An Asymptotic Test for Conditional Independence using Analytic Kernel Embeddings

Meyer Scetbon [*1] **Laurent Meunier** [*2 3] **Yaniv Romano** [4]

Abstract

We propose a new conditional dependence measure and a statistical test for conditional independence. The measure is based on the difference between analytic kernel embeddings of two well-suited distributions evaluated at a finite set of locations. We obtain its asymptotic distribution under the null hypothesis of conditional independence and design a consistent statistical test from it. We conduct a series of experiments showing that our new test outperforms state-of-the-art methods both in terms of type-I and type-II errors even in the high dimensional setting.

1. Introduction

We consider the problem of testing whether two variables X and Y are independent given a set of confounding variables Z, which can be formulated as a hypothesis testing problem of the form:

$$H_0 : X \perp Y|Z \qquad \text{vs.} \qquad H_1 : X \not\perp Y|Z.$$

Testing for conditional independence (CI) is central in a wide variety of statistical learning problems. For example, it is at the core of graphical modeling (Lauritzen, 1996; Koller and Friedman, 2009), causal discovery (Pearl, 2009; Glymour et al., 2019), variable selection (Candès et al., 2018), dimensionality reduction (Li, 2018), and biomedical studies (Richardson and Gilks, 1993; Dobra et al., 2004; Markowetz and Spang, 2007).

Testing for H_0 in such applications is known to be a highly challenging task (Shah and Peters, 2020; Neykov et al., 2021). A large line of work has focused on the design of measures for conditional dependence based for example on kernel methods (Fukumizu et al., 2008; Sheng and Sriperumbudur, 2019; Park and Muandet, 2020; Huang et al., 2020) and rank statistics (Azadkia and Chatterjee, 2021; Shi et al., 2021b). Testing for conditional independence is even more difficult as it requires both designing a test statistic which measures the conditional dependencies and controlling its quantiles. Indeed, existing tests may fail to control the type-I error, especially when the confounding set of variables is high-dimensional with a complex dependency structure (Bergsma, 2004). Furthermore, even if the test is valid, the availability of limited data makes the problem of discriminating between the null and alternative hypotheses extremely difficult, resulting in a test of low power. These challenges has motivated the development of a series of practical methods attempting to reliably test for conditional independence. These include tests based on kernels (Zhang et al., 2012; Doran et al., 2014; Strobl et al., 2019; Zhang et al., 2017), ranks (Runge, 2018; Mittag, 2018), models (Sen et al., 2017; 2018; Chalupka et al., 2018; Shah and Peters, 2020), permutations and samplings (Berrett et al., 2020; Candès et al., 2018; Bellot and van der Schaar, 2019; Shi et al., 2021a; Javanmard and Mehrabi, 2021), and optimal transport (Warren, 2021).

Another line of work aims at building statistical tests for different problems by computing difference of analytic kernel embeddings evaluated at a finite set of locations. Two main strategies are adopted in the literature: either the locations are chosen randomly or are learned in order to maximize the power of the test. In (Epps and Singleton, 1986; Chwialkowski et al., 2015), the authors propose two-sample tests where locations are chosen randomly. In (Zhang et al., 2018), they adopt a similar method for independence testing. In (Jitkrittum et al., 2016; Scetbon and Varoquaux, 2019), the authors propose two-sample tests where the location are leaned instead. Jitkrittum et al. (2017a) learned the location for independence testing and (Jitkrittum et al., 2017b) learned them also to test for goodness-of-fit.

In this paper, we propose a new kernel-based test for conditional independence with asymptotic theoretical guarantees. Taking inspiration from (Chwialkowski et al., 2015; Jitkrittum et al., 2017a; Scetbon and Varoquaux, 2019), we use the

[*]Equal contribution [1]CREST, ENSAE, France [2]Facebook AI Research, Paris, France [3]Université Paris-Dauphine, France [4]Departments of Electrical and Computer Engineering and of Computer Science, Technion, Israel. Correspondence to: Meyer Scetbon <meyer.scetbon@ensae.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

ℓ^p distance between two well-chosen analytic kernel mean embeddings evaluated at a finite set of locations. To the best of our knowledge, it is the first time that this strategy is employed for conditional independence testing. We show that this measure encodes the conditional dependence relation of the random variables under study. Under common assumptions on the richness of the RKHS, we derive the asymptotic null distribution of our measure, and design a simple nonparametric test that is distribution-free under the null hypothesis. Furthermore, we show that our test is consistent. Lastly, we validate our theoretical claims and study the performance of the proposed approach using simulated conditionally (in)dependent data and show that our testing procedure outperforms state-of-the-art methods.

1.1. Related Work

Zhang et al. (2012) propose a kernel based-test (KCIT), by leveraging the characterization of conditional independence derived in (Daudin, 1980) to form a test statistic. The authors of this work obtain the asymptotic null distribution of the proposed statistic and derived a practical procedure from it to test for H_0. However, one main practical issue of the proposed test is that the asymptotic null distribution of their statistic cannot be computed directly as it involved unknown quantities. To address this problem, the authors propose to approximate it either with Monte Carlo simulations or by fitting a Gamma distribution. In our work, we propose a new kernel-based statistic to test for conditional independence and show that its asymptotic null distribution is simply the standard normal distribution. In addition Zhang et al. (2012) extended the Gaussian process (GP) regression framework to the multi-output case, which allowed them to find the hyperparameters involved in the test statistic, maximizing the marginal likelihood. We also deploy a similar optimization procedure to that of Zhang et al. (2012), however, in our case the output of the GP regression is univariate and therefore more computationally efficient. Note also that in (Strobl et al., 2019), the authors propose a relaxed version of KCIT which approximates it using random Fourier features and offer a new method to deal with the tradeoff between the computational cost and the power of the test.

Doran et al. (2014) propose an MMD-based test for conditional independence using a well chosen permutation matrix. The role of this permutation is to simulate samples from the factorized distribution. Once such permutation is obtained, the authors propose to apply an MMD-based two-sample test (Gretton et al., 2012) to detect conditional dependencies between the simulated distribution and the joint one. However the test proposed there can only be applied for small sample sizes as it requires to solve a linear program using the simplex algorithm to compute the permutation matrix. Note also that the authors do not have access directly to the quantiles of the asymptotic null distribution and therefore a bootstrap procedure is required to compute them. In addition, the consistency of their test holds only under some non-trivial conditions on the permutation matrix obtained. In contrast, our test can be applied for large sample sizes, admits a simple asymptotic null distributions from which the quantiles can be directly obtained and is consistent under some mild assumptions on the distributions.

Other CI tests proposed in the literature suggest testing relaxed forms of conditional independence. For instance, Shah and Peters (2020) propose the generalised covariance measure (GCM) which only characterises weak conditional dependence (Daudin, 1980) and Zhang et al. (2017) propose a kernel-based test which focuses only on individual effects of the conditioning variable Z on X and Y. Some other tests are based on the knowledge of the conditional distributions in order to measure conditional dependencies. For example Candès et al. (2018) assume that one has access to the exact conditional distributions, Bellot and van der Schaar (2019); Shi et al. (2021a) approximate them using generative models and Sen et al. (2017) consider model-based methods to generate samples from the conditional distributions. In our work, we design a test statistic which characterizes the exact conditional independence of random variables and obtain its asymptotic null distribution without assuming any knowledge on the conditional distributions. Under some mild assumptions on the RKHSs considered, we also derive an approximate test statistic which admits the same asymptotic distribution and obtain a simple testing procedure from it.

2. Background and Notations

We first recall some notions on kernels and mean embeddings which will be useful in the derivation of our conditional independence test. Let $(\mathcal{D}, \mathcal{A})$ be a Borel measurable space and denote $\mathcal{M}_1^+(\mathcal{D})$ the space of Borel probability measures on $\mathcal{D}$. Let also (H, k) be a measurable RKHS on $\mathcal{D}$, i.e. a functional Hilbert space satisfying the reproducing property: for all $f \in H$, $x \in \mathcal{D}$, $f(x) = \langle f, k_x \rangle_H$. Let $\nu \in \mathcal{M}_1^+(\mathcal{D})$. If $\mathsf{E}_{x \sim \nu}[\sqrt{k(x,x)}]$ is finite, we define for all $t \in \mathcal{D}$ the *mean embedding* as $\mu_{\nu,k}(t) := \int_{x \in \mathcal{D}} k(x,t) d\nu(x)$. Note that $\mu_{\nu,k}$ is the unique element in H satisfying for all $f \in H$, $\mathsf{E}_{x \sim \nu}(f(x)) = \langle \mu_{\nu,k}, f \rangle_H$. If $\nu \mapsto \mu_{\nu,k}$ is injective, then the kernel k is said to be *characteristic*. This property is essential for the separation property to be verified when defining a kernel metric between distributions, such as the MMD (Gretton et al., 2012), or the ℓ^p distance (Scetbon and Varoquaux, 2019).

ℓ^p-distance between mean embeddings. Let k be a definite positive, characteristic, continuous, and bounded kernel on R^d and $p \geq 1$ an integer. Scetbon and Varoquaux (2019) showed that given an absolutely continuous Borel probability measure Γ on R^d, the following function defined for any

$(P, Q) \in \mathcal{M}_1^+(\mathsf{R}^d) \times \mathcal{M}_1^+(\mathsf{R}^d)$ as

$$d_p(P, Q) := \left[\int_{\mathbb{R}^d} |\mu_{P,k}(\mathbf{t}) - \mu_{Q,k}(\mathbf{t})|^p d\Gamma(\mathbf{t}) \right]^{\frac{1}{p}} \quad (1)$$

is a metric on $\mathcal{M}_1^+(\mathsf{R}^d)$. When the kernel k is analytic[1], Scetbon and Varoquaux (2019) also showed that for any $J \geq 1$,

$$d_{p,J}(P, Q) := \left[\frac{1}{J} \sum_{j=1}^{J} |\mu_{P,k}(\mathbf{t}_j) - \mu_{Q,k}(\mathbf{t}_j)|^p \right]^{\frac{1}{p}}, \quad (2)$$

where $(\mathbf{t}_j)_{j=1}^J$ are sampled independently from the Γ distribution, is a random metric[2] on $\mathcal{M}_1^+(\mathsf{R}^d)$.

In what follows, we consider distributions on Euclidean spaces. More precisely, let $d_x, d_y, d_z \geq 1$, $\mathcal{X} := \mathsf{R}^{d_x}$, $\mathcal{Y} := \mathsf{R}^{d_y}$, and $\mathcal{Z} := \mathsf{R}^{d_z}$. Let (X, Z, Y) be a random vector on $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$ with law P_{XZY}. We denote by P_{XY}, P_X, and P_Y the law of (X, Y), X, and Y, respectively. We also denote by $\ddot{\mathcal{X}} := \mathcal{X} \times \mathcal{Z}$, $\ddot{X} := (X, Z)$, and $P_{\ddot{X}}$ its law. Let $P_X \otimes P_Y$ be the product of the two measures P_X and P_Y. Given $(H_{\ddot{X}}, k_{\ddot{X}})$ and (H_Y, k_Y), two measurable reproducing kernel Hilbert spaces (RKHS) on $\ddot{\mathcal{X}}$ and $\mathcal{Y}$, respectively, we define the tensor-product RKHS $H = H_{\ddot{X}} \otimes H_Y$ associated with its *tensor-product kernel* $k = k_{\ddot{X}} \otimes k_Y$, defined for all $\ddot{x}, \ddot{x}^0 \in \ddot{\mathcal{X}}$ and $y, y^0 \in \mathcal{Y}$, as $k((\ddot{x}, y), (\ddot{x}^0, y^0)) = k_{\ddot{X}}(\ddot{x}, \ddot{x}^0) \times k_Y(y, y^0)$.

