An $\ell^p$-based Kernel Conditional Independence Test

Meyer Scetbon$^{*1}$  Laurent Meunier$^{*2,3}$  Yaniv Romano$^4$

Abstract

We propose a new computationally efficient test for conditional independence based on the $\ell^p$ distance between two kernel-based representatives of well suited distributions. By evaluating the difference of these two representatives at a finite set of locations, we derive a finite dimensional approximation of the $\ell^p$ metric, obtain its asymptotic distribution under the null hypothesis of conditional independence and design a simple statistical test from it. The test obtained is consistent and computationally efficient. We conduct a series of experiments showing that the performance of our new tests outperforms state-of-the-art methods both in term of statistical power and type-I error 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 $Z$. This can be formulated as a hypothesis testing problem:

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

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, Markowitz and Spang 2007).

Testing for $H_0$ in such applications is known to be a highly challenging task (Bergsma 2004, Shah et al. 2020). First, 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. Second, 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 motivate the development of a series of practical methods with powerful test statistics, attempting to reliably test for conditional independence. These include kernel-based tests (Fukumizu et al. 2008, Zhang et al. 2012, Doran et al. 2014, Strobl et al. 2019), model-based methods (Sen et al. 2017, 2018, Chalupka et al. 2018, Shah et al. 2020), and sampling strategies (Candès et al. 2018, Bellot and van der Schaar 2019, Shi et al. 2020).

In addition, estimating conditional independence may become computationally expensive as the dimension of the problem increases. For example, the Kernel Conditional Independence Test (KCIT) (Zhang et al. 2012) requires a cubic number of operations with respect to the sample size. Moreover, in this setting, any bootstrap or permutation procedure becomes expensive as one needs to compute multiple times the statistic of interest. Such strategies are often used to estimate the p-value as the (asymptotic) null distribution, which may be parameter dependent (Sen et al. 2017, Shi et al. 2020). In that context, Strobl et al. (2019) propose to reduce the computational complexity of KCIT by considering a Random Fourier features approximations of the kernels involved in the statistic, resulting in a test that scales linearly with respect to the sample size.

In this paper, we propose a new kernel-based test for conditional independence with asymptotic theoretical guarantees. Taking inspiration from Scetbon and Varoquaux (2019), we use the $\ell^p$ distance between two well chosen kernel mean embeddings evaluated at a finite set of locations. We show that this metric encodes the conditional dependence relation of the random variables under study. From this metric, we derive a computationally efficient test statistic which can scales linearly with respect to the number of samples. Under common assumptions on the richness of the RKHS, we derive the

1 CREST, ENSAE 2 LAMSADE, Université Paris-Dauphine 3 Facebook AI Research 4 Technion. Correspondence to: Meyer Scetbon <meyer.scetbon@ensae.fr>.
asymptotic null distribution of our statistic, 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) proposed a kernel based-test (KCT), by leveraging the characterization of conditional independence derived in (Daudin 1980) to form a test statistic. The authors of this work analyzed the asymptotic distribution of the proposed statistic, offering a practical procedure to test for $H_0$. 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. However, this test is computationally expensive since both the estimation of the p-value as well as the computation of the statistic require at least a cubic number of operations. In our work, we propose a new kernel-based statistic that is computationally efficient, which scales linearly with the number of samples. 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. In addition, the p-values of the proposed test are easy to obtain: we show that the asymptotic null distribution of our statistic follows the standard normal distribution.

Other CI tests proposed in the literature suggest testing relaxed forms of conditional independence. For instance, Shah et al. (2020) assumes that $X$ and $Y$ are linked to $Z$ through a regression model, and Sen et al. (2017) consider model-based methods to generate samples from the conditional distribution. However, these methods do not provide theoretical guarantees on the asymptotic null distribution of the underlying test statistic. In contrast, we derive the asymptotic null distribution of our test statistic by adding sufficient assumptions on the RKHS considered, and design 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 $(\mathcal{D}, \mathcal{A})$ a Borel measurable space and denote $\mathcal{M}_+^1(\mathcal{D})$ the space of Borel probability measures on $\mathcal{D}$. Let also $(H, k)$ 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 $\mathbb{E}_{x \sim \nu} [\sqrt{k(x,x)}]$ is finite, we define for all $t \in \mathcal{D}$ the mean embedding as $\mu_{\nu,k}(t) := \int_{\mathcal{D}} \sqrt{k(x,t)} d\nu(x)$. Note that $\mu_{\nu,k}$ is the unique element in $H$ satisfying for all $f \in H$, $\mathbb{E}_{x \sim \nu} (f(x)) = \langle \mu_{\nu,k}, f \rangle_H$. 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 (Chwialkowski et al. 2015; Jitkrittum et al. 2017; Scetbon and Varoquaux 2019).

$L^p$-distance between mean embeddings. Let $k$ be a definite positive, characteristic, and bounded kernel on $\mathbb{R}^d$ and $p \geq 1$ an integer. Scetbon and Varoquaux (2019) showed that given an absolutely continuous Borel probability measure $\Gamma$ on $\mathbb{R}^d$, the following function defined for any $(P,Q) \in \mathcal{M}_+^1(\mathbb{R}^d) \times \mathcal{M}_+^1(\mathbb{R}^d)$ as

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

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

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

where $(t_j)_{j=1}^{J}$ are sampled independently from the $\Gamma$ distribution, is a random metric on $\mathcal{M}_+^1(\mathbb{R}^d)$.

