In the alternative setting 1 GKCM RF has below average power, yet remains competitive with the other kernel-based tests. KCIT and GKCM KRR show comparable power to GKCM RF despite

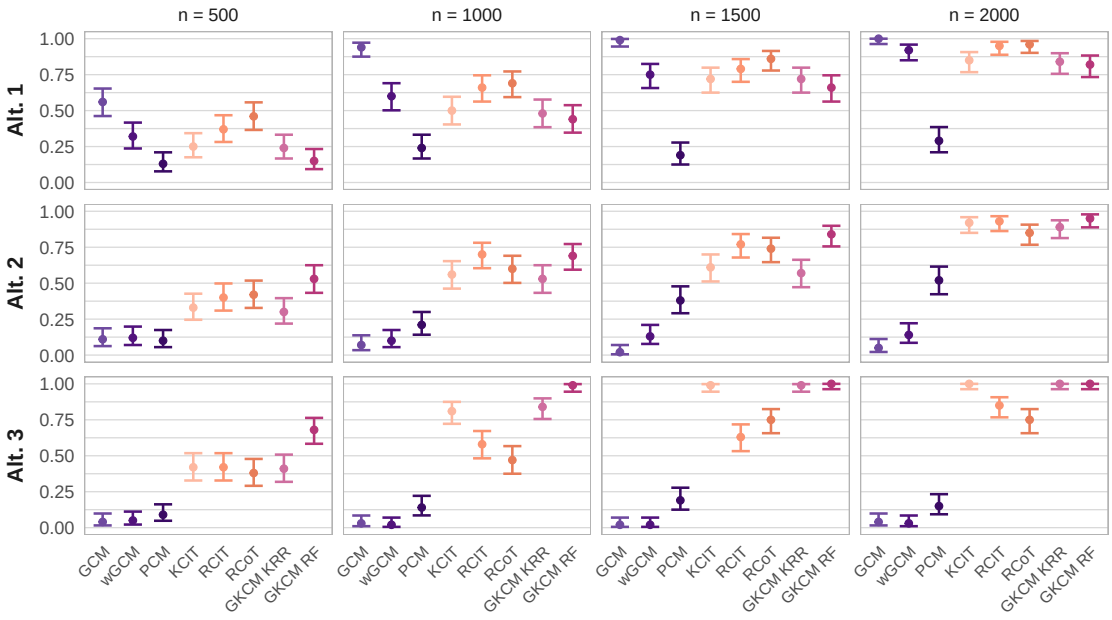

Figure 2: Rejection rates in the alternative settings with rejection threshold $p < 0.05$ (100 iterations). Error bars indicate 95% Wilson confidence intervals.

having inflated type-I error rates in the linear null setting 1. In the nonlinear settings 2 and 3 GKCM RF has power greater than or equal to all other methods. GKCM RF thereby has considerable power under non-linear dependence and good level control across a diverse range of settings. These results indicate that random forests offer a promising alternative to KRR in kernel-based CI testing when hyperparameter tuning is not feasible. Additional simulations, including hyperparameter tuning, are provided in Appendix C.

6. Conclusion

We introduced GKCM, established its theoretical properties and assessed its finite-sample performance using RKHS-valued random forests through a simulation study. GKCM frequently outperforms state-of-the-art alternatives, which shows its potential as a general CI test for statistical and causal inference.

However, several questions related to the practical use of GKCM remain open. Important directions include a systematic evaluation of additional regression methods, an investigation of whether computationally feasible hyperparameter tuning strategies can further improve performance, and an extension of the empirical study to settings with mixed data types.

Furthermore, we provided sufficient conditions for uniform asymptotic type-I error guarantees. An additional direction of future work is to establish examples of families of distributions for which these conditions can be shown to hold.

Acknowledgments

LB was funded by the German Research Foundation (DFG) as part of the Research Unit "Lifespan AI: From Longitudinal Data to Lifespan Inference in Health" (DFG FOR 5347), Grant 459360854, and thanks Anton Rask Lundborg for helpful discussions. DS was supported in part by the Responsible AI Research Centre (RAIR).

References

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, New Jersey, third edition, 2003. ISBN 978-0-471-36091-9.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces In Probability and Statistics*. Kluwer Academic Publishers, 2004. ISBN 978-1-4020-7679-4.