3. A new ℓ^p kernel-based testing procedure

In this section, we present our statistical procedure to test for conditional independence. We begin by introducing a general measure based on the ℓ^p distance d_p between mean embeddings which characterizes the conditional independence. We derive an oracle test statistic for which we obtain its asymptotic distribution under both the null and alternative hypothesis. Then, we provide an efficient procedure to effectively compute an approximation of our oracle statistic and show that it has the exact same asymptotic distribution. To avoid any bootstrap or permutation procedures, we offer a normalized version of our statistic and derive a simple and consistent test from it.

3.1. Conditional Independence Criterion

Let us first introduce the criterion we use to define our statistical test. We define a probability measure $P_{\ddot{X}\ YjZ}$ on $\ddot{\mathcal{X}} \times \mathcal{Y}$ as

$$P_{\ddot{X}\ YjZ}(A \times B) := \mathsf{E}_Z \left[\mathsf{E}_{\ddot{X}}[\mathbf{1}_A|Z] \mathsf{E}_Y[\mathbf{1}_B|Z] \right],$$

for any $(A, B) \in \mathcal{B}(\ddot{\mathcal{X}}) \times \mathcal{B}(\mathcal{Y})$, where $\mathbf{1}_A$ is the characteristic function of a measurable set A and similarly for B. One can now characterize the independence of X and Y given Z as follows: $X \perp Y|Z$ if and only if $P_{XZY} = P_{\ddot{X}\ YjZ}$ (Fukumizu et al., 2004, Theorem 8). Therefore, we have a first simple characterization of the conditional independence: $X \perp Y|Z$ if and only if $d_p(P_{XZY}, P_{\ddot{X}\ YjZ}) = 0$. With this in place, we now state some assumptions on the kernel k considered in the rest of this paper.

Assumption 3.1. The kernel $k : (\ddot{\mathcal{X}} \times \mathcal{Y}) \times (\ddot{\mathcal{X}} \times \mathcal{Y}) \to \mathsf{R}$ is positive definite, characteristic, bounded, continuous and analytic. Moreover, the kernel k is a tensor product of kernels $k_{\ddot{X}}$ and k_Y on $\ddot{\mathcal{X}}$ and $\mathcal{Y}$, respectively.

It is worth noting that a sufficient condition for the kernel k to be characteristic, bounded, continuous and analytic, is that both kernels $k_{\ddot{X}}$ and k_Y are characteristic, bounded, continuous and analytic (Szabó and Sriperumbudur, 2018). For example, if the kernels $k_{\ddot{X}}$ and k_Y are Gaussian kernels[3] on $\ddot{\mathcal{X}}$ and $\mathcal{Y}$ respectively, then $k = k_{\ddot{X}} \otimes k_Y$ satisfies Assumption 3.1 (Jitkrittum et al., 2017a). Using the analyticity of the kernel k, one can work with $d_{p,J}$ defined in (2) instead of d_p to characterize the conditional independence.

Proposition 3.2. *Let $p \geq 1$, $J \geq 1$, k be a kernel satisfying Assumption 3.1, Γ an absolutely continuous Borel probability measure on $\ddot{\mathcal{X}} \times \mathcal{Y}$, and $\{(\mathbf{t}_j^{(1)}, t_j^{(2)})\}_{j=1}^J$ sampled independently from Γ. Then Γ-almost surely, $d_{p,J}(P_{XZY}, P_{\ddot{X}\ YjZ}) = 0$ if and only if $X \perp Y|Z$.*

Proof. Recall that $X \perp Y|Z$ if and only if $P_{XZY} = P_{\ddot{X}\ YjZ}$ (Fukumizu et al., 2008). If k is bounded, characteristic, and analytic, then, by invoking (Scetbon and Varoquaux, 2019, Theorem 2.1) we get that $d_{p,J}^p$ is a random metric on the space of Borel probability measures. This concludes the proof. □

The key advantage of using $d_{p,J}(P_{XZY}, P_{\ddot{X}\ YjZ})$ to measure the conditional dependence is that it only requires to compute the differences between the mean embeddings of P_{XZY} and $P_{\ddot{X}\ YjZ}$ at J locations. In what follows, we derive from it a first oracle test statistic for conditional independence.

3.2. A First Oracle Test Statistic

When the kernel k considered satisfies Assumption 3.1, we can obtain a simple expression of our measure

[1] An *analytic kernel* on R^d is a positive definite kernel such that for all x 2 R^d, $k(x, \ )$ is an analytic function, i.e., a function defined locally by a convergent power series.

[2] A random metric is a random process which satisfies all the conditions for a metric almost-surely.

[3] A gaussian kernel K on W R^d satisfies for all w, w^0 2 W, $K(w, w^0) := \exp\left(\frac{kw \ \ w^0 k_2^2}{2\sigma^2} \right)$ for some $\sigma > 0$.

$d_{p,J}(P_{XZY}, P_{\ddot{X}\ Y\!jZ})$. Indeed, the tensor formulation of the kernel k allows us to write the mean embedding of $P_{\ddot{X}\ Y\!jZ}$ for any $(\mathbf{t}^{(1)}, t^{(2)}) \in \ddot{\mathcal{X}} \times \mathcal{Y}$ as:

$$\mu_{P_{\ddot{X}\ Y\!jZ}, k_{\ddot{X}}\ k_Y}(\mathbf{t}^{(1)}, t^{(2)}) = \mathsf{E}_Z\left[\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\right]. \quad (3)$$

Then, by defining the witness function as

$$\Delta(\mathbf{t}^{(1)}, t^{(2)}) := \mathsf{E}\left[\left(k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X}) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\right) \times \left(k_Y(t^{(2)}, Y) - \mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\right)\right],$$

and by considering $\{(\mathbf{t}_j^{(1)}, t_j^{(2)})\}_{j=1}^J$ sampled independently according to Γ, we get that (see Appendix A.1 for more details)

$$d_{p,J}(P_{XZY}, P_{\ddot{X}\ Y\!jZ}) = \left[\frac{1}{J}\sum_{j=1}^J \left|\Delta(\mathbf{t}_j^{(1)}, t_j^{(2)})\right|^p\right]^{1/p}.$$

Estimation. Given n observations $\{(x_i, z_i, y_i)\}_{i=1}^n$ that are drawn independently from P_{XZY}, we aim at obtaining an estimator of $d_{p,J}^p(P_{XZY}, P_{\ddot{X}\ Y\!jZ})$. To do so, we introduce the following estimate of $\Delta(\mathbf{t}^{(1)}, t^{(2)})$, defined as

$$\Delta_n(\mathbf{t}^{(1)}, t^{(2)}) = \frac{1}{n}\sum_{i=1}^n \left(k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{x}_i) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|z_i\right]\right) \times \left(k_Y(t^{(2)}, y_i) - \mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|z_i\right]\right).$$

With this in place, a natural candidate to estimate $d_{p,J}^p(P_{XZY}, P_{\ddot{X}\ Y\!jZ})$ (up to the constant J) can be expressed as

$$\mathrm{CI}_{n,p} := \sum_{j=1}^J \left|\Delta_n(\mathbf{t}_j^{(1)}, t_j^{(2)})\right|^p,$$

where $(\mathbf{t}_1^{(1)}, t_1^{(2)}), \ldots, (\mathbf{t}_J^{(1)}, t_J^{(2)}) \in \ddot{\mathcal{X}} \times \mathcal{Y}$ are sampled independently from Γ.

We now turn to derive the asymptotic distribution of this statistic. For that purpose, define, for all $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, n\}$,

$$u_i(j) := \left(k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{x}_i) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = z_i\right]\right) \times \left(k_Y(t_j^{(2)}, y_i) - \mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = z_i\right]\right),$$

$\mathbf{u}_i := (u_i(1), \ldots, u_i(J))^T$ and $\mathbf{\Sigma} := \mathsf{E}(\mathbf{u}_1\mathbf{u}_1^T)$. We also denote by $\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{u}_i$. Observe that $\mathrm{CI}_{n,p} = \|\mathbf{S}_n\|_p^p$. In the following proposition we obtain the asymptotic distribution of our statistic $\mathrm{CI}_{n,p}$.

Proposition 3.3. *Suppose that Assumption 3.1 is verified. Let $p \geq 1$, $J \geq 1$ and $((\mathbf{t}_1^{(1)}, t_1^{(2)}), \ldots, (\mathbf{t}_J^{(1)}, t_J^{(2)})) \in (\ddot{\mathcal{X}} \times \mathcal{Y})$. Then, under H_0, we have: $\sqrt{n}\mathbf{S}_n \to \mathcal{N}(0, \mathbf{\Sigma})$. Moreover, under H_1, if $((\mathbf{t}_j^{(1)}, t_j^{(2)}))_{j=1}^J$ are sampled independently according to Γ, then Γ-almost surely, for any $q \in \mathsf{R}$, $\lim_{n\,!\,\infty} P(n^{p/2}\mathrm{CI}_{n,p} \geq q) = 1$.*

Proof. Recall that $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{u}_i$ where $\mathbf{u}_i$ are i.i.d. samples. Under H_0, $\mathsf{E}[\mathbf{u}_i] = 0$. Using the Central Limit Theorem, we get: $\sqrt{n}\mathbf{S}_n \to \mathcal{N}(0, \mathbf{\Sigma})$. Using the analyticity of the kernel k, under H_1, Γ-almost surely, there exists a $j \in \{1, \ldots, J\}$ such that $\mathsf{E}[u_1(j)] \neq 0$. Therefore, we can deduce that Γ-almost surely, $\mathbf{S} := \mathsf{E}[\mathbf{u_1}] \neq 0$. Now, for all $q > 0$, we get: $P(n^{p/2}\mathrm{CI}_{n,p} > q) \to 1$ because $\mathrm{CI}_{n,p} \to \|\mathbf{S}\|_p^p$ when $n \to \infty$. □

From the above proposition, we can define a consistent statistical test at level $0 < \alpha < 1$, by rejecting the null hypothesis if $n^{p/2}\mathrm{CI}_{n,p}$ is larger than the $(1-\alpha)$ quantile of the asymptotic null distribution, which is the law associated with $\|X\|_p^p$, where X follows the multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma})$. However, in practice, $\mathrm{CI}_{n,p}$ cannot be computed as it requires the access to samples from the conditional means involved in the statistic, namely $\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z\right]$ and $\mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z\right]$ for all $j \in \{1, \ldots, J\}$, which are unknown. Below, we show how to estimate these conditional means by using Regularized Least-Squares (RLS) estimators.

3.3. Approximation of the Test Statistic

The oracle statistic defined above involves conditional means that are unknown and cannot be used directly in practice. To alleviate this issue, we provide here a practical test statistic which approximates the oracle one while conserving its asymptotic behavior.

