# KREDIT: HIGH-DIMENSIONAL CONDITIONAL INDEPENDENCE TESTING VIA KERNEL-RESIDUALIZED DISTANCES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Conditional independence testing underpins graphical model learning and causal discovery, yet high-dimensional inference with estimated conditional distance features remains challenging. To this end, we introduce KREDIT (Kernel-Residualized Distance Testing), a nonparametric framework combining two-sided distance residualization, independent conditional-mean fitting, and off-diagonal joint studentization. An exact pair-kernel analysis separates sampling variation from directional regression error. Under explicit moment, spectral, and prediction conditions, we prove feasible Gaussian limits, finite-sample approximation bounds, and dimension-dependent convergence rates. We further establish shrinking-signal power guarantees, empirical-variance inference under nondegenerate alternatives, and testing and estimation lower bounds in a rectangular Gaussian subexperiment. A specified nonlinear family remains detectable despite zero linear covariances. Simulation studies verify linear and nonlinear settings under which stratum-calibrated KREDIT has higher empirical power than the benchmark methods, with size near the nominal level. Wine chemistry and rolling gold–inflation analyses further illustrate its practical utility.

## 1 INTRODUCTION

Testing conditional independence asks whether two random vectors remain associated after accounting for observed covariates. Such question underlies graphical model learning, feature selection, and causal discovery. Distance correlation provides a flexible measure of multivariate association without specifying a regression model (Székely et al., 2007). Its conditional extension, however, requires both a well-defined population target and an inferential treatment of the conditioning step. In particular, estimation error that is negligible for prediction need not be negligible for a degenerate test statistic.

Such distinction becomes consequential in high dimensions. Increasing dimensionality can change the dependence features captured by distance and kernel statistics (Zhu et al., 2020), while also making Gaussian calibration possible through a diffuse spectrum (Gao et al., 2021). Conditioning introduces an additional source of error: the two conditional means underlying every pairwise distance must be learned. We thus seek a distance-based procedure whose sampling variation, regression error, and detectable signal can be analyzed together. Such an analysis necessarily specifies a distribution class; unrestricted conditional independence testing does not admit nontrivial universal validity (Shah & Peters, 2020).

### 1.1 MAJOR CONTRIBUTIONS

The major contributions of our work are fourfold, with theory and experiments developed for the same implemented statistic.

*First, we introduce KREDIT (Kernel-Residualized Distance Testing).* Two-sided distance residualization and off-diagonal evaluation produce a single dependence statistic without explicit feature coordinates. Independent conditional-mean fitting and joint studentization connect such construction to an implementable test, separating score normalization from inferential calibration.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

*Second, we establish quantitative null inference.* An exact pair-kernel analysis separates spectral spread, directional regression error, and empirical variance estimation. Finite-sample bounds yield a dimension-explicit feasible Gaussian rate, and an exact feature-mean derivative sharpens a nonconstant-conditioning example. Complementary multiplier and discrete-stratum results cover different spectral and sampling regimes, including a finite-first-moment model. Calibration experiments evaluate each implementation.

*Third, we develop signal estimation and detection theory.* Projection-relative bounds give consistency for shrinking alternatives and quantitative plug-in variance inference away from degeneracy. A nonlinear sign–radius family has an inverse-dimension signal despite zero linear covariances. A rectangular Gaussian subexperiment supplies two-sided distance-signal bounds and lower bounds for testing, estimation, and confidence-set length. Separate experiments examine nonlinear power and empirical variance estimation.

*Fourth, we provide reproducible empirical comparisons.* Simulation studies demonstrate linear and nonlinear designs in which stratum-calibrated KREDIT exceeds every benchmark construction in Table 1, with size near the nominal level. Complete paired uncertainty and computation times accompany these comparisons. Finite-sample distributions assess Gaussian calibration, while wine chemistry and six- and twelve-month gold–inflation analyses illustrate the stated significance calculations.

### 1.2 Related works

The construction connects two established lines of research. Distance covariance has motivated screening, variable selection, and interaction discovery (Kong et al., 2015; 2017), while kernel methods express association through covariance of feature maps (Gretton et al., 2005). Their relation is formalized by Sejdinovic et al. (2013). Conditional extensions are not interchangeable: partial distance correlation uses projection of distance matrices (Székely & Rizzo, 2014), conditional distance correlation employs localized centering (Wang et al., 2015), and projected dependence measures impose regression structure (Fan et al., 2020). Conditional kernel covariance and the distance–kernel comparison of Bai et al. (2024); Sheng & Sriperumbudur (2023) further distinguish these targets.

Residualization itself predates KREDIT. Conditional-mean-zero characterizations (Daudin, 1980), conditional kernel operators (Fukumizu et al., 2007), and the kernel conditional independence (KCI) test (Zhang et al., 2011) provide foundations. The generalised covariance measure (GCM), weighted GCM (WGCM), and generalised Hilbertian covariance measure (GHCM) analyze residual covariances (Shah & Peters, 2020; Scheidegger et al., 2022; Lundborg et al., 2022). Split kernel conditional independence (SplitKCI) controls explicitly regression-induced bias through splitting (Pogodin et al., 2025); recent work examines conditioning-kernel choice and practical calibration (He et al., 2025). The generalised kernel covariance measure (GKCM) permits a broad class of regressors (Bergen et al., 2026), fast kernel conditional independence (FastKCI) aggregates local tests (Schacht & Huang, 2025), and recent functional and sequential methods address different sampling regimes (Tang & Li, 2026; He & Sutherland, 2026). KREDIT's contribution is the distance-specific joint-growth analysis and its feasible implementation.