Dean A. Bodenham and Niall M. Adams. A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing*, 26(4):917–928, July 2016. URL https://doi.org/10.1007/s11222-015-9583-4.

Céline Brouard, Marie Szafranski, and Florence d'Alché Buc. Input Output Kernel Regression: Supervised and Semi-Supervised Structured Output Prediction with Operator-Valued Kernels. *Journal of Machine Learning Research*, 17(176):1–48, 2016. ISSN 1533-7928. URL http://jmlr.org/papers/v17/15-602.html.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022. URL http://jmlr.org/papers/v23/21-0585.html.

Donald L. Cohn. *Measure Theory: Second Edition*. Birkhäuser Advanced Texts Basler Lehrbücher. Springer, New York, NY, 2013. ISBN 978-1-4614-6955-1. URL https://link.springer.com/10.1007/978-1-4614-6956-8.

Panayiota Constantinou and A. Philip Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618–2653, December 2017. URL https://doi.org/10.1214/16-AOS1537. Publisher: Institute of Mathematical Statistics.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2 edition, 2014. ISBN 978-1-107-05584-1. URL https://doi.org/10.1017/CBO9781107295513.

Jean-Jacques Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980. URL https://doi.org/10.1093/biomet/67.3.581.

A. Philip Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, September 1979. URL https://academic.oup.com/jrsssb/article/41/1/1/7027599.

Tamim El Ahmad, Junjie Yang, Pierre Laforgue, and Florence d'Alché Buc. Deep Sketched Output Kernel Regression for Structured Prediction. In Albert Bifet, Jesse Davis, Tomas Krilavičius, Meelis Kull, Eirini Ntoutsi, and Indrė Žliobaitė, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 93–110, Cham, 2024. Springer Nature Switzerland. URL https://doi.org/10.1007/978-3-031-70352-2_6.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic, October 2018. URL http://arxiv.org/abs/1707.07269. arXiv:1707.07269 [math].

Pierre Geurts, Louis Wehenkel, and Florence d'Alché Buc. Kernelizing the Output of Tree-Based Methods. In *Proceedings of the 23nd International Machine Learning Conference*, ICML'06, pages 345–352, Pittsburgh, Pennsylvania, 2006. Association for Computing Machinery. URL https://doi.org/10.1145/1143844.1143888.

Pierre Geurts, Louis Wehenkel, and Florence d'Alché Buc. Gradient Boosting for Kernelized Output Spaces. In *Proceedings of the 24th international conference on Machine learning*, ICML'07, pages 289–296, New York, NY, USA, June 2007. Association for Computing Machinery. doi: 10.1145/1273496.1273533. URL https://doi.org/10.1145/1273496.1273533.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, June 2019. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00524/full. Publisher: Frontiers.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimilano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1803–1810, Madison, WI, USA, June 2012. Omnipress.

Zheng He, Roman Pogodin, Yazhe Li, Namrata Deka, Arthur Gretton, and Danica J. Sutherland. On the Hardness of Conditional Independence Testing In Practice. In *Advances in Neural Information Processing Systems 39*, San Diego, California, 2025. NeurIPS.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant Causal Prediction for Nonlinear Models. *Journal of Causal Inference*, 6(2), September 2018. URL https://doi.org/10.1515/jci-2017-0016.

Tailen Hsing and Randall L. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, 2015. ISBN 978-1-118-76257-8. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781118762547.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer International Publishing, Cham, 2021. ISBN 978-3-030-61871-1. URL https://link.springer.com/10.1007/978-3-030-61871-1.

Lucas Kook. *comets: Covariance Measure Tests for Conditional Independence*, 2025. URL https://CRAN.R-project.org/package=comets. R package version 0.2-2.

Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, and Florence D'Alché-Buc. Duality in RKHSs with Infinite Dimensional Outputs: Application to Robust Losses. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5598–5607. PMLR, November 2020. URL https://proceedings.mlr.press/v119/laforgue20a.html.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal Learning Rates for Regularized Conditional Mean Embedding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 4433–4445, Red Hook, NY, USA, November 2022. Curran Associates Inc. URL https://doi.org/10.52202/068431-0320.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Towards Optimal Sobolev Norm Rates for the Vector-Valued Regularized Least-Squares Algorithm. *J. Mach. Learn. Res.*, 25 (1):8554–8604, 2024. URL https://www.jmlr.org/papers/volume25/23-1663/23-1663.pdf.