Our goal here is to estimate $\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = \cdot\right]$ and $\mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = \cdot\right]$ for all $j \in \{1, \ldots, J\}$ in order to effectively approximate of our statistic. To do so, we consider kernel-based regularized least squares (RLS) estimators. Let $1 \leq r \leq n$ and $\{(x_i, z_i, y_i)\}_{i=1}^r$ be a subset of r samples. Let also $j \in \{1, \ldots, J\}$, and denote by $H_Z^{1,j}$ and $H_Z^{2,j}$ two separable RKHSs on $\mathcal{Z}$. Denote also by $k_Z^{1,j}$ and $k_Z^{2,j}$ their associated kernels and $\lambda_{j,r}^{(1)},\ \lambda_{j,r}^{(2)} > 0$ the regularization parameters involved in the RLS regressions. Then, the RLS estimators are the unique solutions of the following problems:

$$\min_{h \in H_Z^{2,j}} \frac{1}{r}\sum_{i=1}^r \left(h(z_i) - k_Y(t_j^{(2)}, y_i)\right)^2 + \lambda_{j,r}^{(2)}\|h\|_{H_Z^{2,j}}^2 \text{ and}$$

$$\min_{h \in H_Z^{1,j}} \frac{1}{r}\sum_{i=1}^r \left(h(z_i) - k_{\ddot{X}}(\mathbf{t}_j^{(1)}, (x_i, z_i))\right)^2 + \lambda_{j,r}^{(1)}\|h\|_{H_Z^{1,j}}^2,$$

which we denote by $h^{(2)}_{j,r}$ and $h^{(1)}_{j,r}$, respectively. These estimators have simple expressions in terms of the kernels involved. For example, let $k_{\ddot{X}}(\mathbf{t}^{(1)}_j, \ddot{X}_r) := [k_{\ddot{X}}(\mathbf{t}^{(1)}_j, (x_1, z_1)), \ldots, k_{\ddot{X}}(\mathbf{t}^{(1)}_j, (x_r, z_r))]^T$, then for any $z \in \mathcal{Z}$, the estimator $h^{(1)}_{j,r}$ can be expressed as

$$h^{(1)}_{j,r}(z) = \sum_{i=1}^{r} [\alpha^{(1)}_{j,r}]_i k^{1,j}_Z(z_i, z)\ , \quad \text{with}$$
$$\alpha^{(1)}_{j,r} := (\mathbf{K}^{1,j}_{r,Z} + r\lambda^{(1)}_{j,r}\mathrm{Id}_r)^{\ 1} k_{\ddot{X}}(\mathbf{t}^{(1)}_j, \ddot{X}_r) \in \mathsf{R}^r,$$

where $\mathbf{K}^{1,j}_{r,Z} := (k^{1,j}_Z(z_i, z_j))_{1\ \ i,j\ \ r}$. Similarly, we obtain simple expressions of $h^{(2)}_{j,r}$. We can now introduce our new estimator of the witness function at each location $(\mathbf{t}^{(1)}_j, t^{(2)}_j)$ as follows:

$$\widetilde{\Delta}_{n,r}(\mathbf{t}^{(1)}_j, t^{(2)}_j) := \frac{1}{n}\sum_{i=1}^{n} \left(k_{\ddot{X}}(\mathbf{t}^{(1)}_j, \ddot{x}_i) - h^{(1)}_{j,r}(z_i)\right) \times \left(k_Y(t^{(2)}_j, y_i) - h^{(2)}_{j,r}(z_i)\right),$$

and the proposed test statistic becomes

$$\widetilde{\mathrm{CI}}_{n,r,p} := \sum_{j=1}^{J} \left|\widetilde{\Delta}_{n,r}(\mathbf{t}^{(1)}_j, t^{(2)}_j)\right|^p\ .$$

Asymptotic Distribution. To get the asymptotic distribution, we need to make two extra assumptions. Let us define, for $m \in \{1, 2\}$ and $j \in \{1, \ldots, J\}$, $L^{m,j}_Z$—the operator on $L^2(\mathcal{Z}, P_Z)$ as $L^{m,j}_Z(g)(\cdot) = \int_{\mathsf{Z}} k^{m,j}_Z(\cdot, z)g(z)dP_Z(z)$.

Assumption 3.4. There exists $Q > 0$, and $\gamma \in [0, 1]$ such that for all $\lambda > 0$, $m \in \{1, 2\}$ and $j \in \{1, \ldots, J\}$:

$$\mathrm{Tr}((L^{m,j}_Z + \lambda I)^{\ 1} L^{m,j}_Z) \le Q\lambda^{\ \gamma}.$$

Assumption 3.5. There exists $2 \ge \beta > 1$ such that for any $j \in \{1, \ldots, J\}$, $(\mathbf{t}^{(1)}, t^{(2)}) \in \ddot{\mathcal{X}} \times \mathcal{Y}$,

$$\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})\mathsf{j}Z = \ \right] 2\ R\left(\left[L^{1,j}_Z\right]^{\beta/2}\right),$$
$$\mathsf{E}_{Y}\left[k_Y(t^{(2)}, Y)\mathsf{j}Z = \ \right] 2\ R\left(\left[L^{2,j}_Z\right]^{\beta/2}\right),$$

where $\mathcal{R}\left(\left[L^{m,j}_Z\right]^{\beta/2}\right)$ is the image space of $\left[L^{m,j}_Z\right]^{\beta/2}$. Moreover, there exists $L, \sigma > 0$ such that for all $l \ge 2$ and P_Z-almost all $z \in \mathcal{Z}$

$$\mathsf{E}_{\ddot{X}\mathsf{j}Z=z}\left[\left|k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})\quad \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})\ \mathsf{j}\ Z\right]\right|^l\right] \quad \frac{l!\sigma^2 L^{l\ \ 2}}{2},$$
$$\mathsf{E}_{Y\mathsf{j}Z=z}\left[\left|k_Y(t^{(2)}, Y)\quad \mathsf{E}_{Y}\left[k_Y(t^{(2)}, Y)\mathsf{j}Z\right]\right|^l\right] \quad \frac{l!\sigma^2 L^{l\ \ 2}}{2}.$$

These assumptions are central in our proofs and are common in kernel statistic studies (Caponnetto and De Vito, 2007; Fischer and Steinwart, 2020; Rudi and Rosasco, 2017). Under these assumptions, (Fischer and Steinwart, 2020) proved optimal learning rates for RLS in RKHS norm, which is essential to guarantee that our new statistic $\widetilde{\mathrm{CI}}_{n,r,p}$, estimated with RLS, has the same asymptotic law as our oracle estimator $\mathrm{CI}_{n,p}$.

To derive the asymptotic distribution of our new test statistic, we also need to define for all $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, n\}$, $\widetilde{u}_{i,r}(j) := (k_{\ddot{X}}(\mathbf{t}^{(1)}_j, \ddot{x}_i) - h^{(1)}_{j,r}(z_i))(k_Y(t^{(2)}_j, y_i) - h^{(2)}_{j,r}(z_i))$, $\widetilde{\mathbf{u}}_{i,r} := (\widetilde{u}_{i,r}(1), \ldots, \widetilde{u}_{i,r}(J))^T$, and $\widetilde{\mathbf{S}}_{n,r} := \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathbf{u}}_{i,r}$. Note that $\widetilde{\mathrm{CI}}_{n,r,p} = \|\widetilde{\mathbf{S}}_{n,r}\|^p_p$. In the following proposition, we show the asymptotic behavior of the statistic of interest. The proof of this proposition is given in Appendix A.2.

Proposition 3.6. *Suppose that Assumptions 3.1-3.4-3.5 are verified. Let $p \ge 1$, $J \ge 1$, $((\mathbf{t}^{(1)}_1, t^{(2)}_1), \ldots, (\mathbf{t}^{(1)}_J, t^{(2)}_J)) \in (\ddot{\mathcal{X}} \times \mathcal{Y})^J$, r_n such that $n^{\frac{\beta+\gamma}{2\beta}} \in o(r_n)$ and $\lambda_{r_n} = r_n^{\frac{1}{1+\gamma}}$. Then, under H_0, we have $\sqrt{n}\widetilde{\mathbf{S}}_{n,r_n} \to \mathcal{N}(0, \mathbf{\Sigma})$. Moreover, under H_1, if the $((\mathbf{t}^{(1)}_j, t^{(2)}_j))_{j=1}^J$ are sampled independently according to Γ, then Γ-almost surely, for any $q \in \mathsf{R}$, $\lim_{n\ !\ \infty} P(n^{p/2}\widetilde{\mathrm{CI}}_{n,r_n,p} \ge q) = 1$.*

From the above proposition, we can derive a consistent test at level α for $0 < \alpha < 1$. Indeed, we obtain the asymptotic null distribution of $n^{p/2}\widetilde{\mathrm{CI}}_{n,r_n,p}$ and we show that under the alternative hypothesis H_1, Γ-almost surely, $n^{p/2}\widetilde{\mathrm{CI}}_{n,r_n,p}$ is arbitrarily large as n goes to infinity. For a fixed level α, the test rejects H_0 if $n^{p/2}\widetilde{\mathrm{CI}}_{n,r_n,p}$ exceeds the $(1-\alpha)$-quantile of its asymptotic null distribution and this test is therefore consistent. For example, when $p \in \{1, 2\}$, the asymptotic null distribution of $n^{p/2}\widetilde{\mathrm{CI}}_{n,r_n,p}$ is either a sum of correlated Nakagami variables[4] ($p = 1$) or a sum of correlated chi square variables ($p = 2$). However, computing the quantiles of these asymptotic null distributions can be computationally expensive as it requires a bootstrap or permutation procedure. In the following, we consider a different approach in which we normalize the statistic to obtain a simple asymptotic null distribution.

3.4. Normalization of the Test Statistic

Herein, we consider a normalized variant of our statistic $\widetilde{\mathrm{CI}}_{n,r,p}$ in order to obtain a tractable asymptotic null distribution. Denote $\mathbf{\Sigma}_{n,r} := \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{u}}_{i,r}\widetilde{\boldsymbol{u}}^T_{i,r}$ and let $\delta_n > 0$,

[4] the probability density function of a Nakagami distribution of parameters $m \quad \frac{1}{2}$ and $\omega > 0$ is for all $x \quad 0$, $f(x, m, \omega) = \frac{2m^m}{G(m)\omega^m}x^{2m\ \ 1}\exp(-\frac{m}{\omega}x^2)$ where G is the Euler Gamma function.

then the normalized statistic considered is given by

$$\widetilde{\mathrm{NCI}}_{n,r,p} := \|(\mathbf{\Sigma}_{n,r} + \delta_n \mathrm{Id}_J)^{-1/2} \widetilde{\mathbf{S}}_{n,r}\|_p^p.$$

In the next proposition, we show that our normalized approximate statistic converges in law to the standard multivariate normal distribution. The proof is given in Appendix A.3.

Proposition 3.7. *Suppose that Assumptions 3.1-3.4-3.5 are verified. Let $p \geq 1$, $J \geq 1$, $((\mathbf{t}_1^{(1)}, t_1^{(2)}), \ldots, (\mathbf{t}_J^{(1)}, t_J^{(2)})) \in (\ddot{\mathcal{X}} \times \mathcal{Y})^J$, r_n such that $n^{\frac{\beta+\gamma}{2\beta}} \in o(r_n)$, $\lambda_n = r_n^{-\frac{1}{1+\gamma}}$ and $(\delta_n)_{n>0}$ a sequence of positive real numbers such that $\lim_{n\to\infty} \delta_n = 0$. Then, under H_0, we have $\sqrt{n}(\mathbf{\Sigma}_{n,r} + \delta_n Id_J)^{-1/2}\mathbf{S}_{n,r_n} \to \mathcal{N}(0, Id_J)$. Moreover, under H_1, if the $((\mathbf{t}_j^{(1)}, t_j^{(2)}))_{j=1}^J$ are sampled independently according to Γ, then Γ-almost surely, for any $q \in \mathbb{R}$, $\lim_{n\to\infty} P(n^{p/2}\widetilde{\mathrm{NCI}}_{n,r_n,p} \geq q) = 1$.*

Remark 3.8. We emphasize that J need not increase with n for test consistency. Note also that the regularization parameter δ_n allows to ensure that $(\mathbf{\Sigma}_{n,r} + \delta_n \mathrm{Id}_J)^{-1/2}$ can be stably computed. In practice, δ_n requires no tuning, and can be set to be a very small constant.