A complementary route changes evaluation instead of nuisance fitting. The cross Hilbert–Schmidt independence criterion (cross-HSIC) uses split evaluation for unconditional Gaussian calibration (Shekhar et al., 2023); KREDIT retains the pair statistic and requires spectral spread for its normal limit. Non-null Gaussian central limit theorems (CLTs), power universality, and minimax optimality for unconditional kernel distance covariance are established by Han & Shen (2024). Our conditional plug-in analysis and explicit rectangular signal calculation extend the comparison to independently fitted conditional means. Gaussian independence lower bounds (Ramdas et al., 2016) and smooth conditional-testing limits (Neykov et al., 2021) provide further decision-theoretic context.

Table 1 summarizes the methodological distinctions. In particular, DISCO develops scalable conditional-distance estimators for independence regularization in deep learning (Kavak et al., 2026), whereas our emphasis is inferential calibration and quantitative power. The integrated residual-covariance target studied here can vanish through cancellation over conditioning values; our power claims thus require a nonzero target, not just conditional dependence. Such distinction makes the comparison with pointwise conditional measures precise.

Table 1: Methodological comparison. Each distinction concerns the specified construction.

| Method or literature | Key distinction of KREDIT |
|---|---|
| HDCI; cross-HSIC (Gao et al., 2021; Shekhar et al., 2023) | KREDIT estimates conditional feature means, whereas HDCI uses U-centered unconditional distances and cross-HSIC changes the evaluation construction. |
| Partial distance correlation (Székely & Rizzo, 2014) | KREDIT estimates conditional feature means, instead of projecting onto the span of a centered conditioning-distance matrix. |
| Conditional distance correlation; DISCO (Wang et al., 2015; Kavak et al., 2026) | KREDIT uses one global residual-covariance target, as opposed to pointwise conditional centering and the local-correlation aggregation used by DISCO. |
| Projected dependence (Fan et al., 2020) | KREDIT residualizes distance features nonparametrically, instead of applying distance dependence to regression residuals under a factor-model formulation. |
| GCM; weighted GCM (Shah & Peters, 2020; Scheidegger et al., 2022) | KREDIT uses distance-feature tensor products, not only products of residualized original coordinates or their weighted versions. |
| KCI; SplitKCI (Zhang et al., 2011; Pogodin et al., 2025) | KREDIT uses distance-induced features and off-diagonal joint studentization, as opposed to the regularized kernel statistic and calibration developed for KCI. |
| GHCM; GKCM (Lundborg et al., 2022; Bergen et al., 2026) | KREDIT couples distance-induced residual features with off-diagonal joint studentization and dimension-explicit Gaussian theory, whereas GHCM and GKCM develop broader residual-covariance frameworks. |
| KREDIT (this paper) | Two-sided distance residualization, independent training, and joint studentization connect a single computable statistic to explicit null and signal-relative power guarantees. |

The rest of the paper is organized as follows. Section 2 presents the formulation and algorithm. Section 3 develops the theory, followed by simulation studies and real data applications in Sections 4 and 5, respectively. Section 6 concludes with some discussions. Appendices A–F provide assumptions, complete proofs, implementation details, and additional empirical results.

## 2 Problem formulation and KREDIT

### 2.1 Population target and conditional residualization

Let us write $U = (X, Y, Z) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r$, with dimensions $p, q, r$. Lowercase letters denote realizations, primes independent copies, $\mathbb{E}$ expectation, $\Pr$ probability, and $\|\cdot\|$ the Euclidean norm. For finite first moments, the marginally centered distance kernel is

$$g_X(x, x') = \|x - x'\| - \mathbb{E}\|x - X'\| - \mathbb{E}\|X - x'\| + \mathbb{E}\|X - X'\|. \tag{1}$$

For an integrable function $f$ of two observations, we define $\mathcal{E}_1 f(z, x') = \mathbb{E}[f(X, x') \mid Z = z]$, $\mathcal{E}_2 f(x, z') = \mathbb{E}[f(x, X') \mid Z' = z']$, and $\mathcal{E}_{12} f(z, z') = \mathbb{E}[f(X, X') \mid Z = z, Z' = z']$. These are argumentwise averaging operators. Applying inclusion–exclusion to (1) gives that

$$d_X(u, u') = g_X(x, x') - (\mathcal{E}_1 g_X)(z, x') - (\mathcal{E}_2 g_X)(x, z') + (\mathcal{E}_{12} g_X)(z, z'), \tag{2}$$

with $d_Y$ defined similarly. Thus both observations in a distance pair are adjusted for their respective conditioning values.

The distance-induced kernel gives a useful equivalent representation (Sejdinovic et al., 2013)

$$\begin{aligned} k_X(x, x') = \tfrac{1}{2}\{\|x\| + \|x'\| - \|x - x'\|\} = \langle \phi_X(x), \phi_X(x') \rangle, \\ r_X = \phi_X(X) - m_X(Z), \quad m_X(z) = \mathbb{E}[\phi_X(X) \mid Z = z]. \end{aligned} \tag{3}$$

Here, $\phi_X$ is a Hilbert-space feature map of $k_X$; norms and inner products use the corresponding space. Define $r_Y$ similarly and $t = r_X \otimes r_Y$, with $\otimes$ the Hilbert tensor product. Proposition 6 shows that

$$d_X = -2\langle r_X, r_X' \rangle, \quad H(u, u') = d_X(u, u') d_Y(u, u') = 4\langle t, t' \rangle, \quad \theta = \mathbb{E}H = 4\|\mathbb{E}t\|^2. \tag{4}$$

The population KREDIT score is $\rho = \theta / \sqrt{ab}$, where $a = \mathbb{E}d_X^2$ and $b = \mathbb{E}d_Y^2$ are positive and finite. Then it holds that $0 \le \rho \le 1$. Under the conditional-independence null $H_0 : X \perp\!\!\!\perp Y \mid Z$, $\mathbb{E}[t \mid Z] = 0$, so $\theta = 0$ and $H$ is canonical: $\mathbb{E}[H(U, U') \mid U] = 0$. Lemma 1 proves the stronger argumentwise centering needed for such conclusion.