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

3 A new $\ell_p$ kernel-based testing

In this section, we present our statistical procedure to test for conditional independence. We begin by

---

1 An analytic kernel on $\mathbb{R}^d$ is a positive definite kernel such that for all $x \in \mathbb{R}^d$, $k(x, \cdot)$ 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.
introducing a general measure based on the $L^p$ distance 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 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 $P_{X\otimes Y|Z}$ on $\mathcal{X} \times \mathcal{Y}$ as

$$P_{X\otimes Y|Z}(X \times Y) := E_Z \left[ E_{\tilde{X}\otimes Y|Z}[1_A|Z|E_{Y|Z}[1_B|Z]] \right],$$

for any $(A, B) \in B(\mathcal{X}) \times B(\mathcal{Y})$, where $1_A$ is the characteristic function of a measurable set $A$ and similarly for $B$. One now characterize the independence of $X \otimes Y$ given $Z$ as follows: $X \perp Y|Z$ if and only if $P_{X\otimes Y} = P_{X\otimes Y|Z}$ \cite{Fukumizu2008}. Moreover, for an appropriate kernel $k$ and an absolutely continuous Borel probability measure $\Gamma$ on $\mathcal{X} \times \mathcal{Y}$, $d_p(\cdot, \cdot)$ defined in (1) is a metric on $\mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y})$ \cite{Scetbon2019}. Therefore, we have a first simple characterization of the conditional independence: $X \perp Y|Z$ if and only if $d_p(P_{X\otimes Y}, P_{X\otimes Y|Z}) = 0$. With this in place, we now state some assumptions on the kernel $k$ considered in the rest of this paper.

Assumption 1. The kernel $k : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is definite positive, characteristic, bounded, continuous and analytic. Moreover, the kernel $k$ is a tensor product of kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ on $\mathcal{X}$ 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 $k_\mathcal{X}$ and $k_\mathcal{Y}$ are characteristic, bounded, continuous and analytic \cite{Szabo2018}. If the kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ are Gaussian kernels \cite{Titikrittum2017}, then $k = k_\mathcal{X} \otimes k_\mathcal{Y}$ satisfies Assumption 1 \cite{Titikrittum2017}. In fact, using the analyticity of the kernel $k$, one can work with $d_{p,J}$ from (2) instead of $d_p$ to characterize the conditional independence.

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

Proof. Recall that $X \perp Y|Z$ if and only if $P_{X\otimes Y} = P_{X\otimes Y|Z}$ \cite{Fukumizu2008}. If $k$ is bounded, characteristic, and analytic, then, by invoking \cite{Scetbon2019} Theorem 4) we get that $d_p^p$ is a random metric on the space of Borel probability measures. This concludes the proof.

The key advantage of using $d_{p,J}(P_{X\otimes Y}, P_{X\otimes Y|Z})$ to measure the conditional dependence is that it only requires to compute the differences between the mean embeddings of $P_{X\otimes Y}$ and $P_{X\otimes Y|Z}$ 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 is the tensor product of two kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ on $\mathcal{X}$ and $\mathcal{Y}$, respectively, we can obtain a simple expression of our measure $d_{p,J}(P_{X\otimes Y}, P_{X\otimes Y|Z})$. Indeed, in that case, the mean embedding of $P_{X\otimes Y|Z}$ can be expressed for any $(t^{(1)}, t^{(2)}) \in \mathcal{X} \times \mathcal{Y}$ as:

$$\mu_{P_{X\otimes Y|Z}}(k_{\mathcal{X}}(t^{(1)}, \tilde{X}), k_{\mathcal{Y}}(t^{(2)}, Y)) = E_{\tilde{X}} \left[ k_{\mathcal{X}}(t^{(1)}, \tilde{X})|Z \right] E_{Y} \left[ k_{\mathcal{Y}}(t^{(2)}, Y)|Z \right].$$

Therefore, under Assumption 1 by defining the witness function:

$$\mu(t^{(1)}, t^{(2)}) := E \left[ k_{\mathcal{X}}(t^{(1)}, \tilde{X}) - E_{\tilde{X}} \left[ k_{\mathcal{X}}(t^{(1)}, \tilde{X})|Z \right] \right] \times \left( k_{\mathcal{Y}}(t^{(2)}, Y) - E_{Y} \left[ k_{\mathcal{Y}}(t^{(2)}, Y)|Z \right] \right),$$

we get that

$$d_{p,J}(P_{X\otimes Y}, P_{X\otimes Y|Z}) = \left( \frac{1}{J} \sum_{j=1}^J \left| \mu(t^{(1)}, t^{(2)}) \right|^p \right)^{1/p},$$

where $(\{t^{(1)}_j, t^{(2)}_j\})_{j=1}^J$ are sampled independently according to $\Gamma$.

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

$$\tilde{\mu}_n(t^{(1)}, t^{(2)}) := \frac{1}{n} \sum_{i=1}^n \left( k_{\mathcal{X}}(t^{(1)}, \tilde{x}_i) - E \left[ k_{\mathcal{X}}(t^{(1)}, \tilde{X})|z_i \right] \right) \times \left( k_{\mathcal{Y}}(t^{(2)}, y_i) - E \left[ k_{\mathcal{Y}}(t^{(2)}, Y)|z_i \right] \right).$$
With this in place, a natural candidate to estimate $d_{n,p}^0(P_{XY}, P_{X@Y|Z})$ (up to the constant $J$) can be expressed as
\[
\tilde{C}_{I,n,p} := \frac{1}{J} \sum_{j=1}^{J} \left| \tilde{\mu}_n(t_j^{(1)}, t_j^{(2)}) \right|^p,
\]
where $(t_j^{(1)}, t_j^{(2)}) \in \tilde{X} \times \tilde{Y}$ are sampled independently from $\Gamma$.