Anton Rask Lundborg, Rajen D. Shah, and Jonas Peters. Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5):1821–1850, 2022. URL https://doi.org/10.1111/rssb.12544.

Anton Rask Lundborg, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52 (6):2851–2878, December 2024. URL https://doi.org/10.1214/24-AOS2447. Publisher: Institute of Mathematical Statistics.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7512–7523. PMLR, July 2021. URL https://proceedings.mlr.press/v139/mastouri21a.html.

Dimitri Meunier, Zikai Shen, Mattes Mollenhauer, Arthur Gretton, and Zhu Li. Optimal Rates for Vector-Valued Spectral Regularization Learning Algorithms. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, pages 82514–82559, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Loris Michel and Domagoj Cevid. *drf: Distributional Random Forests*, 2021. URL https://CRAN.R-project.org/package=drf. R package version 1.1.0.

Junhyung Park and Krikamol Muandet. A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f340f1b1f65b6df5b5e3f94d95b11daf-Paper.pdf.

Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University

Press, Cambridge, 2016. ISBN 978-1-107-10409-9. URL https://doi.org/10.1017/CBO9781316219232.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, November 2016. URL https://doi.org/10.1111/rssb.12167.

Roman Pogodin, Antonin Schrab, Yazhe Li, Danica J. Sutherland, and Arthur Gretton. Practical Kernel Tests of Conditional Independence, September 2025. URL http://arxiv.org/abs/2402.13196. arXiv:2402.13196 [cs].

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

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

Rajen D. Shah and Jonas Peters. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. Publisher: Institute of Mathematical Statistics.

Eiki Shimizu, Kenji Fukumizu, and Dino Sejdinovic. Neural-kernel conditional mean embeddings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 45040–45059, Vienna, Austria, July 2024. JMLR.org. URL https://dl.acm.org/doi/10.5555/3692070.3693903.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, New York, NY, 2008. ISBN 978-0-387-77241-7. URL https://link.springer.com/10.1007/978-0-387-77242-4.

Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate Kernel-Based Conditional Independence Tests for Fast Non-Parametric Causal Discovery. *Journal of Causal Inference*, 7 (1), March 2019. URL https://doi.org/10.1515/jci-2018-0017. Publisher: De Gruyter.

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

Haizhang Zhang and Liang Zhao. On the Inclusion Relation of Reproducing Kernel Hilbert Spaces, June 2011. URL http://arxiv.org/abs/1106.4075. arXiv:1106.4075 [math].

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 647–655, Arlington, Virginia, USA, June 2009. AUAI Press. URL https://dl.acm.org/doi/10.5555/1795114.1795190.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Barcelona, Spain, 2011. AUAI Press. URL https://webdav.tuebingen.mpg.de/causality/UAI11_KCItest.pdf.

Qinyi Zhang, Sarah Filippi, Seth Flaxman, and Dino Sejdinovic. Feature-to-Feature Regression for a Two-Step Conditional Independence Test. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, volume 33, Sydney, Australia, 2017. AUAI Press. URL https://www.auai.org/uai2017/proceedings/papers/250.pdf.

Appendix A. Proof of Lemma 1

By Cohn (2013, Lemma 8.3.8), there exists a Borel-measurable function $g : \mathcal{F} \to \mathcal{X}$ such that $(g \circ \phi)(x) = x$ for each $x \in \mathcal{X}$. Thereby $X = (g \circ \phi)(X)$ and $\sigma(X) = \sigma((g \circ \phi)(X))$. Since g and ϕ are Borel-measurable, it holds that $\sigma((g \circ \phi)(X)) = \sigma(X) \subseteq \sigma(\phi(X))$ and $\sigma(\phi(X)) \subseteq \sigma(X)$.

Appendix B. CI testing with joint embeddings

In the following, we review the use of joint embeddings in kernel-based CI testing. Two different approaches have been proposed. The first regresses the joint embedding of (X, Z) directly on Z. The second exploits a decomposition in which only the embedding of X is regressed on Z, while the Z-embedding is incorporated afterwards. Although the latter avoids some of the difficulties of the former, both methods encounter distinct statistical and practical issues, which we discuss below.