### 2.2 ALGORITHM AND INFERENTIAL STATISTIC

We divide independent triples into training observations $\mathcal{T} = \{U_a^{\mathrm{tr}}\}_{a=1}^m$ and evaluation observations $\{U_i\}_{i=1}^n$, totaling $m+n$ observations; superscript $\mathrm{tr}$ marks training coordinates. For training-only fitted weights satisfying that $\sum_a v_a(z) = 1$, let us set

$$\widehat{m}_X(z) = \sum_{a=1}^m v_a(z)\phi_X(X_a^{\mathrm{tr}}), \quad \widehat{m}_Y(z) = \sum_{a=1}^m v_a(z)\phi_Y(Y_a^{\mathrm{tr}}). \tag{5}$$

For a nonnegative smoothing kernel $K$ and bandwidth $h > 0$, the local weights have the explicit form

$$v_a(z) = \frac{K((z - Z_a^{\mathrm{tr}})/h)}{\sum_{b=1}^m K((z - Z_b^{\mathrm{tr}})/h)}, \quad a = 1, \ldots, m, \tag{6}$$

whenever the denominator is positive. Such construction estimates conditional feature means instead of conditional distributions. The same matrix thus residualizes every pairwise distance without solving a separate regression for each evaluation pair. Training-only tuning leaves the evaluation sample independent of the fitted functions, although its prediction rate still has to satisfy the stated assumptions.

The two fits may share training observations. We employ the Nadaraya–Watson (NW) weights (Nadaraya, 1964; Watson, 1964); Appendix D specifies normalization, bandwidths, and empty-neighborhood handling. Independent evaluation, as opposed to independence between the two fits, is what the theory requires.

No feature coordinates need to be computed. Concatenate evaluation and training observations, let $A_X$ be their raw distance matrix, and denote by $V_{ia} = v_a(Z_i)$ and $L = [I_n, -V]$, where $I_n$ is the $n \times n$ identity matrix and $\top$ represents transpose. Proposition 7 provides the exact implementation

$$\widehat{D}_X = LA_XL^\top = A_X^{ee} - VA_X^{te} - A_X^{et}V^\top + VA_X^{tt}V^\top, \quad \widehat{D}_Y = LA_YL^\top. \tag{7}$$

Superscripts $e, t$ index evaluation/training blocks; $\mathbf{1}_k$ is the all-ones vector in $\mathbb{R}^k$ (dimension omitted when clear). Since $L\mathbf{1} = 0$, ordinary double-centering of $A_X$ leaves (7) unchanged. In particular, $\widehat{D}_{X,ij} = -2\langle \phi_X(X_i) - \widehat{m}_X(Z_i), \phi_X(X_j) - \widehat{m}_X(Z_j)\rangle$ exactly. Such identity connects implementation to the population construction in (4).

For $C_n = \binom{n}{2} = n(n-1)/2$, we retain only distinct evaluation pairs

$$\begin{aligned} &\widehat{H}_{ij} = \widehat{D}_{X,ij}\widehat{D}_{Y,ij}, \quad \widehat{U}_n = C_n^{-1}\sum_{i<j}\widehat{H}_{ij}, \quad \widehat{\sigma}_n^2 = C_n^{-1}\sum_{i<j}\widehat{H}_{ij}^2, \\ &\widehat{a}_n = C_n^{-1}\sum_{i<j}\widehat{D}_{X,ij}^2, \quad \widehat{b}_n = C_n^{-1}\sum_{i<j}\widehat{D}_{Y,ij}^2, \quad \widehat{\rho}_n = \widehat{U}_n/\sqrt{\widehat{a}_n\widehat{b}_n}. \end{aligned} \tag{8}$$

We define the KREDIT test statistic as

$$T_n = \frac{\sqrt{C_n}\widehat{U}_n}{\widehat{\sigma}_n} = \frac{\sqrt{C_n}\widehat{\rho}_n}{\sqrt{\widehat{\kappa}_n}}, \quad \widehat{\kappa}_n = \frac{\widehat{\sigma}_n^2}{\widehat{a}_n\widehat{b}_n} \tag{9}$$

with $\widehat{\sigma}_n$ the nonnegative square root of $\widehat{\sigma}_n^2$. Joint studentization accounts for variation shared by $d_X^2$ and $d_Y^2$ through $Z$, even under conditional independence: $\mathbb{E}(d_X^2 d_Y^2)$ need not equal $ab$. Negative scores are retained; a zero denominator means no rejection. For level $\alpha \in (0,1)$ and standard normal distribution function $\Phi$, the Gaussian rule rejects when $T_n > \Phi^{-1}(1-\alpha)$ under Theorem 1. Propositions 11 and 5 provide multiplier and exact-stratum alternatives.

With $O(\cdot)$ denoting asymptotic order, Algorithm 1 for KREDIT uses $O((m+n)^2(p+q))$ operations for distances and $O(nm^2 + n^2m)$ for dense residualization, with $O((m+n)^2)$ storage. Normal calibration adds $O(n^2)$ work; $B$ independent multiplier draws add $O(Bn^2)$. Appendix D specifies exact block reuse and all benchmark comparators.