We now turn to derive the asymptotic distribution of this statistic. For that purpose, define, for all $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, n\}$,
\[
u_i(j) := \left( k_X(t_j^{(1)}, \tilde{x}_i) - \mu_X \left[ k_X(t_j^{(1)}, \tilde{X}) | Z = z_i \right] \right) \times \left( k_Y(t_j^{(2)}, y_i) - \mu_Y \left[ k_Y(t_j^{(2)}, Y) | Z = z_i \right] \right),
\]
\[
u_i := (\nu_1(1), \ldots, \nu_n(J))^T\text{ and } \Sigma := \mathbb{E}(\nu_i \nu_i^T).
\]
We also denote by $\tilde{S}_n := \frac{1}{n} \sum_{i=1}^n \nu_i$. Now, observe that $\tilde{C}_{I,n,p} = \|\tilde{S}_n\|^p_p$.

**Proposition 2.** Suppose that Assumption 7 is verified. Let $((t_1^{(1)}, t_1^{(2)}), \ldots, (t_J^{(1)}, t_J^{(2)})) \in (\tilde{X} \times \tilde{Y})$. Then, under $H_0$, we have: $\sqrt{n} \tilde{S}_n \rightarrow \mathcal{N}(0, \Sigma)$. Moreover, under $H_1$, if $(t_1^{(1)}, t_1^{(2)})^{T}_{j=1} = \tilde{X}$ are sampled independently according to $\Gamma$, then $\Gamma$-almost surely, for any $q \in \mathbb{R}$, $\lim_{n \rightarrow \infty} P(n^{p/2} \tilde{C}_{I,n,p} > q) = 1$.

Proof. Recall that $\tilde{S}_n = \frac{1}{n} \sum_{i=1}^n \nu_i$ where $\nu_i$ are i.i.d. samples. Under $H_0$, $\mathbb{E}[\nu_i] = 0$. Using the Central Limit Theorem, we get: $\sqrt{n} \tilde{S}_n \rightarrow \mathcal{N}(0, \Sigma)$. Using the analyticity of the kernel $k$, under $H_1$, $\Gamma$-almost surely, there exists a $j \in \{1, \ldots, J\}$ such that $\mathbb{E}[\nu_j] \neq 0$. Therefore, we can deduce that $\Gamma$-almost surely, $S = \mathbb{E}[\nu_1] \neq 0$. Now, for all $r > 0$, we get: $P(n^{p/2} \tilde{C}_{I,n,p} > r) \rightarrow 1$ because $\tilde{C}_{I,n,p} \rightarrow \|S\|^p_p$ when $n \rightarrow \infty$.

From the above proposition, we can define a consistent statistical test at level $0 < \alpha < 1$, by rejecting the null hypothesis if $n^{p/2} \tilde{C}_{I,n,p}$ 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 $\mathcal{N}(0, \Sigma)$. However, in practice, $\tilde{C}_{I,n,p}$ cannot be computed as it requires the access to samples from the conditional means involved in the statistic, 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

Our goal here is to estimate $\mathbb{E}_X \left[ k_X(t_j^{(1)}, \tilde{X}) | Z = \cdot \right]$ and $\mathbb{E}_Y \left[ k_Y(t_j^{(2)}, Y) | Z = \cdot \right]$ 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 $H_{H_j}^1$ and $H_{H_j}^2$ two separable RKHSs on $\mathcal{Z}$. Denote also by $k_{H_j}^1$ and $k_{H_j}^2$ their associated kernels and $\lambda_{H_j}^r, \lambda_{H_j}^r > 0$ the regularization parameters involved in the RLS regressions. Then, the RLS estimators are the unique solutions of the following problems:
\[
\min_{h \in H_{H_j}^2} \frac{1}{r} \sum_{i=1}^r \left( h(z_i) - k_{H_j}^1(x_i, z_i) \right)^2 + \lambda_{H_j}^r \|h\|_{H_{H_j}^2}^2,
\]
\[
\min_{h \in H_{H_j}^2} \frac{1}{r} \sum_{i=1}^r \left( h(z_i) - k_{H_j}^2(x_i, z_i) \right)^2 + \lambda_{H_j}^r \|h\|_{H_{H_j}^2}^2,
\]
which we denote by $h_{H_j}^1$ and $h_{H_j}^2$, respectively. These estimators have simple expressions in term of the kernels involved. For example, let $k_{H_j}^1(t_j^{(1)}, \tilde{X}_r) := [k_{H_j}^1(t_j^{(1)}, (x_1, z_1)), \ldots, k_{H_j}^1(t_j^{(1)}, (x_r, z_r))]^T$, for any $z \in \mathcal{Z}$, the estimator $h_{H_j}^1$ can be expressed as
\[
h_{H_j}^1(z) = \sum_{i=1}^r \alpha_{H_j}^1(z_i) k_{H_j}^1(z_i, z),
\]
with
\[
\alpha_{H_j}^1 := (k_{H_j}^1(z_i, z_i))_{1 \leq i, j \leq r},
\]
where $k_{H_j}^1 := (k_{H_j}^1(z_i, z_j))_{1 \leq i, j \leq r}$. Note that once $\alpha_{H_j}^1$ is available, evaluating the RLS estimator requires only $O(rd)$ operations, where $d$ corresponds to the computational cost of evaluating the kernel $k_{H_j}^1$. Moreover, $\alpha_{H_j}^1$ can be evaluated in at most $O(r^2 d + r^3)$ algebraic operations. We will use the above for the complexity analysis of our methods, although one can apply the Coppersmith–Winograd algorithm [Coppersmith and Winograd, 1990] that reduces the computational cost of order $O(r^2d + r^{2.376})$ to compute $\alpha_{H_j}^1$.