The direct approach is used in the original definition of the KCIT (Zhang et al., 2011). To construct the joint embedding, Zhang et al. combine the Gaussian kernel k with a (possibly differently scaled) Gaussian kernel $m' : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ with RKHS $\mathcal{H}'$ and canonical feature map $\psi'(z) := m'(\cdot, z)$. They use the Gaussian tensor-product kernel

$$k'\big((x, z), (x', z')\big) := k(x, x')m'(z, z')$$

whose RKHS is the completed Hilbert tensor product $\mathcal{F} \otimes \mathcal{H}'$ with canonical feature map

$$\phi'(x, z) := \phi(x) \otimes \psi'(z)$$

(Paulsen and Raghupathi, 2016, Section 5.5).

However, direct regression of $\phi'(X, Z)$ is generally problematic: Not only is the output space large, but the regression function additionally has the form $z \mapsto F_P(z) \otimes \psi'(z)$. The regression method thereby needs to simultaneously learn $z \mapsto F_P(z)$ and reconstruct $z \mapsto \psi'(z)$, even though the latter is deterministic and fully known. Furthermore, the latter part of the problem behaves like

learning an identity operator between infinite-dimensional RKHSs, which is typically not Hilbert-Schmidt (Pogodin et al., 2025, Appendix B.2). As a consequence, the regression function is misaligned with the hypothesis spaces of KRR (cf. Section 4.1). This phenomenon is illustrated by Mastouri et al. (2021, Appendix B.9), who show that regularisation leads to considerable shrinkage especially in data-sparse regions. The resulting bias is an important source of Type-I error inflation in the original KCIT.

To avoid these difficulties, Pogodin et al. (2025) propose to modify KCIT by regressing the joint embedding indirectly. Since $\psi'(Z)$ is $\sigma(Z)$-measurable, it can be pulled out of the conditional mean embedding:

$$\mathbb{E}[\phi'(X,Z) \mid Z] = \mathbb{E}[\phi(X) \otimes \psi'(Z) \mid Z] = \mathbb{E}[\phi(X) \mid Z] \otimes \psi'(Z)$$

By the bilinearity of the tensor product map, the residual then factorises as

$$\phi'(X,Z) - \mathbb{E}[\phi'(X,Z) \mid Z] = \big(\phi(X) - \mathbb{E}[\phi(X) \mid Z]\big) \otimes \psi'(Z).$$

Accordingly, one estimates the $\mathcal{F}$-valued residual $\phi(X) - \mathbb{E}[\phi(X) \mid Z]$ and subsequently tensors it with $\psi'(Z)$. This removes the identity-like component from the regression and thereby eliminates a major source of type-I error inflation in the original KCIT.

Additionally, the decomposition has the benefit that the regression-rate requirements of the GHCM framework are not strengthened by using the joint embedding. If we define the joint conditional mean embedding function $F'_P(z) := F_P(z) \otimes \psi'(z)$ and its estimate by $\hat{F}'_n(z) := \hat{F}_n(z) \otimes \psi'(z)$, then the in-sample MSPE

$$\mathcal{E}_n^{F'} := \frac{1}{n}\sum_{i=1}^{n} \|F'_P(Z_i) - \hat{F}'_n(Z_i)\|^2_{\mathcal{F}\otimes\mathcal{H}'}$$

satisfies

$$\mathcal{E}_n^{F'} = \frac{1}{n}\sum_{i=1}^{n} \|(F_P(Z_i) - \hat{F}_n(Z_i)) \otimes \psi'(Z_i)\|^2_{\mathcal{F}\otimes\mathcal{H}'} = \frac{1}{n}\sum_{i=1}^{n} \|F_P(Z_i) - \hat{F}_n(Z_i)\|^2_{\mathcal{F}} m'(Z_i, Z_i),$$

since $\|f \otimes h'\|^2_{\mathcal{F}\otimes\mathcal{H}'} = \|f\|^2_{\mathcal{F}}\|h'\|^2_{\mathcal{H}'}$ for simple tensors and $\|\psi'(z)\|^2_{\mathcal{H}'} = m'(z,z)$. For a Gaussian kernel normalised so that $m'(z,z) = 1$, $\mathcal{E}_n^{F'}$ and $\mathcal{E}_n^{F}$ thereby coincide.