An empty conditioning set is handled directly by HDCI (Gao et al., 2021). U-center the evaluation distance matrices as $G_X = \mathcal{U}(A_X)$ and $G_Y = \mathcal{U}(A_Y)$ using (D.1). Its signed statistic adopts $\langle A, B\rangle_F = \sum_{ij} A_{ij}B_{ij}$ and $\|A\|_F^2 = \langle A, A\rangle_F$

$$T_{\mathrm{HDCI}} = \sqrt{\frac{n(n-1)}{2}}\,\frac{\langle G_X, G_Y\rangle_F}{\|G_X\|_F\|G_Y\|_F}. \tag{10}$$

**Algorithm 1** KREDIT with independent training and joint studentization

**Require:** Training observations, evaluation observations, level $\alpha$, and a training-only bandwidth rule.

1: Fit weights $V_{ia} = v_a(Z_i)$, normalizing each row; use uniform weights for an empty neighborhood.
2: Form distance blocks for $X$ and $Y$ and compute (7).
3: Exclude diagonal evaluation pairs and compute (8).
4: Return $\widehat{\rho}_n$, $T_n$, and $1 - \Phi(T_n)$; optionally use (C.90) for multiplier calibration.
5: Do not reject when $\widehat{\sigma}_n = 0$; otherwise apply the chosen upper-tail critical value.

Equation (10) uses the original bias-corrected distance correlation directly, without training, a square root of its signed numerator, or KREDIT's joint normalizer.

## 3 THEORY

All technical assumptions are detailed in Appendix A. Sections 3.1 and 3.2 address calibration and power; Section 3.3 establishes a testing lower bound. We allow a triangular array of laws $P_n$ with increasing tested dimensions and fixed conditioning dimension. We write $\|F\|_k = (\mathbb{E}\|F\|^k)^{1/k}$ for an $L^k$ norm, conditional on $\mathcal{T}$ for fitted functions. Every moment and prediction constant may depend on dimension. The results below distinguish the oracle distribution, the fitted-mean perturbation, and the alternative signal; Appendix B gives their complete proofs.

### 3.1 END-TO-END NULL CALIBRATION

Write $\xrightarrow{P}$, $\xrightarrow{d}$ for convergence in probability/distribution, and $N(0,1)$ for the standard normal law. Denote by $U_n^o = C_n^{-1} \sum_{i<j} H(U_i, U_j)$ and $\sigma_n^2 = \mathbb{E}H^2$, with $\sigma_n > 0$. The oracle conditions utilize a fourth moment and a kernel contraction

$$\Gamma_n(u,v) = \mathbb{E}[H(u,U')H(v,U')], \quad A_n = \frac{\mathbb{E}H^4}{n\sigma_n^4}, \quad B_n = \frac{\mathbb{E}\Gamma_n(U,U')^2}{\sigma_n^4}. \tag{11}$$

Here, $A_n$ controls large increments and $B_n$ controls concentration of the predictable variance. For eigenvalues $\lambda_{n,j}$ of the operator $f \mapsto \mathbb{E}[H(\cdot,U)f(U)]$ on $L^2(P_n)$, $B_n = \sum_j \lambda_{n,j}^4 / (\sum_j \lambda_{n,j}^2)^2$. Hence, the Gaussian regime concerns a diffuse effective spectrum, not dimension growth by itself.

To isolate estimation error, let us write $e_X = \widehat{m}_X - m_X$, $e_Y = \widehat{m}_Y - m_Y$, and $\widehat{t} = (r_X - e_X(Z)) \otimes (r_Y - e_Y(Z)) = t + \delta$, with tensor error $\delta$ expanded in (A.1). Its mean error is $b_n^{\mathrm{err}} = \mathbb{E}_Z(e_X \otimes e_Y)$, where $\mathbb{E}_Z$ averages over a fresh $Z$ with training fixed, so the population perturbation is second order in regression error. Assumption 3 controls this mean, the pair-kernel error, and its directional first projection separately. This is weaker than controlling that projection solely by the total feature energy.

**Theorem 1** (Feasible Gaussian calibration)**.** *Under Assumptions 1–3, the KREDIT statistic in* (9) *satisfies that*

$$\frac{\sqrt{C_n}(\widehat{U}_n - U_n^o)}{\sigma_n} \xrightarrow{P} 0, \quad \frac{\widehat{\sigma}_n^2}{\sigma_n^2} \xrightarrow{P} 1, \quad T_n \xrightarrow{d} N(0,1). \tag{12}$$

*Consequently, the Gaussian rejection probability converges to the prespecified significance level $\alpha \in (0,1)$ for every null sequence satisfying those assumptions.*

The assumptions separate analytic requirements from a particular smoothing model. Assumption 1 specifies independent evaluation and integrability; Assumption 2 supplies the relative moments and spectral condition, whose eigenvalue identity is proved in (C.56). Assumption 3 then constrains the fitted means through measurable error quantities. Compact support, kernel positivity, and Hölder smoothness enter only the sufficient regression result in Lemma 3, not the general transfer theorem. Such modular formulation permits alternative learners when their prediction errors satisfy the same conditions.

Proposition 1 strengthens such limit to a finite-sample Kolmogorov bound that controls separately the numerator perturbation and empirical studentization. In the independent isotropic Gaussian

model with $p = q = d$ and global means trained on $m \asymp n$ observations (bounded positive sample-size ratio), Corollary 2 shows that for a constant $C$ independent of $n, d$,

$$d_{\mathrm{K}}(\mathcal{L}(T_n), N(0,1)) \le C\{(n^{-1} + d^{-2})^{1/5} + (d/n)^{2/3}\}. \tag{13}$$

Here, $\mathcal{L}$ denotes law, $d_{\mathrm{K}}$ the supremum distance between distribution functions, and $d = o(n)$ means $d/n \to 0$. The rate in (13) covers fitted means and empirical joint variance for $d \to \infty$ with $d = o(n)$. For the fitted pair function $\widehat{H}$ evaluated in (8), the error $\ell_n = \|\widehat{H} - H\|_2$ can be smaller than the feature-energy bound in (A.1). Controlling that error and its directional projection separately is essential to this range.