Our normalization procedure allows us to derive a simple statistical test, which is distribution-free under the null hypothesis.

Statistical test at level α: Compute $n^{p/2}\widetilde{\mathrm{NCI}}_{n,r,p}$, choose the threshold τ corresponding to the $(1-\alpha)$ quantile of the asymptotic null distribution, and reject the null hypothesis whenever $n^{p/2}\widetilde{\mathrm{NCI}}_{n,r,p}$ is larger than τ. For example, if $p = 2$, the threshold τ is the $(1-\alpha)$-quantile of $\chi^2(J)$, i.e., a sum of J *independent* standard χ^2 variables.

Total Complexity: Our normalized statistic $\widetilde{\mathrm{NCI}}_{n,r,p}$ requires first to compute $\alpha_{j,r}^{(1)}$ and $\alpha_{j,r}^{(2)}$. These quantities can be evaluated in at most $\mathcal{O}(r^2 d + r^3)$ algebraic operations where d corresponds to the computational cost of evaluating the kernels involved in the RLS regressions. We will use the above for the complexity analysis of our method, although one can consider the theoretical estimation given by the Coppersmith–Winograd algorithm (Coppersmith and Winograd, 1987) that reduces the computational cost to $\mathcal{O}(r^2 d + r^{2.376})$. Once $\alpha_{j,r}^{(1)}$ and $\alpha_{j,r}^{(2)}$ are available, evaluating the RLS estimators $h_{j,r}^{(1)}$ and $h_{j,r}^{(2)}$ requires only $\mathcal{O}(rd)$ operations. Then $\widetilde{\Delta}_{n,r}$ can be evaluated in $\mathcal{O}(nrd + r^2 d + r^3)$ operations and $\widetilde{\mathrm{CI}}_{n,r,p}$ has therefore a computational complexity of $\mathcal{O}(J(nrd + r^2 d + r^3))$. The computation of $\widetilde{\mathrm{NCI}}_{n,r,p}$ requires inverting a $J \times J$ matrix $\mathbf{\Sigma}_{n,r} + \delta_n \mathrm{Id}_J$, but this is fast and numerically stable: we empirically observe that only a small value of J is required (see Section 4), e.g. less than 10. Finally the total computational cost to evaluate $\widetilde{\mathrm{NCI}}_{n,r,p}$ is $\mathcal{O}(J(nrd + r^2 d + r^3) + nJ^2 + J^3)$.

3.5. Hyperparameters

The hyperparameters of our statistics $\widetilde{\mathrm{NCI}}_{n,r,p}$ fall into two categories: those directly involved with the test and those of the regression. We assume from now on that all the kernels involved in the computation of our statistics are *Gaussian kernels*, and consider n i.i.d. observations $\{(x_i, z_i, y_i)\}_{i=1}^n$.

The first category includes both the choice of the locations $((t_x, t_z)_j, (t_y)_j))_{j=1}^J$ on which differences between the mean embeddings are computed and the choice of the kernels $k_{\ddot{X}}$ and k_Y. Each location t_x, t_y, t_z is randomly chosen according to a Gaussian variable with mean and covariance of $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$, and $\{z_i\}_{i=1}^n$, respectively. As we consider Gaussian kernels, we should also choose the bandwidths. Here, we restrict ourselves to one-dimensional kernel bandwidths σ_X, σ_Y, and σ_Z for the kernels k_X, k_Y, and k_Z, respectively. More precisely, we select the median of $\{\|x_i - x_j\|_2\}_{1 \leq i<j \leq n}$, $\{\|y_i - y_j\|_2\}_{1 \leq i<j \leq n}$, and $\{\|z_i - z_j\|_2\}_{1 \leq i<j \leq n}$ for σ_X, σ_Y, and σ_Z, respectively.

The other category contains all the kernels $k^{m,j}$ and the regularization parameters $\lambda_{j,r}^{(m)}$ involved in the RLS problems. These parameters should be selected carefully to avoid either underfitting of the regressions, which may increase the type-I error, or overfitting, which may result in a large type-II error. To optimize these, similarly to (Zhang et al., 2012), we consider a GP regression that maximizes the likelihood of the observations. While carrying out a precise GP regression can be prohibitive, in practice, we run this method only on a batch of size 200 observations randomly selected and we perform only 10 iterations for choosing the hyperparameters involved in the RLS problems. Hence, our optimization procedure does not affect the total computational cost as it is independent of the number of observations n.

Remark 3.9. Note that here we select the locations $((t_x, t_z)_j, (t_y)_j))_{j=1}^J$ randomly. If one wants to choose the locations by maximizing the power the test, then a bi-level optimization problem appears as the RLS estimators depend on the locations chosen and we believe that it is out of the scope of this paper.

4. Experiments

The goal of this section is three fold: (i) to investigate the effects of the parameters J and p on the performances of our method, (ii) to validate our theoretical results depicted in Propositions 3.3 and 3.7, and (iii) to compare our method with those proposed in the literature. In more detail, we first compare the performance of our method, both in terms of both power and type-I error, by varying the hyperparameters J and p. We show that our method is robust to the choice of p, and also show that the power increases as J increases. Then, we explore synthetic toy problems where one can derive an explicit formulation of

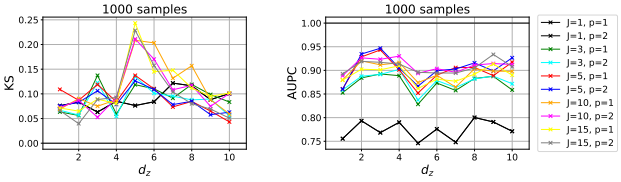

Figure 1. Comparison of the KS statistic (*left*) and the AUPC (*right*) of our test statistic $\widetilde{\text{NCI}}_{n,r,p}$ when the data is generated respectively from the models defined in (4) and (5) with Gaussian noises for multiple p and J. For each problem, we draw $n = 1000$ samples and repeat the experiment 100 times. We set $r = 1000$ and report the results obtained when varying the dimension d_z of each problem from 1 to 10. Observe that when $J = 1$, for all p $\widetilde{\text{NCI}}_{n,r,1} = \widetilde{\text{NCI}}_{n,r,p}$, therefore there is only one common black curve.

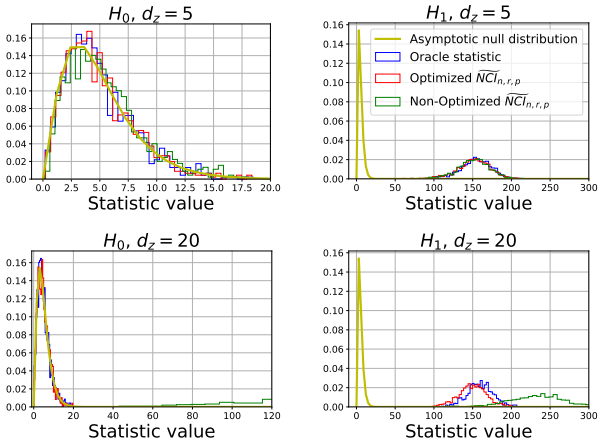

Figure 2. Comparisons between the empirical distributions of the normalized version of the oracle statistic $\widehat{\text{CI}}_{n,p}$ and the approximate normalized statistic $\widetilde{\text{NCI}}_{n,r,p}$, with the theoretical asymptotic null distribution when the data is generated either from the model defined in (6) (*left*) or the one defined in (7) (*right*). We set the dimension of Z to be either $d_z = 5$ (*top row*) or $d_z = 20$ (*bottom row*). For each problem, we draw $n = 1000$ samples and repeat the experiment 1000 times. In all the experiments, we set $J = 5$ and $p = 2$, thus the asymptotic null distribution follows a $\chi^2(5)$. Observe that both the oracle statistic and the approximated one recover the true asymptotic distribution under the null hypothesis. When H_1 holds, we can see that the two statistics manage to reject the null hypothesis. This figure also illustrates the empirical distribution of our approximate statistic when we do not optimize the hyperparameters involved in the RLS estimators: in this case we do not control the type-I error in the high dimensional setting.

the conditional means involved in our test statistic. In these cases, we can compute our proposed oracle statistic $\widehat{\text{CI}}_{n,p}$ and its normalized version, allowing us to show that under the null hypothesis we recover the theoretical asymptotic null distribution obtained in Proposition 3.3. We also reach similar conclusions regarding our approximate normalized test statistic, $\widetilde{\text{NCI}}_{n,r,p}$. In addition, in this experiment, we investigate the effect of the proposed optimization procedure for choosing the hyperparameters involved in the RLS estimators of $\widetilde{\text{NCI}}_{n,r,p}$, and show its benefits. Finally, we demonstrate on several synthetic experiments that our proposed testing procedure outperforms state-of-the-art (SoTA) methods both in terms of statistical power and type-I error, even in the high dimensional setting. The code is available at https://github.com/meyerscetbon/lp-ci-test[5].

Benchmarks. We consider 6 synthetic data sets and compare the power and type-I error of our test $\widetilde{\text{NCI}}_{n,r,p}$ to the following 6 existing CI methods: **KCIT** (Zhang et al., 2012), **RCIT** (Strobl et al., 2019), **CCIT** (Sen et al., 2017), **CRT** (Candès et al., 2018) using correlation statistic from (Bellot and van der Schaar, 2019), **FCIT** (Chalupka et al., 2018) and **GCM** (Shah and Peters, 2020). Software packages of all the above tests are freely available online and each experiment was run on a single CPU.

Evaluation. To evaluate the performance of the tests, we consider four metrics. Under H_0, we report either the Kolmogorov-Smirnov (KS) test statistic between the distribution of p-values returned by the tests and the uniform distribution on $[0, 1]$, or the type-I errors at level $\alpha = 0.05$. Note that a valid conditional independence test should control the type-I error rate at any level α. Here, a test that generates a p-value that follows the uniform distribution over $[0, 1]$ will achieve this requirement. The latter property of the p-values translates to a small KS statistic value. Under H_1, we compute either the area under the power curve (AUPC) of the empirical cumulative density function of the p-values returned by the tests, or the resulting type-II error. A conditional test has higher power when its AUPC is closer to one. Alternatively, the smaller the type-II error is, the more powerful the test is.

Effects of p and J. Our first experiment studies the effects of p and J on our proposed method. To do so, we follow the synthetic experiment proposed in (Strobl et al., 2019). To evaluate the type-I error, we generate data that follows the model:

$$X = f_1(\varepsilon_x),\ Y = f_2(\varepsilon_y),\ \text{and Z} \sim \mathcal{N}(0_d, I_{d_z}), \quad (4)$$

where Z, ε_x, and ε_y are samples from jointly independent standard Gaussian or Laplace distributions, and f_1 and f_2 are smooth functions chosen uniformly from the set $\{(\cdot), (\cdot)^2, (\cdot)^3, \tanh(\cdot), \exp(-|\cdot|)\}$. To compare the power of the tests, we also consider the model:

$$X = f_1(\varepsilon_x + 0.8\varepsilon_b), Y = f_2(\varepsilon_y + 0.8\varepsilon_b), \quad (5)$$

where ε_b is sampled from a standard Gaussian or Laplace distribution. In Figure 1, we compare the KS statistic and

[5]Our code requires a slight modification of the Gaussian Process Regression implemented in scikit-learn (Pedregosa et al., 2011) to limit the number of iterations involved in the optimization procedure.