The decomposed joint embedding does, however, modify the test statistic through off-diagonal weights, and may thereby introduce a new source of finite-sample type-I errors. If we define the (uncentered) joint residuals by $\hat{\varepsilon}'_i := (\phi(X_i) - \hat{F}_n(Z_i)) \otimes \psi'(Z_i)$, then

$$\langle \hat{\varepsilon}'_i, \hat{\varepsilon}'_j \rangle_{\mathcal{F}\otimes\mathcal{H}'} = m'(Z_i, Z_j)\langle \phi(X_i) - \hat{F}_n(Z_i), \phi(X_j) - \hat{F}_n(Z_j)\rangle_{\mathcal{F}},$$

so replacing $\hat{\varepsilon}_i$ by $\hat{\varepsilon}'_i$ is equivalent to inserting the kernel weights $m'(Z_i, Z_j)$ and the corresponding centering corrections into the quadratic form defining the test statistic T_n. This reweighting does not change the form of the null distribution asymptotically when the regression errors satisfy the in-sample error conditions. In finite samples, however, there are typically non-negligible regression errors, and the weights can make their effect on the test statistic more pronounced. This problem is highlighted theoretically and empirically by He et al. (2025, Section 5) for a version of the KCIT based on a U-statistic and the wild bootstrap: they show that using a joint embedding can

improve power under the alternative, but also amplify "leaked dependence" under the null. This phenomenon may be especially pronounced when tuning m' to maximise power. While leaked dependence should become negligible as the regression accuracy improves with growing sample size, He et al. demonstrate that it can remain practically relevant at low and moderate sample sizes at which kernel-based CI tests are typically used. Although He et al. only establish this phenomenon for tests based on U-statistics, the underlying mechanism also suggests potential finite-sample distortions for GHCM-style tests.

Appendix C. Additional simulation results

In the following, we include additional simulations. In Section C.1, we replicate the simulation study of Zhang et al. (2011) using the methods and hyperparameter settings as described in the main paper. In Section C.2, we repeat the simulation from the main paper with alternative hyperparameter settings for the regression methods of the tests we compare to GKCM RF.

C.1. Simulation study from Zhang et al. (2011)

The simulation study comprises four scenarios (cases I and II, each under the null and alternative) and considers varying sample sizes ($n \in \{200, 400\}$) and numbers of conditioning variables ($d \in \{1, \dots, 5\}$). These four data-generating processes differ along two dimensions. First, cases I and II differ in how X and Y depend on the covariates Z, in particular in how many components of Z are relevant. In case I, both X and Y depend only on Z_1, while the remaining covariates are irrelevant. In case II, X and Y depend on all components of $Z = (Z_1, \dots, Z_d)$. Second, the null and the alternative differ in whether there is an unobserved additive common cause of X and Y: under the null, no latent common cause is added, whereas under the alternative, an additional latent common cause is added to both variables.

Let $(\varepsilon_X, \varepsilon_Y) \sim \mathcal{N}_2(0, \mathrm{I}_2)$, $Z = (Z_1, \dots, Z_d) \sim \mathcal{N}_d(0, \mathrm{I}_d)$, and $C \sim \mathcal{N}(0, 0.25)$ be mutually independent. For each fixed scenario, the random vectors are independent across observations and the Monte Carlo iterations are independent of each other. However, within a given iteration, the datasets generated for different values of d are not independent, since those for larger d reuse and extend quantities generated for smaller d. To simplify notation, define

$$\phi(u) = u + \frac{u^3}{3} + \frac{1}{2}\tanh\left(\frac{u}{3}\right), \qquad \psi(v) = v + \tanh\left(\frac{v}{3}\right), \qquad h(u) = \frac{u}{2} + 0.7\tanh(u).$$

Additionally, let

$$f_1(z) = 0.7\left(\frac{z_1^3}{5} + \frac{z_1}{2}\right), \qquad g_1(z) = \frac{z_1^3/4 + z_1}{3}.$$

For $j = 2, \dots, d$, define recursively

$$f_j(z) = h(a_j f_{j-1}(z) + b_j z_j), \qquad g_j(z) = h(a_j g_{j-1}(z) + b_j z_j),$$

where

$$a_j = \begin{cases} \frac{1}{2}, & j = 2, \\ \frac{2}{3}, & j \geq 3, \end{cases} \qquad b_j = \begin{cases} 1, & j = 2, \\ \frac{5}{6}, & j \geq 3. \end{cases}$$

The scenarios are defined as follows.