The primitive results retain dimensional constants. Lemmas 2 and 3 translate prediction accuracy into inference conditions; Proposition 8 provides a compatible smoothing window. Proposition 9 verifies the oracle contraction, and Corollary 3 covers nonconstant scalar conditioning with $p = q = o(n^{1/3})$. Lemma 4 evaluates the feature-mean derivative exactly, and Proposition 9 bounds the conditional feature-covariance operator. These calculations yield the latter range; it is sufficient, not necessarily optimal. Proposition 11 treats fixed spectra using only a square-integrable pair kernel; Corollary 5 verifies it with finite marginal first moments. Proposition 5 gives exact discrete-stratum calibration. Proposition 4 also identifies a shifted Gaussian limit when fitted-mean bias persists at the dimension boundary, with an explicit centering for that model.

### 3.2 Signal-relative power

Under an alternative, let us define $\theta_n = \mathbb{E}H > 0$, $\tau_n = \|H\|_2$, and $\nu_n = \|\mathbb{E}[H(U, U') \mid U] - \theta_n\|_2$. The first projection fluctuates at scale $\nu_n/\sqrt{n}$, whereas the canonical remainder fluctuates at scale $\tau_n/n$. Assumption 4 retains such distinction and bounds the additional fitted first projection as specified in (A.4). It does not impose null degeneracy under an alternative.

**Theorem 2** (Power with a shrinking signal)**.** *Under Assumptions 1 and 4, we have that*

$$\widehat{U}_n/\theta_n \xrightarrow{P} 1, \quad T_n \xrightarrow{P} \infty. \tag{14}$$

*Consequently, the KREDIT test rejects with probability tending to one at every fixed upper-tail Gaussian critical value.*

Proposition 2 gives finite-sample estimation and rejection bounds in terms of these same signal and projection quantities. Unlike the null result, Theorem 2 requires neither spectral spread nor fourth moments of the pair kernel. Its sufficient conditions compare the fitted mean shift, the first-projection fluctuation, and the second-order fluctuation to the same population signal. In particular, a small first projection allows a stronger dimensional regime than a generic square-root-sample concentration bound. The proof bounds the fitted denominator from above under the alternative itself; it does not import an oracle approximation established only under conditional independence. Appendix B.2 supplies each signal-relative comparison.

Theorem 2 applies whenever estimation and sampling errors are small relative to the signal. A concrete nonlinear example makes such requirement quantitative. Let $X_j = S_j R_j$ and $Y_j = R_j^2$, with mutually independent uniform signs $S_j \in \{-1, 1\}$ and uniform radii $R_j \in \{1, 2\}$. Every cross-coordinate linear covariance is zero. Proposition 10 nevertheless shows that $\theta_d \ge c_*/d$, $\tau_d = O(1)$, and $\nu_d = O(d^{-1/2})$, for a universal $c_* > 0$ specified in (C.77). With global feature means learned from $m \asymp n$ observations, Corollary 4 yields feasible power one for $d = o(n^{3/4})$. The improvement comes from controlling the first projection directly instead of bounding all fluctuations at the same square-root-sample scale.

Proposition 3 complements detection with an empirical-variance CLT and finite-sample error bounds under a third moment of the first projection. Its assumption requires a nonzero first projection, unlike the canonical null. Table 4 compares these conditional results to the established U-statistic and high-dimensional theory (Hoeffding, 1948; Haeusler, 1988; Han & Shen, 2024). Appendix C contains their intermediate calculations.

### 3.3 A distance-signal minimax lower bound

A lower bound separates difficulty intrinsic to dependence testing from conditional-mean estimation. Let $N = m + n$ be the full sample budget, including training. Assumption 12 specifies Gaussian

rank-one alternatives with independent conditioning and sign directions; it is a subexperiment with no nuisance-learning difficulty.

**Theorem 3** (Rectangular Gaussian subexperiment lower bound)**.** *Under Assumption 12, let $P_0$ be the independent Gaussian null, and $P_{u,v}$ its alternatives indexed by unit sign vectors $u \in \{\pm p^{-1/2}\}^p$ and $v \in \{\pm q^{-1/2}\}^q$. For $0 < \gamma < 1/4$, set $b_\gamma = \gamma^2(2-\gamma)/(1-\gamma)$. Then every randomized test $0 \le \psi \le 1$ with $\mathbb{E}_{P_0^{\otimes N}}\psi \le \alpha$ satisfies that*

$$\frac{\gamma}{\pi N} \le \theta(P_{u,v}) \le \frac{\gamma}{N(1-\gamma)^{3/2}}, \quad \inf_{u,v} \mathbb{E}_{P_{u,v}^{\otimes N}}\psi \le \alpha + \frac{1}{2}\sqrt{(1-b_\gamma)^{-1/2} - 1}. \tag{15}$$

Here, $\theta(P)$ is the target under law $P$, and $P^{\otimes N}$ represents $N$ independent observations. Theorem 3 places every mixture component on the same inverse-total-sample distance-signal scale, even for unequal dimensions. Its proof preserves cancellation in the signed likelihood overlap and evaluates the distance signal by a positive integral. The Gaussian testing principle is classical (Ramdas et al., 2016; Han & Shen, 2024); Corollary 1 translates the same mixture into absolute- and squared-error risk bounds and a lower bound on honest interval length. These are subexperiment information limits.