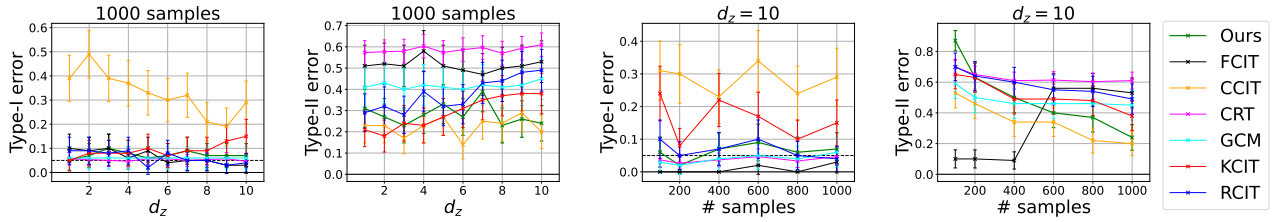

Figure 3. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line) and the type-II error (lower is better) of our test procedure with other SoTA tests on the two problems presented in (4) and (5) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): type-I and type-II errors obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals 1000. (*Middle-right, right*): type-I and type-II errors obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals 10.

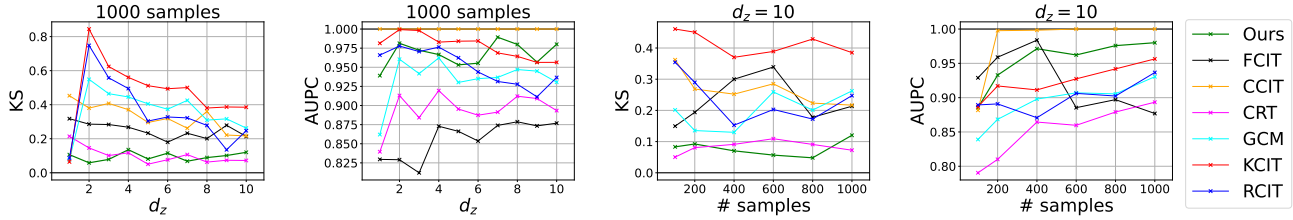

Figure 4. Comparison of the KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (8) and Eq. (9) with Laplace noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): the KS statistic and AUPC (respectively) obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals 1000. (*Middle-right, right*): the KS and AUPC (respectively), obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals 10.

the AUPC of our method when varying p and J. That figure shows that (i) our method is robust to the choice of p, and (ii) the performances of the test do not necessarily increase as J increases. Armed with theses observations, in the following experiments, we always set $p = 2$ and $J = 5$ for our method.

Effect of the rank r. In this experiment, we investigate the effect of the rank regression r on our proposed method both in terms of performance and time. For that purpose, in Figure 5, we consider the two problems presented in (4) and (5) with Gaussian noises and show the type-I and type-II when varying the ratio r/n for multiple sample size n. We observe that the rank r does not affect the power of the method, however we observe that the type-I error decreases as the ratio increases. Therefore the rank r allows in practice to deal with the tradeoff between the computational time and the control of the type-I error. In the following experiment we always set $r = n$ for simplicity.

Illustrations of our theoretical findings. The following experiment confirms that validity of our theoretical results from Propositions 3.3 and 3.7. For that purpose, we generate two synthetic data sets for which either H_0 or H_1 holds. Concretely, we define a first triplet (X, Y, Z) as follows:

$$X = \mathrm{P}_1(Z) + \varepsilon_x, \quad Y = \mathrm{P}_1(Z) + \varepsilon_y. \tag{6}$$

Above, ε_x and ε_y follow two independent standard normal distributions, $Z \sim \mathcal{N}(0_{d_z}, \Sigma)$ with $\Sigma \in \mathsf{R}^{d_z \ d_z}$. The covariance matrix Σ is obtained by multiplying a random matrix whose entries are independent and follow standard normal distribution, by its transpose, and P_1 is a projection onto the first coordinate. As a result, in this case, we have that $X \perp Y \mid Z$. We also consider a modification of the above data generating function for which H_1 holds. This is done by adding a noise component ε_b that is shared across X and Y as follows:

$$X = \mathrm{P}_1(Z) + \varepsilon_x + \varepsilon_b, \quad Y = \mathrm{P}_1(Z) + \varepsilon_y + \varepsilon_b, \tag{7}$$

where ε_b follows the standard normal distribution. Since we consider *Gaussian kernels*, we can obtain an explicit formulation of $\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = \cdot\right]$ and $\mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = \cdot\right]$ for both data generation functions. See Appendix B for more details. Consequently, we are able to compute both the normalized version of our oracle statistic $\widehat{\mathrm{CI}}_{n,p}$ and our approximate normalized statistic $\widetilde{\mathrm{NCI}}_{n,r,p}$. In Figure 2, we show that both statistics manage to recover the asymptotic distribution under H_0, and reject the null hypothesis under H_1. In addition, we show that in the high dimensional setting, only our optimized version of $\widetilde{\mathrm{NCI}}_{n,r,p}$—obtained by optimizing the hyperparameters involved in the RLS estimators of our statistic—manages to recover the asymptotic distribution under H_0.

Comparisons with existing tests. In our next experiments,

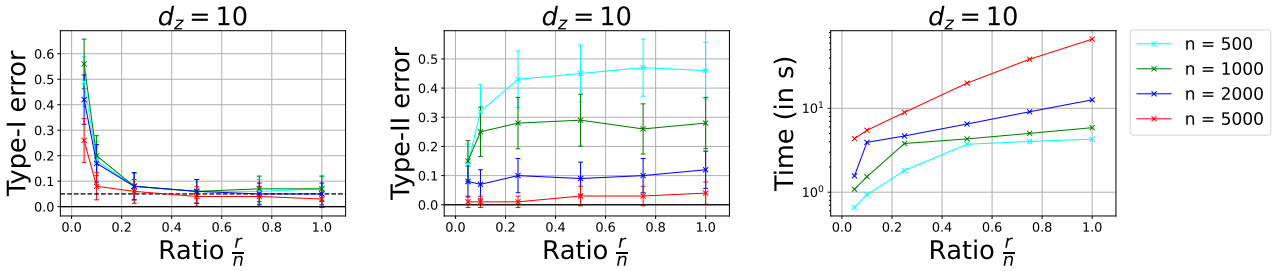

Figure 5. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line) and the type-II error (lower is better) of our test procedure on the two problems presented in (4) and (5) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, Middle*): type-I and type-II errors obtained by each test when varying the ratio regression rank/total numbers of samples for different numbers of samples. (*Right*): time in seconds (log-scale) to compute the statistic when varying the ratio regression rank/total number of samples for different number of samples.

we compare the performance of our method (implemented with the optimized version of our statistic) with state-of-the-art techniques for conditional independence testing. We first study the two data generating functions from (4) and (5). For each of these problems, we consider two settings. In the first, we fix the dimension d_z while varying the number of samples n. In the second, we fix the number of samples while varying the dimension of the problem. To evaluate the performance of the tests, we compare the type-I errors at level $\alpha = 0.05$ under the first model (4), and, for the second model (5), we evaluate the power of the test by presenting the type-II error. Figures 3 (Gaussian case) and 10 (Laplace case) demonstrate that our method consistently controls the type-I error and obtains a power similar to the best SoTA tests. In Figures 8 and 11, we compare the KS statistic and the AUPC of the different tests, and obtain similar conclusions. In addition, in Figure 6, we consider the same setting as in Figure 8 where we samples noises randomly according to a non-symmetric mixture of Gaussians and obtain the same results. In Figure 9 and 12, we investigate the high dimensional regime and show that our test is the only one which manages to control the type-I error while being competitive in term of power with other methods. See Appendix B.1 for more details.

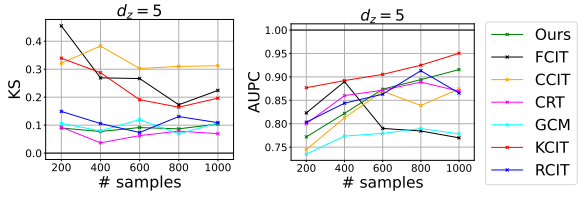

Figure 6. In this experiment we compare the KS statistic and the AUPC of our test procedure with other SoTA tests on the two problems presented in (4) and (5) where noises are randomly sampled according to a non-symmetric mixture of Gaussians. Each point in the figures is obtained by repeating the experiment for 100 independent trials. Here we fix the dimension to be $d = 5$ and we vary the number of samples n.

We now conduct another series of experiments that build upon the synthetic data sets presented in (Zhang et al., 2012; Li and Fan, 2020; Doran et al., 2014; Bellot and van der Schaar, 2019). To compare type-I error rates, we generate simulated data for which H_0 is true:

$$X = f_1\left(\bar{Z} + \varepsilon_x\right), Y = f_2\left(\bar{Z} + \varepsilon_y\right). \tag{8}$$

Above, $\bar{Z}$ is the average of $Z = (Z_1, \cdots, Z_{d_z})$, ε_x and ε_y are sampled independently from a standard Gaussian or Laplace distribution, and f_1 and f_2 are smooth functions chosen uniformly from the set $\{(\cdot), (\cdot)^2, (\cdot)^3, \tanh(\cdot), \exp(-|\cdot|)\}$. To evaluate the power, we consider the following data generating function:

$$X = f_1\left(\bar{Z} + \varepsilon_x\right) + \varepsilon_b, Y = f_2\left(\bar{Z} + \varepsilon_y\right) + \varepsilon_b, \tag{9}$$

where ε_b is a standard Gaussian or Laplace distribution.

As in the previous experiment, for each model, we study two settings by either fixing the dimension d_z, or the sample size n. In Figure 4 (Laplace case) and 14 (Gaussian case), we compare the KS and the AUPC of our method with the SoTA tests and demonstrate that our procedure manages to be powerful while controlling the type-I error. In Figures 13 and 16, we also compare the type-I and type-II errors of the different tests, and obtain similar conclusions. In addition, we investigate the high dimensional regime and show in Figure 15 and 18 that our test outperforms all the other proposed methods in most of the settings. See Appendix B.2 for more details.

Acknowledgements

M.S. was supported by a "Chaire d'excellence de l'IDEX Paris Saclay". Y.R. was supported by the Israel Science Foundation (grant 729/21) and thanks the Career Advancement Fellowship, Technion, for providing research support.

References

Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6): 3070–3102, 2021.

Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems 32*, pages 2199–2208, 2019.

Wicher Pieter Bergsma. *Testing conditional independence for continuous random variables*. Citeseer, 2004.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80 (3):551–577, 2018.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28:1981–1989, 2015.

Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 1–6, 1987.

JJ Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980.

Adrian Dobra, Chris Hans, Beatrix Jones, Joseph R Nevins, Guang Yao, and Mike West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004.

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141. Citeseer, 2014.

TW Epps and Kenneth J Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4):177–203, 1986.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *J. Mach. Learn. Res.*, 21:205–1, 2020.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Zhen Huang, Nabarun Deb, and Bodhisattva Sen. Kernel partial correlation coefficient–a measure of conditional dependence. *arXiv preprint arXiv:2012.14804*, 2020.

Adel Javanmard and Mohammad Mehrabi. Pearson chi-squared conditional randomization test. *arXiv preprint arXiv:2111.00027*, 2021.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29, 2016.

Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. JMLR, 2017a.

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. *Advances in Neural Information Processing Systems*, 30, 2017b.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. CRC Press, 2018.

Chun Li and Xiaodan Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3): e1489, 2020.

Florian Markowetz and Rainer Spang. Inferring cellular networks–a review. *BMC bioinformatics*, 8(6):1–17, 2007.

Nikolas Mittag. A nonparametric k-sample test of conditional independence. 2018.