With this in place, we can now introduce our new estimator of the witness function $\mu$ at each location $(t_j^{(1)}, t_j^{(2)})$ as follows:
\[
\tilde{\mu}_n,r(t_j^{(1)}, t_j^{(2)}) := \frac{1}{n} \sum_{i=1}^n \left( k_X(t_j^{(1)}, \tilde{x}_i) - h_{H_j}^1(z_i) \right) \times \left( k_Y(t_j^{(2)}, y_i) - h_{H_j}^2(z_i) \right),
\]
which can be evaluated in $O(nrd + r^2d + r^3)$ operations. Here, the proposed test statistic is
\[
\tilde{C}_{I,n,p,r} := \left| \tilde{\mu}_n,r(t_j^{(1)}, t_j^{(2)}) \right|^p_p,
\]
whose computational complexity is in $O(J(nrd + r^2d + r^3))$. Note that the total cost of computing $\tilde{C}_{I,n,p,r}$ scales linearly with the number of samples $n$, however, to obtain theoretical guarantees, we will see (Proposition 3) that $r$ depends on $n$. 

**An $\ell^p$-based Kernel Conditional Independence Test**
Asymptotic Distribution. To get the asymptotic distribution, we need to make two assumptions. Let us define, for $m \in \{1, 2\}$ and $j \in \{1, \ldots, J\}$, $I_{Z}^{m,j}=\text{the operator on } L^{2}(Z, P_{Z})$ as $I_{Z}^{m,j}(g)(\cdot) = \int_{Z}^{m,j}(\cdot, z) g(z) dP_{Z}(z)$.

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

$$\text{Tr}(I_{Z}^{m,j} + \lambda I_{Z}^{m,j}) \leq Q\lambda^{-\gamma}.$$  

Assumption 3. There exists $2 \geq \beta > 1$ such that for any $j \in \{1, \ldots, J\}$, $(t^{(1)}, t^{(2)}) \in \tilde{X} \times \mathcal{Y}$,

$$\mathbb{E}_{\tilde{X}}\left[k_{x}(t^{(1)}, \tilde{X})|Z = \cdot\right] \in \mathcal{R}\left([L_{Z}^{m,j}]^{\beta/2}\right),$$

$$\mathbb{E}_{\mathcal{Y}}\left[k_{y}(t^{(2)}, Y)|Z = \cdot\right] \in \mathcal{R}\left([L_{Z}^{m,j}]^{\beta/2}\right),$$

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

$$\mathbb{E}_{Z \sim z}\left[k_{x}(t^{(1)}, \tilde{X}) - \mathbb{E}_{\tilde{X} \sim z}\left[k_{x}(t^{(1)}, \tilde{X})\right]\right]^{l} \leq \frac{l \sigma^{2} L_{z}^{1-2}}{2},$$

$$\mathbb{E}_{Z \sim z}\left[k_{y}(t^{(2)}, Y) - \mathbb{E}_{Y \sim z}\left[k_{y}(t^{(2)}, Y)\right]\right]^{l} \leq \frac{l \sigma^{2} L_{z}^{1-2}}{2}.$$  

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 for both $L^{2}$ and RKHS norms, which are essential to guarantee that our new statistic $\tilde{C}_{n,r,p}$, estimated with RLS, has the same asymptotic law as our oracle estimator $\tilde{C}_{n,p}$.

To derive the asymptotic distribution of our new test statistic, we need to define for all $j \in \{1, \ldots, J\}$ and $i \in \{1, \ldots, n\}$, $\tilde{u}_{i,j}(r) := (k_{x}(t^{(1)}, \tilde{x}_{i}) - h_{j,r}(z_{i}))(k_{y}(t^{(2)}, y_{i}) - h_{j,r}(z_{i}))$, $\tilde{u}_{i,r} := (\tilde{u}_{i,r}(1), \ldots, \tilde{u}_{i,r}(J)^{T}$, and $S_{n,r} := \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_{i,r}$. 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. Suppose that Assumptions [1]–[3] are verified. Let $p \geq 1$, $J \geq 1$, $((t_{i}^{(1)}, t_{i}^{(2)}), \ldots, (t_{j}^{(1)}, t_{j}^{(2)})) \in (\tilde{X} \times \mathcal{Y})^{J}$, $r_{n}$ such that $n^{-\gamma} \in o(r_{n})$, $\lambda_{n} = r_{n}^{-1}$, and $\tilde{u}_{i,r} = (\tilde{u}_{i,r}(1), \ldots, \tilde{u}_{i,r}(J)^{T}$, and $S_{n,r} := \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_{i,r}$. 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 4. Suppose that Assumptions [1]–[3] are verified. Let $p \geq 1$, $J \geq 1$, $((t_{i}^{(1)}, t_{i}^{(2)}), \ldots, (t_{j}^{(1)}, t_{j}^{(2)})) \in (\tilde{X} \times \mathcal{Y})^{J}$, $r_{n}$ such that $n^{-\gamma} \in o(r_{n})$, $\lambda_{n} = r_{n}^{1-\gamma}$, and $\tilde{u}_{i,r} = (\tilde{u}_{i,r}(1), \ldots, \tilde{u}_{i,r}(J)^{T}$, and $S_{n,r} := \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_{i,r}$. 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.