Null and alternative. Under the null

$$X = \phi\big(f(Z) + \tanh(\varepsilon_X)\big), \qquad Y = \psi\big(g(Z) + \varepsilon_Y\big).$$

Under the alternative

$$X = \phi\big(f(Z) + \tanh(\varepsilon_X)\big) + C, \qquad Y = \psi\big(g(Z) + \varepsilon_Y\big) + C.$$

Cases I and II. In case I, the functions $f(Z)$ and $g(Z)$ depend only on the first covariate:

$$f(Z) = f_1(Z), \qquad g(Z) = g_1(Z).$$

In case II, the functions are built recursively from all d covariates:

$$f(Z) = f_d(Z), \qquad g(Z) = g_d(Z).$$

While these equations describe the structural form of the data-generating process, in the implemented simulation X, Y, and Z are standardised before adding the common cause C under the alternative.

The results are shown in Figure 3. In case I, GKCM RF has excellent power and type I error rate, almost consistently outperforming the other methods. In case II, like the other tests, GKCM RF has slightly inflated type-I error rates for larger conditioning sets, yet is still able to outperform most other methods in many settings both with regard to level and power. The additional simulation thereby confirms the good performance of GKCM RF reported in the main text.

C.2. New hyperparameter settings

In order to investigate the impact of the hyperparameter settings on the regression methods, we repeat the simulation study from the main paper with alternative, potentially improved settings for the regression methods of the competing tests. For all kernel-based methods using kernel ridge regression we used the same settings as in the main paper except for the regularisation parameter λ, which we tuned via leave-one-out cross-validation over the set $\{10^{-5}/n, ..., 10^{3}/n\}$ in each iteration (using a subsample of maximum size 1000). The aim is to compare the RKHS-valued regression methods for settings in which light parameter tuning for kernel ridge regression is feasible. For the random forests employed in the residual-based methods, we use the same fixed parameter values as in the Distributional Random Forests (i.e., `num.trees` = 700, `mtry` = 7, and `min.node.size` = 5) in order to investigate the impact of the auto-tuning on their performance (note that these settings differ from the default fixed hyperparameter-settings used in the `comets` package).

The results for the null settings are shown in Figure 4. For the kernel-based methods using KRR, tuning λ notably improved the regressions in settings 1 and 2, resulting in approximately nominal type-I error rates for all methods and sample sizes. The type-I error rates also improved in setting 3, yet remain mildly inflated in some sample sizes. Only in setting 4 the performance did not improve.

The performance of the residual-based methods improved as well, especially for PCM, which had type-I error rates exceeding 0.5 throughout settings 1, 3, and 4 in the main text. Particularly in setting 4, PCM now stands out with approximately nominal type-I error rates for all sample sizes, thereby outperforming all other methods. For GCM and wGCM, the type-I error rates have improved in settings 1 and 4.