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

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, 2020.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Sylvia Richardson and Walter R Gilks. A bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138(6):430–442, 1993.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *NIPS*, pages 3215–3225, 2017.

Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

Meyer Scetbon and Gael Varoquaux. Comparing distributions: ℓ_1 geometry improves kernel two-sample testing. In *Advances in Neural Information Processing Systems*, volume 32, pages 12327–12337, 2019.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test, 2017.

Rajat Sen, Karthikeyan Shanmugam, Himanshu Asnani, Arman Rahimzamani, and Sreeram Kannan. Mimic and classify: A meta-algorithm for conditional independence testing. *arXiv preprint arXiv:1806.09708*, 2018.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48, Jun 2020.

Tianhong Sheng and Bharath K. Sriperumbudur. On distance and kernel measures of conditional independence. *arXiv: Statistics Theory*, 2019.

Chengchun Shi, Tianlin Xu, Wicher Bergsma, and Lexin Li. Double generative adversarial networks for conditional independence testing. *Journal of Machine Learning Research*, 22(285):1–32, 2021a.

Hongjian Shi, Mathias Drton, and Fang Han. On azadkia-chatterjee's conditional dependence coefficient. *arXiv preprint arXiv:2108.06827*, 2021b.

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), 2019.

Zoltán Szabó and Bharath Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:233, 2018.

Andrew Warren. Wasserstein conditional independence testing. *arXiv preprint arXiv:2107.14184*, 2021.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Qinyi Zhang, Sarah Filippi, Seth Flaxman, and D. Sejdinovic. Feature-to-feature regression for a two-step conditional independence test. In *UAI*, 2017.

Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.

Supplementary Material

A. Proofs

A.1. On the Formulation of the Witness Function

Let $(\mathbf{t}_j)_{j=1}^J$ sampled independently from the Γ distribution, then by definition of $d_{p,J}(\cdot,\cdot)$, we have that

$$d_{p,J}(P_{XZY}, P_{\ddot{X}\ \ YjZ}) := \left[\frac{1}{J}\sum_{j=1}^{J}\left|\mu_{P_{XZY},k_{\ddot{X}}\ k_Y}(\mathbf{t}_j) - \mu_{P_{\ddot{X}}\ \ YjZ,k_{\ddot{X}}\ k_Y}(\mathbf{t}_j)\right|^p\right]^{\frac{1}{p}},$$

Moreover thanks to Assumption 3.1, we have that for any $(\mathbf{t}^{(1)}, t^{(2)}) \in \ddot{\mathcal{X}} \times \mathcal{Y}$

$$\begin{aligned}\mu_{P_{\ddot{X}}\ \ YjZ,k_{\ddot{X}}\ k_Y}(\mathbf{t}^{(1)}, t^{(2)}) &= \mathsf{E}_Z\left[\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\right]\ , \text{ and}\\ \mu_{P_{XZY},k_{\ddot{X}}\ k_Y}(\mathbf{t}^{(1)}, t^{(2)}) &= \mathsf{E}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})k_Y(t^{(2)}, Y)\right]\ .\end{aligned}$$

Let us now introduce the following witness function

$$\Delta(\mathbf{t}^{(1)}, t^{(2)}) := \mathsf{E}\left[\left(k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X}) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\right) \times \left(k_Y(t^{(2)}, Y) - \mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\right)\right]\ .$$

Therefore we obtain that

$$\begin{aligned}\Delta(\mathbf{t}^{(1)}, t^{(2)}) = &\mathsf{E}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})(k_Y(t^{(2)}, Y)\right]\\ &- \mathsf{E}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\right]\\ &+ \mathsf{E}\left[\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\right]\\ &- \mathsf{E}\left[\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]k_Y(t^{(2)}, Y)\right]\ .\end{aligned}$$

Now remark that

$$\begin{aligned}\mathsf{E}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\right] &= \mathsf{E}\left[\mathsf{E}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]|Z\right]\right]\\ &= \mathsf{E}\left[\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\right]\ .\end{aligned}$$

Similarly, we have that

$$\mathsf{E}\left[\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]k_Y(t^{(2)}, Y)\right] = \mathsf{E}\left[\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\right]$$

from which follows that

$$\begin{aligned}\Delta(\mathbf{t}^{(1)}, t^{(2)}) &= \mathsf{E}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})(k_Y(t^{(2)}, Y)\right] - \mathsf{E}\left[\mathsf{E}_Y\left[k_Y(t^{(2)}, Y)|Z\right]\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}^{(1)}, \ddot{X})|Z\right]\right]\\ &= \mu_{P_{XZY},k_{\ddot{X}}\ k_Y}(\mathbf{t}^{(1)}, t^{(2)}) - \mu_{P_{\ddot{X}}\ \ YjZ,k_{\ddot{X}}\ k_Y}(\mathbf{t}^{(1)}, t^{(2)})\ .\end{aligned}$$

A.2. Proof of Proposition 3.6

Proof. For all $j \in [J]$:

$$\sqrt{n}\widetilde{\Delta}_{n,r}(\mathbf{t}_j^{(1)}, t_j^{(2)}) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{x}_i) - h_{j,r}^{(1)}(z_i)\right)\left(k_Y(t_j^{(2)}, y_i) - h_{j,r}^{(2)}(z_i)\right)$$

$$= \sqrt{n}\Delta_n(\mathbf{t}_j^{(1)}, t_j^{(2)}) \tag{10}$$

$$+ \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{x}_i) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = z_i\right]\right)\left(\mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = z_i\right] - h_{j,r}^{(2)}(z_i)\right) \tag{11}$$

$$+ \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = z_i\right] - h_{j,r}^{(1)}(z_i)\right)\left(k_Y(t_j^{(2)}, y_i) - \mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = z_i\right]\right) \tag{12}$$

$$+ \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = z_i\right] - h_{j,r}^{(1)}(z_i)\right)\left(\mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = z_i\right] - h_{j,r}^{(2)}(z_i)\right) \tag{13}$$

Let us treat the four terms of this decomposition. The term (10) has been treated by Propostion 3.3, and satisfies, under the null hypothesis H_0

$$\sqrt{n}\Delta_n(\mathbf{t}_j^{(1)}, t_j^{(2)}) \to_{n\,!\,1} \mathcal{N}\left(0, \mathsf{E}\left[\left(k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X}) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z\right]\right)\left(k_Y(t_j^{(2)}, Y) - \mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z\right]\right)\right]\right) .$$

Let us now show that the last term (13) converges towards 0 in probability. Let us denote for all j, $e_j^{(1)} : z \to \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = z\right]$ and $e_j^{(2)} : z \to \mathsf{E}_Y\left[k_{\ddot{X}}(t_j^{(2)}, Y)|Z = z\right]$, both elements of H_Z by Assumption 3.5. Then we have, for all $i \in [n]$:

$$\left(e_j^{(1)}(z_i) - h_{j,r}^{(1)}(z_i)\right)\left(e_j^{(2)}(z_i) - h_{j,r}^{(2)}(z_i)\right) = \langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), k_Z(z_i, \cdot) \otimes k_Z(z_i, \cdot)\rangle.$$

Then we deduce, by denoting: $\mu_{ZZ} := \mathsf{E}\left[k_Z(Z, \cdot)k_Z(Z, \cdot)\right]$ and $\hat{\mu}_{ZZ} := \frac{1}{n}\sum_{i=1}^{n} k_Z(z_i, \cdot)k_Z(z_i, \cdot)$, that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = z_i\right] - h_{j,r}^{(1)}(z_i)\right)\left(\mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = z_i\right] - h_{j,r}^{(2)}(z_i)\right)$$

$$= \langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \frac{1}{n}\sum_{i=1}^{n} k_Z(z_i, \cdot) \otimes k_Z(z_i, \cdot)\rangle$$

$$= \langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \mu_{ZZ}\rangle + \langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \hat{\mu}_{ZZ} - \mu_{ZZ}\rangle .$$

Then remark that:

$$|\langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \mu_{ZZ}\rangle| = |\mathsf{E}_Z\left[\left(e_j^{(1)}(Z) - h_{j,r}^{(1)}(Z)\right)\left(e_j^{(2)}(Z) - h_{j,r}^{(2)}(Z)\right)\right]|$$

$$\leq \|e_j^{(1)} - h_{j,r}^{(1)}\|_{L^2(P_Z)}\|e_j^{(2)} - h_{j,r}^{(2)}\|_{L^2(P_Z)} .$$

Under the Assumptions 3.4-3.5, for $\lambda_r = \frac{1}{r^{\beta+\gamma}}$, we have, using the results from (Fischer and Steinwart, 2020, Theorem 1): $\|e_j^{(1)} - h_{j,r}^{(1)}\|_{L^2(P_Z)}^2 \leq \frac{C\tau^2}{r^{\frac{\beta}{\beta+\gamma}}}$ with probability $1 - 4e^{\ \tau}$ and $\|e_j^{(2)} - h_{j,r}^{(2)}\|_{L^2(P_Z)}^2 \leq \frac{C\tau^2}{r^{\frac{\beta}{\beta+\gamma}}}$ with probability $1 - 4e^{\ \tau}$, for some constant C independent from n and τ. then by union bound, we deduce with probability $1 - 8e^{\ \tau}$ we have:

$$\sqrt{n}|\langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \mu_{ZZ}\rangle| \leq \sqrt{n}\frac{C^2\tau^4}{r^{\frac{\beta}{\beta+\gamma}}} .$$

Then, if $\sqrt{n} \in o(r^{\frac{\beta}{\beta+\gamma}})$, we have: $\sqrt{n}|\langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \mu_{ZZ}\rangle| \to 0$ in probability when $n \to \infty$. Moreover:

$$|\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \hat{\mu}_{ZZ} - \mu_{ZZ}\rangle| \leq \|e_j^{(1)} - h_{j,r}^{(1)}\|_{H_Z} \|e_j^{(2)} - h_{j,r}^{(2)}\|_{H_Z} \|\hat{\mu}_{ZZ} - \mu_{ZZ}\|_{H_Z \quad H_Z},$$

and by Markov inequality, $\|\hat{\mu}_{ZZ} - \mu_{ZZ}\|_{H_Z \quad H_Z} \leq \sqrt{\frac{C'}{n\delta}}$ with probability $1-\delta$ for some constant C'. Moreover, under Assumption 3.4-3.5, we have $\|e_j^{(1)} - h_{j,r}^{(1)}\|_{H_Z} \to 0$ and $\|e_j^{(2)} - h_{j,r}^{(2)}\|_{H_Z} \to 0$ in probability. Then, we deduce that $\sqrt{n}|\langle\left(e_j^{(1)} - h_{j,r}^{(1)}\right) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \hat{\mu}_{ZZ} - \mu_{ZZ}\rangle| \to 0$ in probability. Finally, the term (13) goes to 0 in probability.