3.4 Normalization of the Test Statistic

Herein, we consider a normalized variant of our statistic $\tilde{C}_{n,r,p}$. Denote for $\delta_{n} > 0$, $\tilde{S}_{n,r} := \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_{i,r} \tilde{u}_{i,r}^{T}$, and $\delta_{n} \tilde{I}_{d,j}$, then the normalized statistic considered is given by $\tilde{NCl}_{n,r,p} := \|	ilde{S}_{n,r}^{-1/2} \tilde{S}_{n,r}^{1/2}||^{p}$, which can be evaluated in $O(J(nrd+r^{2}d+r^{3}) + nJ^{2} + J^{3})$ algebraic operations. 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.2.

Proposition 4. Suppose that Assumptions [1]–[3] are verified. Let $p \geq 1$, $J \geq 1$, $((t_{i}^{(1)}, t_{i}^{(2)}), \ldots, (t_{j}^{(1)}, t_{j}^{(2)})) \in (\tilde{X} \times \mathcal{Y})^{J}$, $r_{n}$ such that $n^{-\gamma} \in o(r_{n})$, $\lambda_{n} = r_{n}^{-1}$, and $\tilde{u}_{i,r} = (\tilde{u}_{i,r}(1), \ldots, \tilde{u}_{i,r}(J)^{T}$, and $S_{n,r} := \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_{i,r}$. 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.

Statistical test at level $\alpha$: Compute $n^{p/2} \tilde{NCl}_{n,r,p}$, choose the threshold $\tau$ corresponding to the $(1 - \alpha)$ quantile of the asymptotic null distribution, and reject the null hypothesis whenever $n^{p/2} \tilde{NCl}_{n,r,p}$ is larger than $\tau$. For example, if $p = 2$, the threshold is the probability density function of a Nakagami distribution of parameters $m \geq 2$ and $\omega > 0$ is for all $\tau \geq 0$,

$$f(x, m, \omega) = \frac{\omega^{m}}{\Gamma(m)} x^{m-1} \exp(-\frac{\omega}{\tau} x^2),$$

where $G$ is the Euler Gamma function.
(1 − 𝜉)-quantile of χ²(𝐽), i.e., a sum of 𝐽 independent standard χ² variables.

3.5 Hyperparameters

The hyperparameters of our statistics N̂ CI_{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 {(𝑥ᵢ, 𝑧ᵢ, 𝑦ᵢ)}_{𝑖=1}^{𝑛}.

The first category includes both the choice of the locations ((𝑡ₓ, 𝑡𝑧)ᵢ, (𝑡𝑦)ᵢ)_{𝑖=1}^{𝐽} on which differences between the mean embeddings are computed and the choice of the kernels 𝑘ₓ and 𝑘𝑧. Each location 𝑡ₓ, 𝑡𝑧, 𝑡𝑦 is randomly chosen according to a Gaussian variable with mean and covariance of {𝑥ᵢ}_{𝑖=1}^{𝑛}, {𝑦ᵢ}_{𝑖=1}^{𝑛}, and {𝑧ᵢ}_{𝑖=1}^{𝑛}, respectively. As we consider Gaussian kernels, we should also choose the bandwidths. Here, we restrict ourselves to one-dimensional kernel bandwidths 𝜎ₓ, 𝜎𝑦, and 𝜎𝑧 for the kernels 𝑘ₓ, 𝑘𝑦, and 𝑘𝑧, respectively. More precisely, we select the median of {∥𝑥ᵢ−𝑥ⱼ∥}_{𝑖,𝑗=1}^{𝑛}, {∥𝑦ᵢ−𝑦ⱼ∥}_{𝑖,𝑗=1}^{𝑛}, and {∥𝑧ᵢ−𝑧ⱼ∥}_{𝑖,𝑗=1}^{𝑛} for 𝜎ₓ, 𝜎𝑦, and 𝜎𝑧, respectively.

The other category contains all the kernels 𝑘^{𝑚,j} and the regularization parameters 𝜆^{𝑚,j} 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 for only a few iterations (around 15) for choosing the hyperparameters involved in the RLS problems.

4 Experiments

The goal of this section is threefold: (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 2 and 4, 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 the conditional means involved in our test statistic. In these cases, we can compute our proposed oracle statistic Ĉ CI_{𝑛,𝑟,𝑝} and its normalized version, allowing us to show that under the null hypothesis we recover the theoretical asymptotic null distribution obtained in Proposition 2. We also reach similar conclusions regarding our approximate normalized test statistic, N̂ CI_{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 N̂ CI_{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.

Benchmarks. We consider 6 synthetic data sets and compare the power and type-I error of our test N̂ CI_{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 (Belbôt 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 H₀, 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 𝛼 = 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 H₁, we compute either the area under the power curve (AUPC) of the empirical cumulative density function.
We set the dimension of $\bar{\theta}$ 1000 times. In all the experiments, we set $\bar{\theta}$ to evaluate the statistical power, we consider instead we follow the synthetic experiment proposed in (Zhang et al., 2012; Li and Fan, 2020; Doran et al., 2014; Bellot and van der Schaaf, 2019). To compare type-I error rates, we generate simulated data for which $H_0$ is true:

$$X = f_1 (\tilde{Z} + \varepsilon_x), \quad Y = f_2 (\tilde{Z} + \varepsilon_y). \quad (4)$$

Above, $\tilde{Z}$ is the average of $Z = (Z_1, \cdots, Z_d)$, $\varepsilon_x$ and $\varepsilon_y$ are sampled independently from the standard normal distribution, and $f_1$ and $f_2$ are smooth functions chosen uniformly from the set $\{\cdot, \cdot^2, \cdot^3, \text{tanh}(\cdot), \exp(-|\cdot|)\}$. To evaluate the statistical power, we consider instead the following data generating function:

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

where $\varepsilon_b \sim \mathcal{N}(0, 1)$. In Figure 1 we compare the KS statistic and the AUPC of our method by varying its parameters to measure the type-I error and the power of the test. That figure shows that (i) our method is robust to the choice of $p$, and (ii) the power increases as $J$ increases. Armed with this observation, in the following experiments, we always set $p = 2$ and $J = 5$ for our method.