# 4 Simulation studies

## 4.1 Designs and comparisons

Let $Z \sim \mathrm{Bernoulli}(1/2)$, $\varepsilon_X, \varepsilon_Y$ be independent standard $d$-dimensional Gaussian vectors, and $C$ equal 6 with probability $0.08$ and 1 otherwise, all mutually independent. With $p = q = d$, generate

$$X = 2Z\mathbf{1}_d + s^Z\varepsilon_X, \quad Y = -2Z\mathbf{1}_d + s^{1-Z}\{C\varepsilon_Y + \vartheta f(\varepsilon_X)\}, \tag{16}$$

where $s$ is the scale ratio and $\vartheta$ is the signal strength. The coordinatewise maps are $f(x) = x$ with $(d, s, \vartheta) = (8, 8, 0.35)$, and $f(x) = x + 1.5\sin(2x)$ with $(d, s, \vartheta) = (16, 8, 0.2)$. Matched nulls set $\vartheta = 0$.

All methods use $m = 240$, $n = 100$, and 1,000 common independent datasets per hypothesis. Suffixes P and N stand for permutation and normal calibration; P uses 199 common within-stratum permutations with recomputed studentization. Benchmark labels abbreviate distance covariance/correlation (DC), conditional distance covariance (CDC), ordinary least squares (OLS), kernel ridge regression (KRR), and cross-HSIC (xHSIC). Appendix D.2 defines these implementations.

## 4.2 Calibration, power, and finite-sample distributions

Table 2 gives KREDIT-P sizes $0.053, 0.042$ and powers $0.957, 0.784$, respectively. Its minimum advantages over the twelve benchmark comparators are $0.172, 0.211$, with positive simultaneous paired lower bounds (Table 5). Table 6 provides computation times. KREDIT-N is a same-method calibration diagnostic, not a benchmark comparator.

Figure 1 examines finite-sample normal approximation for $X = \sin(2\pi Z)\mathbf{1}_d + 2\varepsilon_X$ and $Y = \cos(2\pi Z)\mathbf{1}_d + 2\varepsilon_Y$, with $Z$ uniform on $[0, 1]$ and independent standard Gaussian errors. Appendix E.2 also studies linear conditional means; Appendices E.3 and E.4 investigate nonlinear power and empirical-variance inference.

# 5 Real data applications

## 5.1 Wine chemistry conditional on cultivar

The UCI Wine dataset contains 178 chemical profiles from three cultivars (Aeberhard & Forina, 1992). We test association between four phenolic measurements and three optical measurements, conditional on cultivar, using a 88/90 training/evaluation random split. KREDIT-P yields p-value $p = 0.0064$ with 9,999 within-cultivar permutations. This illustrates Proposition 5 under independent sampling with the observed strata. Table 10 reports every comparator; Appendix F.1 specifies the variables, training-only standardization, and split.

Table 2: Size and power at significance level $\alpha = 0.05$, using 1,000 datasets per design, $m = 240$, and $n = 100$. In (16), $(d, s, \vartheta) = (8, 8, 0.35)$ and $(16, 8, 0.2)$ define the linear and nonlinear alternatives; matched nulls set $\vartheta = 0$. P employs 199 common within-stratum permutations, and N uses normal calibration. Appendix E.1 supplies uncertainty and timing.

| | Linear | | Nonlinear | |
|---|---|---|---|---|
| Method | Size | Power | Size | Power |
| HDCI-P | 0.049 | 0.177 | 0.041 | 0.121 |
| PartialDC-P | 0.049 | 0.177 | 0.041 | 0.121 |
| LocalCDC-P | 0.041 | 0.746 | 0.043 | 0.359 |
| DISCO-local-P | 0.043 | 0.766 | 0.044 | 0.403 |
| ProjectedDC-OLS-P | 0.053 | 0.353 | 0.051 | 0.202 |
| GCM-max-P | 0.052 | 0.681 | 0.049 | 0.573 |
| WGCM-fixed-P | 0.052 | 0.575 | 0.058 | 0.473 |
| GHCM-linear-P | 0.036 | 0.418 | 0.047 | 0.198 |
| GKCM-NW-P | 0.050 | 0.785 | 0.051 | 0.480 |
| KCI-KRR-P | 0.051 | 0.734 | 0.055 | 0.414 |
| SplitKCI-KRR-P | 0.049 | 0.746 | 0.052 | 0.414 |
| xHSIC-residual-P | 0.044 | 0.730 | 0.042 | 0.489 |
| KREDIT-N | 0.068 | 0.967 | 0.062 | 0.832 |
| KREDIT-P | 0.053 | 0.957 | 0.042 | 0.784 |

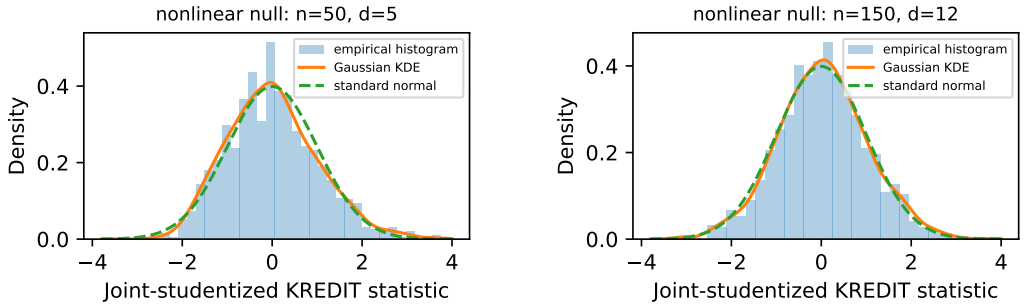


Figure 1: Null distributions of $T_n$ from 1,000 datasets per panel, with $r = 1$, $m = 4n$, and $(n, d) = (50, 5)$ (left) or $(150, 12)$ (right). Conditional means are $\sin(2\pi Z)\mathbf{1}_d$ and $\cos(2\pi Z)\mathbf{1}_d$, with uniform $Z$ on $[0, 1]$ and independent Gaussian errors of coordinate variance four. Histograms and Gaussian-kernel density estimates (KDEs) are compared to the dashed standard normal density. Fits use the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)\mathbf{1}\{|u| \le 1\}$, where $\mathbf{1}$ represents an indicator, and bandwidth $h = 1.5m^{-1/3}$. No empirical centering or rescaling is applied.