The terms (11) and (12) are similar and can be treated the same way. We only focus on the term (11). For all $i \in [n]$:

$$\begin{aligned}
&|\frac{1}{n}\sum_{i=1}^{n}\left(k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{x}_i) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z = z_i\right]\right)\left(\mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z = z_i\right] - h_{j,r}^{(2)}(z_i)\right)| \\
&= |\frac{1}{n}\sum_{i=1}^{n}\langle k_{\ddot{X}}(t_j^{(1)}, \cdot), k_{\ddot{X}}(\ddot{x}_i, \cdot) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\ddot{X}, \cdot)|Z = z_i\right]\rangle_{H_{\ddot{X}}} \langle e_j^{(2)} - h_{j,r}^{(2)}, k_Z(z_i, \cdot)\rangle_{H_Z}| \\
&= |\frac{1}{n}\sum_{i=1}^{n}\langle k_{\ddot{X}}(t^{(1)}, \cdot) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \left(k_{\ddot{X}}(\ddot{x}_i, \cdot) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\ddot{X}, \cdot)|Z = z_i\right]\right) \otimes k_Z(z_i, \cdot)\rangle_{H_{\ddot{X}} \quad H_Z}| \\
&= |\langle k_{\ddot{X}}(t^{(1)}, \cdot) \otimes \left(e_j^{(2)} - h_{j,r}^{(2)}\right), \frac{1}{n}\sum_{i=1}^{n}\left(k_{\ddot{X}}(\ddot{x}_i, \cdot) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\ddot{X}, \cdot)|Z = z_i\right]\right) \otimes k_Z(z_i, \cdot)\rangle_{H_{\ddot{X}} \quad H_Z}| \\
&\leq \|k_{\ddot{X}}(t^{(1)}, \cdot)\|_{H_{\ddot{X}}} \|e_j^{(2)} - h_{j,r}^{(2)}\|_{H_Z} \left(\|\hat{\mu}^1_{\ddot{X}Z} - \mu_{\ddot{X}Z}\|_{H_{\ddot{X}} \quad H_Z} + \|\hat{\mu}^2_{\ddot{X}} - \mu_{\ddot{X}Z}\|_{H_{\ddot{X}} \quad H_Z}\right)
\end{aligned}$$

where: $\hat{\mu}^1_{\ddot{X}Z} := \frac{1}{n}\sum_{i=1}^{n} k_{\ddot{X}}(\ddot{x}_i, \cdot) \otimes k_Z(z_i, \cdot)$, $\hat{\mu}^2_{\ddot{X}Z} := \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\ddot{X}, \cdot)|Z = z_i\right] \otimes k_Z(z_i, \cdot)$, and $\mu_{\ddot{X}Z} := \mathsf{E}\left[k_Y(y, \cdot)k_Z(z, \cdot)\right]$.

By the law of large numbers, we have: $\hat{\mu}^1_{\ddot{X}Z}$ and $\hat{\mu}^2_{\ddot{X}Z}$ converge almost surely towards $\mu_{\ddot{X}Z}$. Moreover by Markov inequality, $\|\hat{\mu}^1_{\ddot{X}Z} - \mu_{\ddot{X}Z}\|_{H_{\ddot{X}} \quad H_Z} \leq \sqrt{\frac{C}{n\delta}}$ with probability $1-\delta$, and $\|\hat{\mu}^2_{\ddot{X}Z} - \mu_{\ddot{X}Z}\|_{H_{\ddot{X}} \quad H_Z} \leq \sqrt{\frac{C}{n\delta}}$ with probability $1-\delta$. Then with probability $1-2\delta$, $\sqrt{n}\left(\|\hat{\mu}^1_{\ddot{X}Z} - \mu_{\ddot{X}Z}\|_{H_{\ddot{X}} \quad H_Z} + \|\hat{\mu}^2_{\ddot{X}Z} - \mu_{\ddot{X}Z}\|_{H_{\ddot{X}} \quad H_Z}\right) \leq 2\sqrt{\frac{C}{\delta}}$. Moreover, under Assumption 3.4-3.5, using the results from (Fischer and Steinwart, 2020), we have that $\|e_j^{(2)} - h_{j,r}^{(2)}\|_{H_Z}$ converges towards 0 in probability. Then the term (11) converges in probability towards 0. The same reasoning holds for (12).

Finally, by Slutsky's Lemma:

$$\sqrt{n}\widetilde{\Delta}_{n,r}(\mathbf{t}_j^{(1)}, t_j^{(2)}) \to_{n\,!\,1} \mathcal{N}\left(0, \mathsf{E}\left[\left(k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X}) - \mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_j^{(1)}, \ddot{X})|Z\right]\right)\left(k_Y(t_j^{(2)}, Y) - \mathsf{E}_Y\left[k_Y(t_j^{(2)}, Y)|Z\right]\right)\right]\right).$$

Now we have $\widetilde{\mathbf{S}}_{n,r} = \left(\widetilde{\Delta}_{n,r}(\mathbf{t}_j^{(1)}, t_j^{(2)})\right)_{j2[J]} = \left(\Delta_n(\mathbf{t}_j^{(1)}, t_j^{(2)})\right)_{j2[J]} + \left(\widetilde{\Delta}_{n,r}(\mathbf{t}_j^{(1)}, t_j^{(2)}) - \Delta_n(\mathbf{t}_j^{(1)}, t_j^{(2)})\right)_{j2[J]}$ and we have shown that $\sqrt{n}\left(\widetilde{\Delta}_{n,r_n}(\mathbf{t}_j^{(1)}, t_j^{(2)}) - \Delta_n(\mathbf{t}_j^{(1)}, t_j^{(2)})\right)_{j2[J]}$ goes to 0 in probability. Then by Slutsky's Lemma and Proposition 3.3, we get: $\widetilde{\mathbf{S}}_{n,r_n} \to \mathcal{N}(0, \mathbf{\Sigma})$.

Let $r > 0$. Under H_1, $\mathbf{S}_{n,r_n} \to \mathbf{S} \neq 0$. Let us consider a realization of $(\mathbf{t}_j^{(1)}, t_j^{(2)})_{j2[J]}$ such that $\|\mathbf{S}\|_p \neq 0$. So $P(n^{p/2}\|\mathbf{S}_{n,r_n}\|_p \geq r) \to 1$ as $n \to \infty$ because $\|\mathbf{S}\|_p \neq 0$. □

A.3. Proof of Proposition 3.7

Proof. First notice that:

$$\begin{aligned}
\widetilde{\mathbf{\Sigma}}_{n,r} &:= \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathbf{u}}_{i,r}\widetilde{\mathbf{u}}_{i,r}^T + \delta_n \mathrm{Id}_J \\
&= \widehat{\mathbf{\Sigma}}_n + \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{u}}_i\left(\widetilde{\mathbf{u}}_{i,r} - \widehat{\mathbf{u}}_r\right)^T + \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\mathbf{u}}_{i,r} - \widehat{\mathbf{u}}_r\right)\widehat{\mathbf{u}}_i^T + \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\mathbf{u}}_{i,r} - \widehat{\mathbf{u}}_r\right)\left(\widetilde{\mathbf{u}}_{i,r} - \widehat{\mathbf{u}}_r\right)^T + \delta_n \mathrm{Id}_J\,.
\end{aligned}$$

By the law of large numbers, we get that under H_0: $\widehat{\mathbf{\Sigma}}_n \to \mathbf{\Sigma}$. Moreover:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{u}}_i\left(\widetilde{\mathbf{u}}_{i,r}-\widehat{\mathbf{u}}_r\right)^T\right]_{kl}=\frac{1}{n}\sum_{i=1}^{n}\left(k_Y(t_k^{(2)},y_i)-\mathsf{E}_Y\left[k_Y(t_k^{(2)},Y)|Z=z_i\right]\right)\left(\mathsf{E}_{\ddot{X}}\left[k_{\ddot{X}}(\mathbf{t}_l^{(1)},\ddot{X})|Z=z_i\right]-h_{l,r}^{(1)}(z_i)\right)$$

which has been proven to converge in probability to 0 in the proof of Proposition 3.6. Then $\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{u}}_i\left(\widetilde{\mathbf{u}}_{i,r}-\widehat{\mathbf{u}}_r\right)^T$ converges in probability to 0. Similarly $\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\mathbf{u}}_{i,r}-\widehat{\mathbf{u}}_r\right)\widehat{\mathbf{u}}_i^T$ and $\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{\mathbf{u}}_{i,r}-\widehat{\mathbf{u}}_r\right)\left(\widetilde{\mathbf{u}}_{i,r}-\widehat{\mathbf{u}}_r\right)^T$ also converge in probability to 0. Then by Slutsky's Lemma, $\widetilde{\mathbf{\Sigma}}_{n,r}$ converges in probability to $\mathbf{\Sigma}$. By Slutsky's Lemma (again) and by Propostion 3.6, we have that: $\widetilde{\mathbf{\Sigma}}_{n,r}^{\frac{1}{}}\widetilde{\boldsymbol{S}}_{n,r}$ converges to a standard gaussian distribution $\mathcal{N}(0,\mathrm{Id})$. The second part of the proposition is the same as the proof of Proposition 3.6. □

B. On the computation of Oracle statistic in Figure 2

To compute the oracle statistic we needed to compute exactly the conditional expectation implied in our statistic. In the case of gaussian kernels and gaussian distributed data for Z, the computation of this conditional expectation is reduced to the computation of moment-generating function of a non-centered χ^2 distribution.

B.1. Additional experiments on Problems (4) and (5)

B.1.1. Gaussian Case

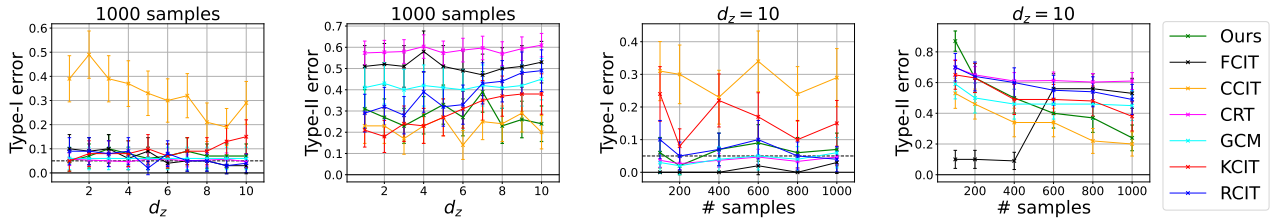

Figure 7. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line) and the type-II error (lower is better) of our test procedure with other SoTA tests on the two problems presented in (4) and (5) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): type-I and type-II errors obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals to 1000. (*Middle-right, right*): type-I and type-II errors obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals to 10.

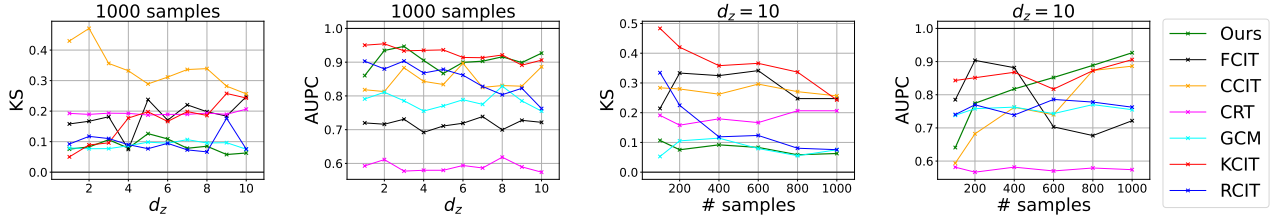

Figure 8. Comparison of the KS statistic (lower is better) and the AUPC (higher is better) of our testing procedure with other SoTA tests on the two problems presented in (4) and (5) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): the KS and AUPC obtained by each test when varying the dimension d_z from 1 to 10, while fixing the number of samples n to 1000. (*Middle-right, right*): the KS and AUPC obtained by each test when varying the number of samples n from 100 to 1000, while fixing the dimension d_z to 10.