**Illustrations of our theoretical findings.** The following experiment confirms that validity of our theoretical results from Propositions 2 and 4. 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:

$$X = P_1(Z) + \varepsilon_x, \quad Y = P_1(Z) + \varepsilon_y. \quad (6)$$

Above, $\varepsilon_x$ and $\varepsilon_y$ follow two independent standard normal distributions, $Z \sim \mathcal{N}(0_d, \Sigma)$ with $\Sigma \in \mathbb{R}^{d_z \times d_z}$. The covariance matrix $\Sigma$ is obtained by multiplying product of 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 $X \perp Y | Z$. We also consider a modification of the above data generating function for which $H_1$ holds. This is done by adding a noise component $\varepsilon_b$ that is shared across $X$ and $Y$ as follows:

$$X = P_1(Z) + \varepsilon_x + \varepsilon_b, \quad Y = P_1(Z) + \varepsilon_y + \varepsilon_b, \quad (7)$$

where $\varepsilon_b$ follows the standard normal distribution. Since we consider Gaussian kernels, we can obtain an explicit formulation of $E_X [k_X(\mathbf{t}_j^{(1)}, \tilde{X}) | Z = \cdot]$ and $E_Y [k_Y(\mathbf{t}_j^{(2)}, Y) | Z = \cdot]$ for both data generation functions. See Appendix B.3 for more details. As a consequence, we are able to compute both the normalized version of our oracle statistic $\widetilde{C}_{n,p}$ as well as our approximate normalized statistic $\tilde{C}_{n,r,p}$. In Figure 2 we show that both statistics manage to recover the asymptotic null distribution under the null hypothesis, and reject the null hypothesis under $H_1$. In addition, we show that in the high dimensional setting, only our optimized version of $\tilde{C}_{n,r,p}$—obtained by optimizing the hyperparameters involved in the RLS estimators of our statistic—manages to recover the theoretical asymptotic null distribution under $H_0$.

**Comparisons with existing tests.** In our next experiments, we compare the performance of our method (implemented with the optimized version of our test
We now conduct another series of experiments that focus on the high dimensional setting, which lead to similar conclusions. In addition, we present in the Appendix B.3 similar experiments that focus on the high dimensional setting, which lead to similar conclusions.

We introduced a new kernel-based statistic for testing conditional independence. We derived its asymptotic distribution and showed that it outperforms other tests both in terms of KS and AUPC in most of the settings.
totic null distribution and designed a simple testing procedure that emerges from it. We also showed that our test is consistent and computationally efficient. Using various synthetic experiments, we demonstrated that our approach outperforms state-of-the-art methods both in terms of type-I error and power, even in the high dimensional setting.

References

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.

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 Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample test with analytic representations of probability measures. arXiv preprint arXiv:1506.04725, 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.

J.J. 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.

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

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.

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

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.

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

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.

Meyer Scetbon and Gael Varoquaux. Comparing distributions: $\ell_1$ 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.
An $\ell^p$-based Kernel Conditional Independence Test

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), Jun 2020. ISSN 0090-5364. doi: 10.1214/19-aos1857. URL [http://dx.doi.org/10.1214/19-AOS1857](http://dx.doi.org/10.1214/19-AOS1857).

Rajen D Shah, Jonas Peters, et al. The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, 48(3):1514–1538, 2020.

Chengchun Shi, Tianlin Xu, Wicher Bergsma, and Lexin Li. Double generative adversarial networks for conditional independence testing, 2020.

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.

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.
Supplementary Material

A  Proofs

A.1  Proof of Proposition 3

Proof. For all \( j \in [J] \):

\[
\sqrt{n \mu_{n,r}(t_j^{(1)}, t_j^{(2)})} = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \left( k_{\tilde{X}}(t_j^{(1)}, \tilde{x}_i) - h_{j,r}(z_i) \right) \left( k_{\tilde{Y}}(t_j^{(2)}, y_i) - h_{j,r}^2(z_i) \right) \\
= \sqrt{n \mu_{n,r}(t_j^{(1)}, t_j^{(2)})} \\
+ \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \left( k_{\tilde{X}}(t_j^{(1)}, \tilde{x}_i) - k_{\tilde{Y}}(t_j^{(1)}, \tilde{X}) | Z = z_i \right) \left( k_{\tilde{Y}}(t_j^{(2)}, Y) | Z = z_i - h_{j,r}'(z_i) \right) \\
+ \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{E}_X \left[ k_{\tilde{X}}(t_j^{(1)}, \tilde{X}) | Z = z_i \right] - h_{j,r}(z_i) \right) \left( k_{\tilde{Y}}(t_j^{(2)}, y_i) - \mathbb{E}_Y \left[ k_{\tilde{Y}}(t_j^{(2)}, Y) | Z = z_i \right] \right) \\
+ \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{E}_X \left[ k_{\tilde{X}}(t_j^{(1)}, \tilde{X}) | Z = z_i \right] - h_{j,r}(z_i) \right) \left( \mathbb{E}_Y \left[ k_{\tilde{Y}}(t_j^{(2)}, Y) | Z = z_i \right] - h_{j,r}^2(z_i) \right)
\]