### 5.2 Six- and twelve-month gold–inflation analyses

We further analyze the 2015–2025 gold exchange-traded fund closing prices (GLD), Consumer Price Index (CPI), 10-year Treasury yields (DGS10), and unemployment rate (UNRATE) (Yahoo Finance, 2026; Federal Reserve Bank of St. Louis, 2026). Gold returns and CPI growth are consecutive-month percentage changes. HDCI tests unconditional dependence via (10); KREDIT conditions on DGS10 alone or jointly with UNRATE. Conditional fits use 24 prior calendar months, a one-month gap, and a Gaussian bandwidth equal to half the lower quartile of positive training conditioning distances.

Figures 2 and 9 depict the six- and twelve-month results. HDCI employs the signed U-centered ratio without a square root or truncation, so its normal upper-tail p-value may exceed 0.5; KREDIT uses 9,999 Gaussian multiplier draws. Dashed curves show within-family Benjamini–Hochberg (BH) adjustments for false discovery rate (FDR) control (Benjamini & Hochberg, 1995). Proposition 12 identifies the additional dependence control, and Appendix F.2 details calendar alignment and unavailable windows.

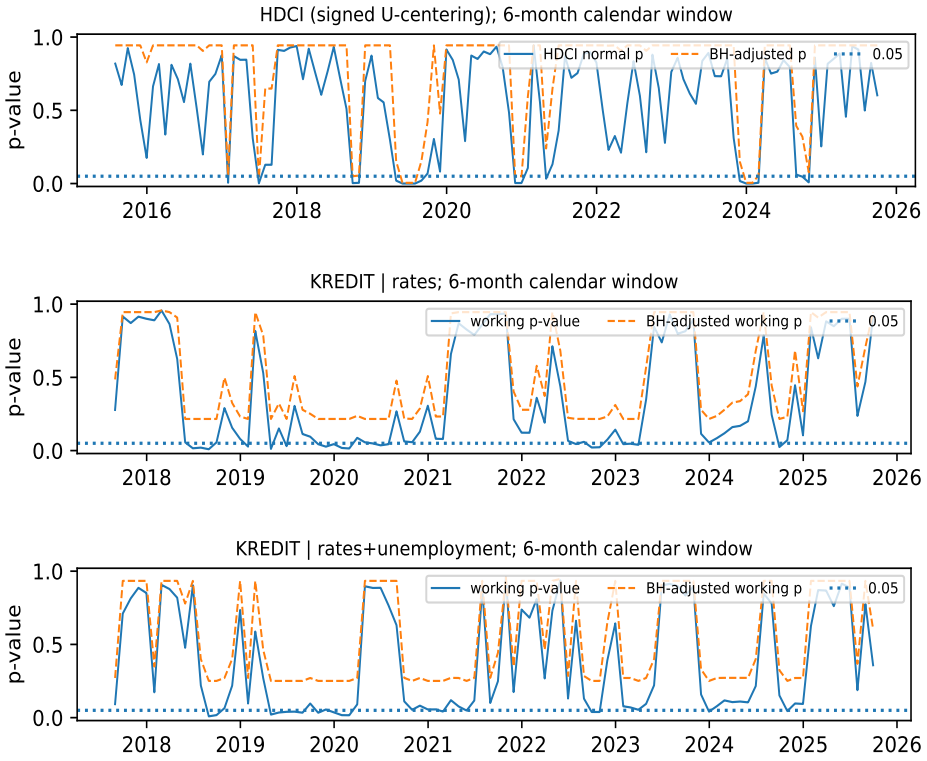


Figure 2: Six-calendar-month GLD–CPI rolling analysis. Top: direct HDCI using the signed U-centered statistic in (10) and its normal upper tail. Middle: KREDIT given the 10-year Treasury yield (DGS10). Bottom: KREDIT given DGS10 and unemployment (UNRATE). Solid curves are raw p-values; KREDIT uses 9,999 Gaussian multiplier draws. Dashed curves are within-panel BH adjustments, and dotted lines mark $0.05$. HDCI employs six paired observations without training; KREDIT uses 24 preceding training months and a one-month gap. Missing calendar months are not bridged. The HDCI panel has 123 valid windows, with 18 raw and nine adjusted p-values below $0.05$.

Table 11 shows that among 123 six-month windows, HDCI has 18 raw and nine BH-adjusted p-values below $0.05$; among 98 conditional windows, KREDIT has 21 raw crossings with rates and 17 with both covariates, but none after the BH adjustment. At twelve months, rates-only KREDIT retains 24 adjusted crossings, versus none after adding unemployment. These contrasts suggest a horizon- and conditioning-dependent gold–inflation association.

## 6 Conclusions

We have introduced KREDIT for inference on a global conditional residual-dependence target. Independent training, off-diagonal evaluation, and joint studentization connect one distance-block algorithm to quantitative null, power, and signal-estimation guarantees. Explicit Gaussian and nonlinear examples distinguish spectral, regression, and signal requirements. Simulation comparisons and two real-data applications illustrate the resulting procedures.

Future work includes structured learners for larger conditioning sets, conditioning-weighted features that reduce target cancellation, and data-adaptive training allocation. Observable prediction and dependence bounds would connect the present error quantities to practical calibration. Nuisance-dependent minimax theory and cross-statistic constructions provide further directions beyond the Gaussian subexperiment and diffuse-spectrum setting considered in the current work.