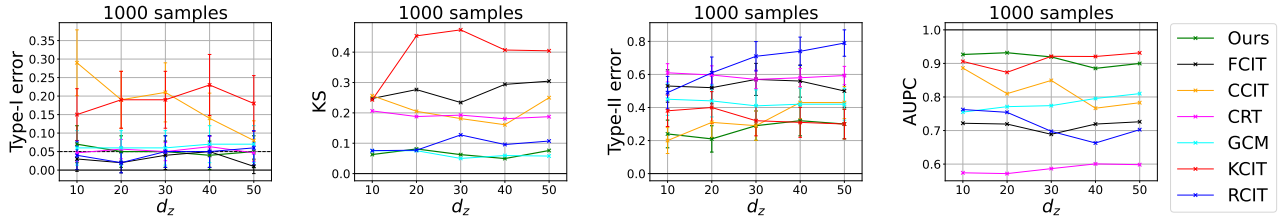

Figure 9. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line), type-II error (lower is better), KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (4) and Eq. (5) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. In each plot the dimension d_z is varying from 10 to 50; here, the number of samples n is fixed and equals to 1000.

B.1.2. LAPLACE CASE

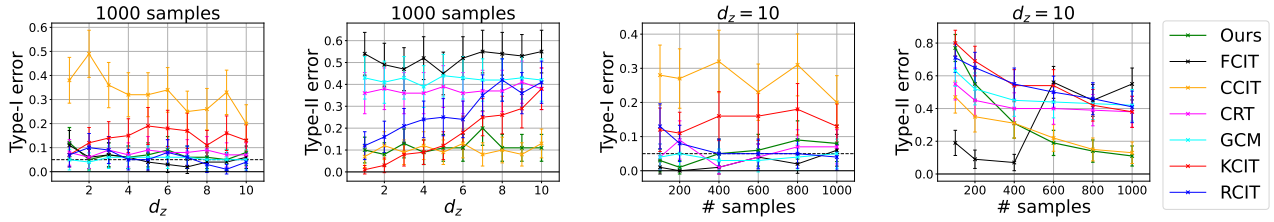

Figure 10. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line) and the type-II error (lower is better) of our test procedure with other SoTA tests on the two problems presented in (4) and (5) with Laplace noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): type-I and type-II errors obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals to 1000. (*Middle-right, right*): type-I and type-II errors obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals to 10.

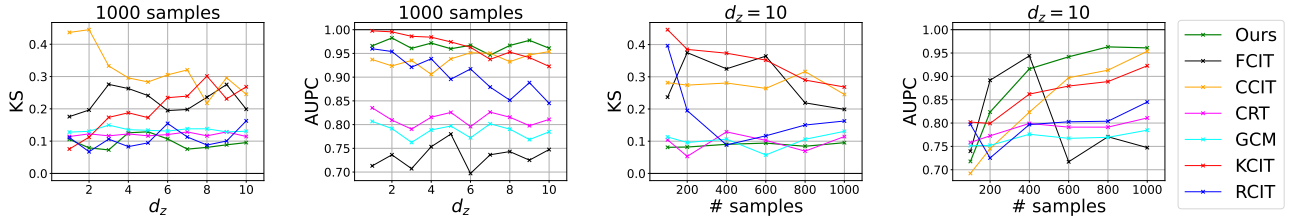

Figure 11. Comparison of the KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (4) and Eq. (5) with Laplace noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): the KS statistic and AUPC (respectively) obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals to 1000. (*Middle-right, right*): the KS and AUPC (respectively), obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals to 10.

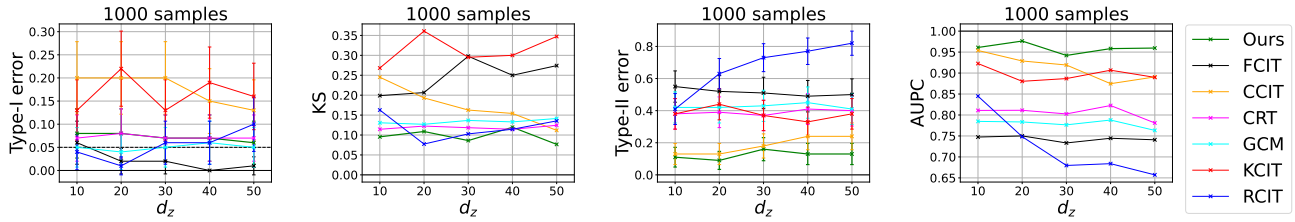

Figure 12. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line), type-II error (lower is better), KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (4) and Eq. (5) with Laplace noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. In each plot the dimension d_z is varying from 10 to 50; here, the number of samples n is fixed and equals to 1000.

B.2. Additional experiments on Problems (8) and (9)

B.2.1. Gaussian Case

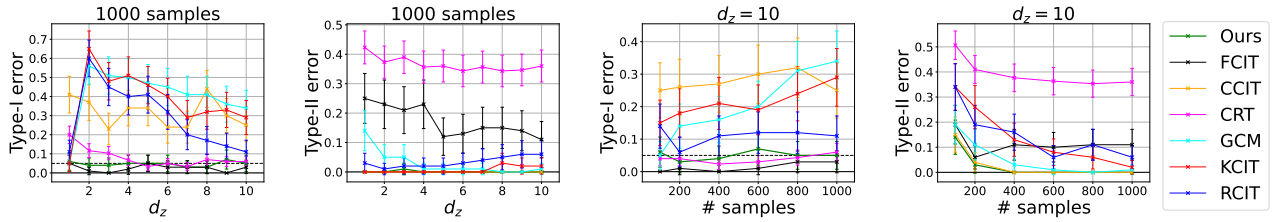

Figure 13. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line) and the type-II error (lower is better) of our test procedure with other SoTA tests on the two problems presented in (8) and (9) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): type-I and type-II errors obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals to 1000. (*Middle-right, right*): type-I and type-II errors obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals to 10.

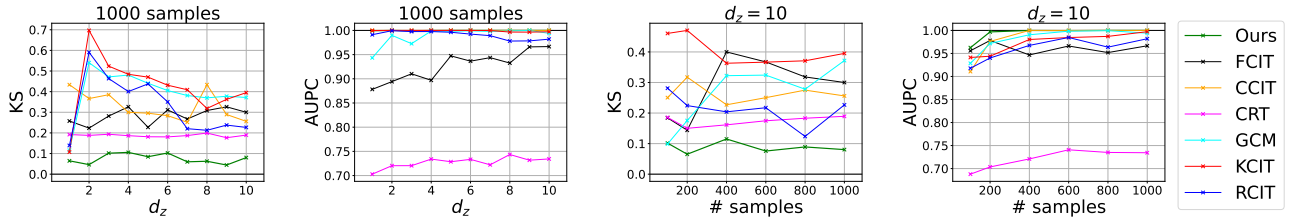

Figure 14. Comparison of the KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (8) and Eq. (9) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): the KS statistic and AUPC (respectively) obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals to 1000. (*Middle-right, right*): the KS and AUPC (respectively), obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals to 10.

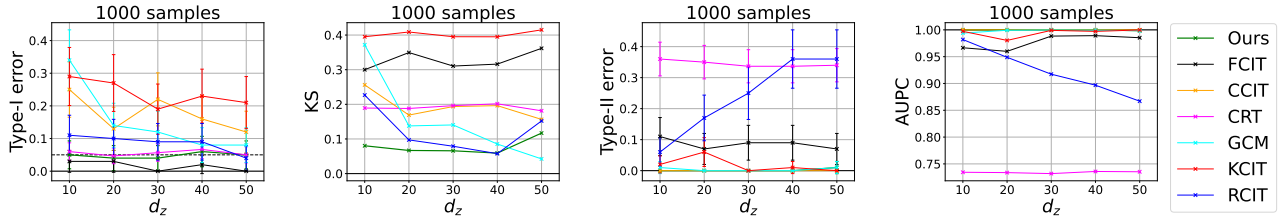

Figure 15. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line), type-II error (lower is better), KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (8) and Eq. (9) with Gaussian noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. In each plot the dimension d_z is varying from 10 to 50; here, the number of samples n is fixed and equals to 1000.

B.2.2. LAPLACE CASE

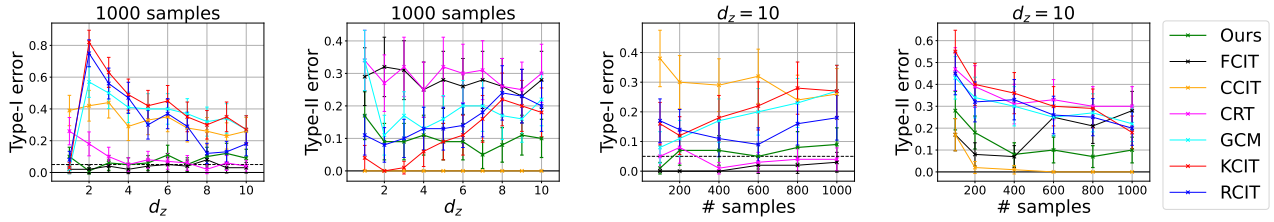

Figure 16. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line) and the type-II error (lower is better) of our test procedure with other SoTA tests on the two problems presented in (8) and (9) with Laplace noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): type-I and type-II errors obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals to 1000. (*Middle-right, right*): type-I and type-II errors obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals to 10.

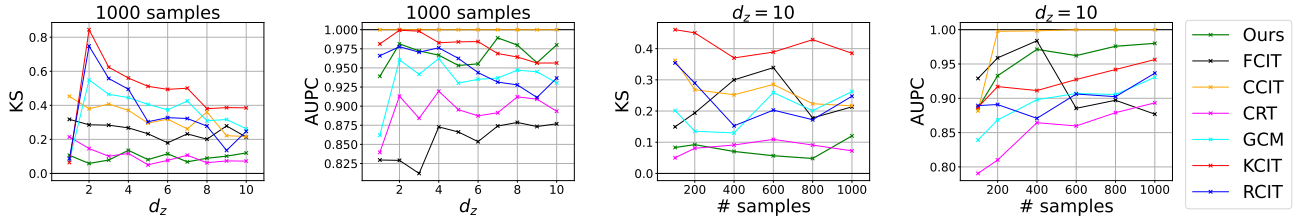

Figure 17. Comparison of the KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (8) and Eq. (9) with Laplace noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. (*Left, middle-left*): the KS statistic and AUPC (respectively) obtained by each test when varying the dimension d_z from 1 to 10; here, the number of samples n is fixed and equals to 1000. (*Middle-right, right*): the KS and AUPC (respectively), obtained by each test when varying the number of samples n from 100 to 1000; here, the dimension d_z is fixed and equals to 10.

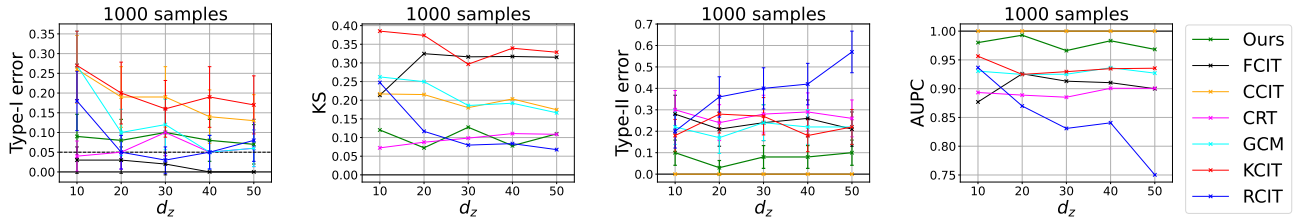

Figure 18. Comparison of the type-I error at level $\alpha = 0.05$ (dashed line), type-II error (lower is better), KS statistic and the AUPC of our testing procedure with other SoTA tests on the two problems presented in Eq. (8) and Eq. (9) with Laplace noises. Each point in the figures is obtained by repeating the experiment for 100 independent trials. In each plot the dimension d_z is varying from 10 to 50; here, the number of samples n is fixed and equals to 1000.