(10)

(11)

(12)

(13)

Let us treat the four terms of this decomposition. The term (10) has been treated by Proposition 2 and satisfies, under the null hypothesis \( H_0 \):

\[
\sqrt{n \mu_{n,r}(t_j^{(1)}, t_j^{(2)})} \to_{n \to \infty} \mathcal{N} \left( 0, \mathbb{E} \left[ \left( k_{\tilde{X}}(t_j^{(1)}, \tilde{X}) - \mathbb{E}_X \left[ k_{\tilde{X}}(t_j^{(1)}, \tilde{X}) | Z = z_i \right] \right) \left( k_{\tilde{Y}}(t_j^{(2)}, Y) - \mathbb{E}_Y \left[ k_{\tilde{Y}}(t_j^{(2)}, Y) | Z = z_i \right] \right) \right] \right)
\]

Let us now show that the last term (13) converges towards 0 in probability. Let us denote for all \( j \), \( e_j^1 : z \to \mathbb{E}_X \left[ k_{\tilde{X}}(t_j^{(1)}, \tilde{X}) | Z = z \right] \) and \( e_j^2 : z \to \mathbb{E}_Y \left[ k_{\tilde{Y}}(t_j^{(2)}, Y) | Z = z \right] \), both elements of \( H_Z \) by Assumption 3. Then we have, for all \( i \in [n] \):

\[
(e_j^1(z_i) - h_{j,r}(z_i)) (e_j^2(z_i) - h_{j,r}^2(z_i)) = \left( (e_j^1 - h_{j,r}) \otimes (e_j^2 - h_{j,r}^2), k_Z(z_i, \cdot) \otimes k_Z(z_i, \cdot) \right)
\]

Then we deduce, by denoting: \( \mu_{ZZ} = \mathbb{E} [k_Z(Z, \cdot) k_Z(Z, \cdot)] \) and \( \mu_{ZZ} = \frac{1}{n} \sum_{i=1}^{n} k_Z(z_i, \cdot) k_Z(z_i, \cdot) \):

\[
\frac{1}{n} \sum_{i=1}^{n} \left( \mathbb{E}_X \left[ k_{\tilde{X}}(t_j^{(1)}, \tilde{X}) | Z = z_i \right] - h_{j,r}(z_i) \right) \left( \mathbb{E}_Y \left[ k_{\tilde{Y}}(t_j^{(2)}, Y) | Z = z_i \right] - h_{j,r}^2(z_i) \right)
\]

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

\[
= \langle e_j^1 - h_{j,r}^1 \otimes (e_j^2 - h_{j,r}^2), \mu_{ZZ} \rangle + \langle (e_j^1 - h_{j,r}^1) \otimes (e_j^2 - h_{j,r}^2), \mu_{ZZ} - \mu_{ZZ} \rangle
\]

We have:

\[
\| (e_j^1 - h_{j,r}^1) \otimes (e_j^2 - h_{j,r}^2), \mu_{ZZ} \| = \mathbb{E}_Z \left[ (e_j^1(Z) - h_{j,r}^1(Z))(e_j^2(Z) - h_{j,r}^2(Z)) \right] \leq \| e_j^1 - h_{j,r}^1 \|_{L^2(P_Z)} \| e_j^2 - h_{j,r}^2 \|_{L^2(P_Z)}
\]

Under the Assumptions 2, for \( \lambda_r = \frac{1}{\sqrt{n}_r} \), we have, using the results from 7: \( \| e_j^1 - h_{j,r}^1 \|_{L^2(P_Z)} \leq \frac{C_2^2}{\lambda_r \sqrt{n}_r} \) with probability \( 1 - 4e^{-7} \) and \( \| e_j^2 - h_{j,r}^2 \|_{L^2(P_Z)} \leq \frac{C_2^2}{\lambda_r \sqrt{n}_r} \) with probability \( 1 - 4e^{-7} \), for some constant \( C \) independent
from $n$ and $\tau$. Then by union bound, we deduce with probability $1 - 8e^{-\tau}$ we have:

$$\sqrt{n}\left(\left(\hat{e}_j - \hat{h}_{j,r}\right) \otimes \left(\hat{e}_j - \hat{h}_{j,r}^2\right), \mu_{ZZ}\right) \leq \sqrt{n}\frac{C^2r^4}{r^{\frac{3}{2}}\tau^4}$$

Then, if $\sqrt{n} \in o\left(n\frac{r^2}{\tau^2}\right)$, we have: $\sqrt{n}\left(\left(\hat{e}_j - \hat{h}_{j,r}\right) \otimes \left(\hat{e}_j - \hat{h}_{j,r}^2\right), \mu_{ZZ}\right) \to 0$ in probability when $n \to \infty$. Moreover:

$$\left(\left(\hat{e}_j - \hat{h}_{j,r}\right) \otimes \left(\hat{e}_j - \hat{h}_{j,r}^2\right), \mu_{ZZ} - \mu_{ZZ}\right) \leq \left\|\hat{e}_j - \hat{h}_{j,r}\right\|_{H_x} \left\|\hat{e}_j - \hat{h}_{j,r}^2\right\|_{H_x} \left\|\hat{\mu}_{ZZ} - \mu_{ZZ}\right\|_{H_x \otimes H_x}$$

By Markov inequality, $\left\|\hat{\mu}_{ZZ} - \mu_{ZZ}\right\|_{H_x \otimes H_x} \leq \sqrt{\frac{C}{n\tau}}$ with probability $1 - \delta$ for some constant $C'$. Moreover, under Assumption $\mathbb{A}$, we have $\left\|\hat{e}_j - \hat{h}_{j,r}\right\|_{H_x} \to 0$ and $\left\|\hat{e}_j - \hat{h}_{j,r}^2\right\|_{H_x} \to 0$ in probability. Then, we deduce that $\sqrt{n}\left(\left(\hat{e}_j - \hat{h}_{j,r}\right) \otimes \left(\hat{e}_j - \hat{h}_{j,r}^2\right), \mu_{ZZ} - \mu_{ZZ}\right) \to 0$ in probability. Finally, the term $\mathbb{B}$ goes to 0 in probability.