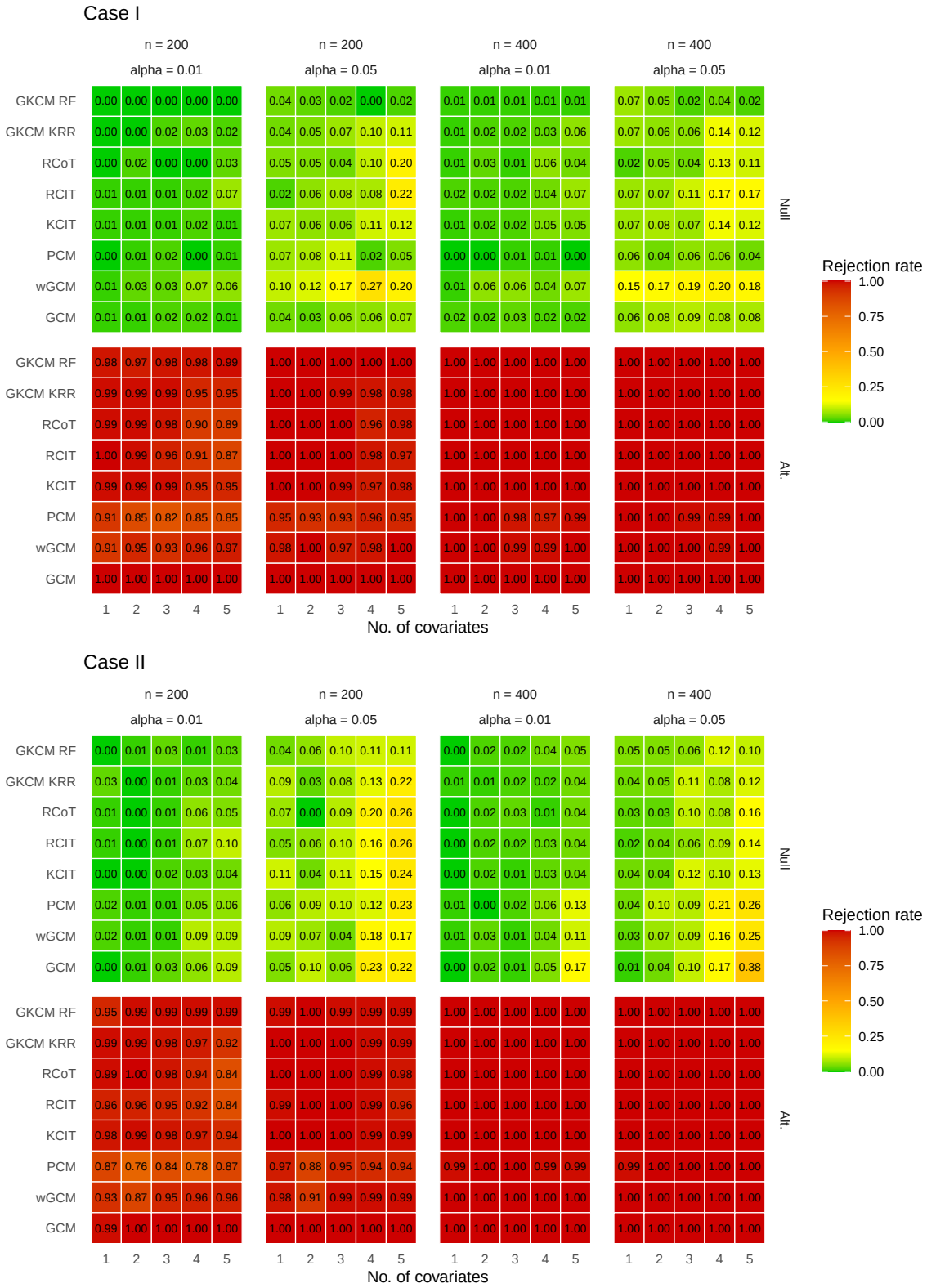

Figure 3: Rejection rates in the scenarios by Zhang et al. (2011) (100 iterations).