## AI USE STATEMENT

Generative artificial intelligence (AI) assistance was used in preparing this manuscript for literature search, editorial changes, bibliography preparation, and LaTeX formatting and cross-reference checks. The mathematical and empirical claims, and the final content of the submission, remain the responsibility of the authors.

## REPRODUCIBILITY STATEMENT

All assumptions are presnted in Appendix A, main proofs in Appendix B, and remaining theoretical details in Appendix C. Appendix D specifies the algorithm, benchmark comparators, calibration, and timing protocol. Appendices E and F document the numerical designs and transformations.

## REFERENCES

Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1992. URL https://archive.ics.uci.edu/dataset/109/wine. Dataset. CC BY 4.0. Numerical copy distributed with scikit-learn; archived for reproducibility.

Qianxue Bai, Yuke Shi, Qing Yang, and Qizhai Li. Conditional kernel covariance and correlation. *Advances in Mathematics (China)*, 53(6):1158–1172, 2024. doi: 10.11845/sxjz.2024017b.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Luca Bergen, Dino Sejdinovic, and Vanessa Didelez. The generalised kernel covariance measure. In *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 1198–1220. PMLR, 2026. URL https://proceedings.mlr.press/v323/bergen26a.html.

Andrew C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941. doi: 10.1090/S0002-9947-1941-0003498-3.

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005. doi: 10.1214/154957805100000104.

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

Jianqing Fan, Yang Feng, and Lucy Xia. A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*, 218(1): 119–139, 2020. doi: 10.1016/j.jeconom.2019.12.016.

Federal Reserve Bank of St. Louis. FRED: CPIAUCSL, DGS10, and UNRATE, 2026. URL https://fred.stlouisfed.org/.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/3a0772443a0739141292a5429b952fe6-Abstract.html.

Lan Gao, Yingying Fan, Jinchi Lv, and Qi-Man Shao. Asymptotic distributions of high-dimensional distance correlation inference. *The Annals of Statistics*, 49(4):1999–2020, 2021. doi: 10.1214/20-AOS2024.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70):2075–2129, 2005. URL https://jmlr.org/papers/v6/gretton05a.html.

Erich Haeusler. On the rate of convergence in the central limit theorem for martingales with discrete and continuous time. *The Annals of Probability*, 16(1):275–299, 1988. doi: 10.1214/aop/1176991901.

Qiyang Han and Yandi Shen. Generalized kernel distance covariance in high dimensions: Non-null CLTs and power universality. *Information and Inference: A Journal of the IMA*, 13 (3):iaae017, 2024. doi: 10.1093/imaiai/iaae017. URL https://academic.oup.com/imaiai/article/13/3/iaae017/7732133.

Zheng He and Danica J. Sutherland. Sequential kernel-based conditional independence testing via adaptive betting. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.18993v2. arXiv:2606.18993v2.

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*, volume 38, 2025. URL https://arxiv.org/abs/2512.14000.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Emre Kavak, Tom Nuno Wolf, and Christian Wachinger. DISCO: Mitigating bias in deep learning with conditional distance correlation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.11653v3. arXiv:2506.11653v3; first posted in 2025.

Jing Kong, Sijian Wang, and Grace Wahba. Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine*, 34(10):1708–1720, 2015. doi: 10.1002/sim.6441.

Yinfei Kong, Daoji Li, Yingying Fan, and Jinchi Lv. Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics*, 45(2):897–922, 2017. doi: 10.1214/16-AOS1474.

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. doi: 10.1111/rssb.12544.

E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020.

Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021. doi: 10.1214/20-AOS2030.

Roman Pogodin, Antonin Schrab, Yazhe Li, Danica J. Sutherland, and Arthur Gretton. Practical kernel tests of conditional independence. arXiv:2402.13196v2, 2025. URL https://arxiv.org/abs/2402.13196v2. First posted in 2024; revised September 2025.

Aaditya Ramdas, David Isenberg, Aarti Singh, and Larry Wasserman. Minimax lower bounds for linear independence testing. arXiv:1601.06259, 2016. URL https://arxiv.org/abs/1601.06259.

Oliver Schacht and Biwei Huang. A fast kernel-based conditional independence test with application to causal discovery. arXiv:2505.11085v2, 2025. URL https://arxiv.org/abs/2505.11085v2.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022. URL https://jmlr.org/papers/v23/21-1328.html.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

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. doi: 10.1214/19-AOS1857.

Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel independence test. *Journal of Machine Learning Research*, 24(369):1–68, 2023. URL https://jmlr.org/papers/v24/23-0248.html.

Tianhong Sheng and Bharath K. Sriperumbudur. On distance and kernel measures of conditional dependence. *Journal of Machine Learning Research*, 24(7):1–16, 2023. URL https://jmlr.org/papers/v24/20-238.html.

Gábor J. Székely and Maria L. Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Yin Tang and Bing Li. A reproducing-kernel-based nonparametric test for conditional independence of functional data. arXiv:2603.13704v4, 2026. URL https://arxiv.org/abs/2603.13704v4. Revised September 2026.

Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015. doi: 10.1080/01621459.2014.993081.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964. URL https://www.jstor.org/stable/25049340.

Yahoo Finance. SPDR Gold Shares (GLD): Historical Data, 2026. URL https://finance.yahoo.com/quote/GLD/history/.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813. AUAI Press, 2011. URL https://arxiv.org/abs/1202.3775.

Changbo Zhu, Xianyang Zhang, Shun Yao, and Xiaofeng Shao. Distance-based and RKHS-based dependence metrics in high dimension. *The Annals of Statistics*, 48(6):3366–3394, 2020. doi: 10.1214/19-AOS1934.