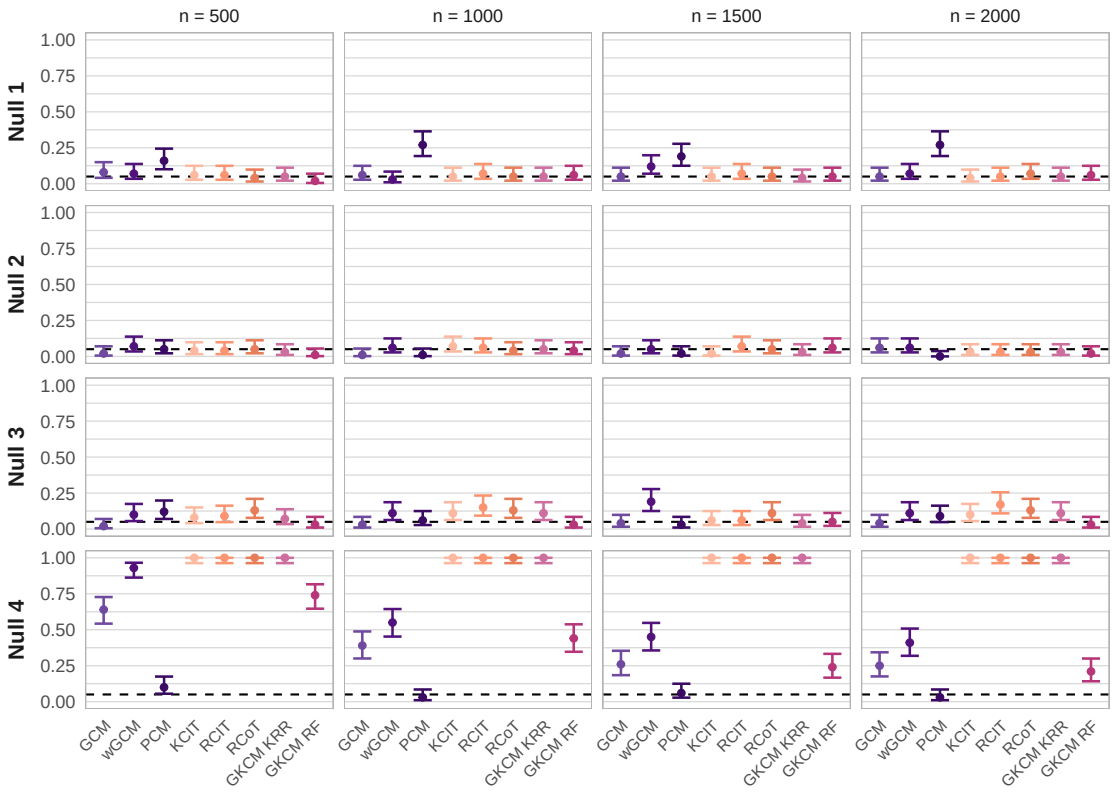

Figure 4: Rejection rates in the null settings with rejection threshold $p < 0.05$ (100 iterations). Error bars indicate 95% Wilson confidence intervals and dashed lines the nominal level.

The results for the alternative settings are shown in Figure 5. Like the type-I error rates, the power of the kernel-based tests using KRR has become more similar to the power of GKCM RF as well; hence, for the former, the reduction in type-I error rates came at the price of a reduction in power. In particular, in settings 1 and 2, KCIT, GKCM KRR and GKCM RF now perform comparably (for most sample sizes with a small lead for the former). While RCIT and RCoT had superior power in the main paper, they now have lower power than KCIT, GKCM KRR, and GKCM RF in these settings. Except for setting 1, where GCM outperforms all other tests, the residual-based tests are outperformed by the kernel-based tests throughout.

We conclude that even with parameter tuning of λ, distributional random forests exhibit better level and comparable or superior power than KRR in kernel-based testing. Furthermore, while the new parameter settings improved the performance of the residual-based tests under the alternative, they are still often outperformed by the kernel-based tests.

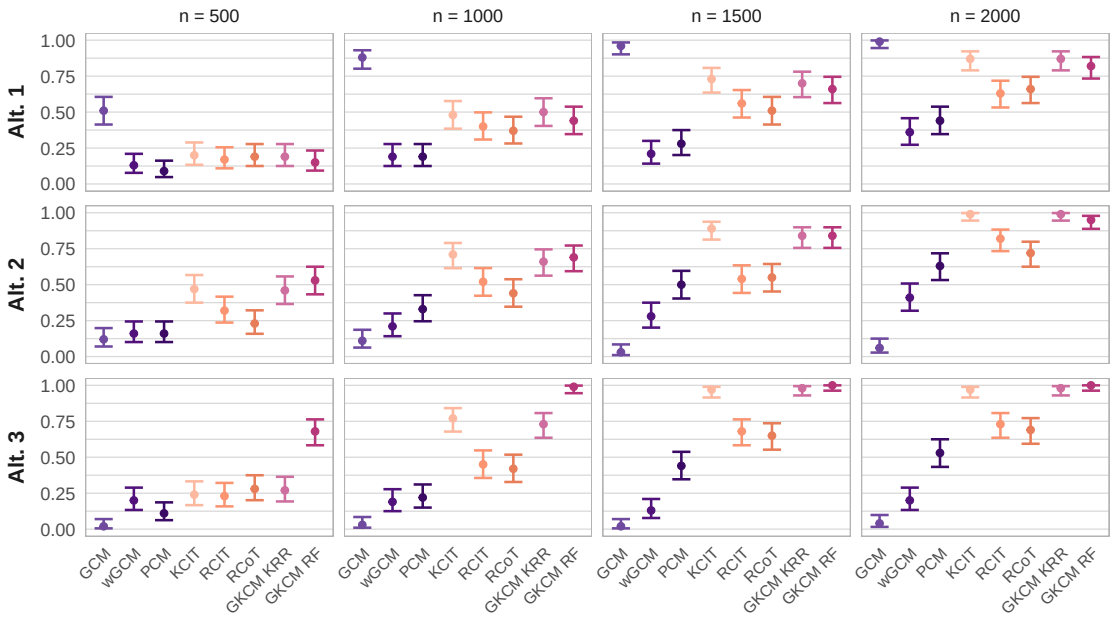

Figure 5: Rejection rates in the alternative settings with rejection threshold $p < 0.05$ (100 iterations). Error bars indicate 95% Wilson confidence intervals.