The terms $\mathbb{B}$ and $\mathbb{C}$ are similar and can be treated the same way. We only focus on the term $\mathbb{B}$. For all $i \in [n]$

$$\frac{1}{n} \sum_{i=1}^{n} \left(k_X(t_{1j}^{(1)}, \hat{X}_i) - \mathbb{E}_X \left(k_X(t_{1j}^{(1)}), \hat{X}\right)\right) \left(k_Y(t_{2j}^{(2)}, Y)\right) = \frac{1}{n} \sum_{i=1}^{n} \left(k_X(t_{1j}^{(1)}, \hat{X}_i) - \mathbb{E}_X \left(k_X(t_{1j}^{(1)}), \hat{X}\right)\right) \left(k_Y(t_{2j}^{(2)}, Y)\right)$$

where: $\hat{\mu}^{1}_{XX} = \frac{1}{n} \sum_{i=1}^{n} k_X(\hat{X}_i, \cdot) \otimes k_X(\cdot, \cdot)$, $\hat{\mu}^{2}_{XX} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_X \left(k_X(\hat{X}_i, \cdot)\right) \otimes k_X(\cdot, \cdot)$, and $\mu_{XX} = \mathbb{E}_X \left(\{\mu_X(t, \cdot)\}_{t \in [J]}\right)$.

By the law of large numbers, we have: $\hat{\mu}^{1}_{XX}$ and $\hat{\mu}^{2}_{XX}$ converge almost surely towards $\mu_{XX}$. Moreover by Markov inequality, $\left\|\hat{\mu}^{1}_{XX} - \mu_{XX}\right\|_{H_x \otimes H_x} \leq \sqrt{\frac{C}{n\tau}}$ with probability $1 - \delta$, and $\left\|\hat{\mu}^{2}_{XX} - \mu_{XX}\right\|_{H_x \otimes H_x} \leq \sqrt{\frac{C}{n\tau}}$ with probability $1 - \delta$. Then with probability $1 - 2\delta$, $\sqrt{n}\left(\left(\hat{\mu}^{1}_{XX} - \mu_{XX}\right)_{H_x \otimes H_x} + \left(\hat{\mu}^{2}_{XX} - \mu_{XX}\right)_{H_x \otimes H_x}\right) \leq 2\sqrt{\frac{C}{n\tau}}$. Moreover, under Assumption $\mathbb{A}$ using the results from $\mathbb{A}$, we have that $\left\|\hat{e}_j - \hat{h}_{j,r}\right\|_{H_x} \to 0$ in probability. Then the term $\mathbb{A}$ converges in probability towards 0. The same reasoning holds for $\mathbb{C}$.

Finally, by Slutzky’s Lemma:

$$\sqrt{n}\tilde{S}_{n,r}(t_{1j}^{(1)}, t_{2j}^{(2)}) \to_{n \to \infty} \mathcal{N}\left(0, 0, \mathbb{E}_X \left(k_X(t_{1j}^{(1)}, \hat{X}) - \mathbb{E}_X \left(k_X(t_{1j}^{(1)}), \hat{X}\right)\right) \left(k_Y(t_{2j}^{(2)}, Y) - \mathbb{E}_Y \left(k_Y(t_{2j}^{(2)}), \hat{Y}\right)\right)\right)$$

Now we have $\tilde{S}_{n,r} = \left(\hat{\mu}_{n,r}(t_{1j}^{(1)}, t_{2j}^{(2)})\right)_{j \in [J]} = \left(\hat{\mu}_n(t_{1j}^{(1)}, t_{2j}^{(2)})\right)_{j \in [J]} + \left(\hat{\mu}_{n,r}(t_{1j}^{(1)}, t_{2j}^{(2)}) - \hat{\mu}_n(t_{1j}^{(1)}, t_{2j}^{(2)})\right)_{j \in [J]}$. We have shown that $\sqrt{n}\left(\left(\hat{\mu}_{n,r}(t_{1j}^{(1)}, t_{2j}^{(2)}) - \hat{\mu}_n(t_{1j}^{(1)}, t_{2j}^{(2)})\right)_{j \in [J]}\right)$ goes to 0 in probability. Then by Slutzky Lemma and Proposition $\mathbb{A}$ we get: $\tilde{S}_{n,r} \to \mathcal{N}(0, \Sigma)$.

Let $r > 0$. Under $H_1$, $\tilde{S}_{n,r} \to \mathbb{S} \neq 0$. Let consider a realization of $(t_{1j}^{(1)}, t_{2j}^{(2)})_{j \in [J]}$ such that $\left\|\mathbb{S}\right\|_p \neq 0$. So $P(n^{p/2}\left\|\tilde{S}_{n,r}\right\|_p \geq r) \to 1$ as $n \to \infty$ because $\left\|\mathbb{S}\right\|_p \neq 0$. 

A.2 Proof of Proposition 4

Proof. First notice that:

\[
\tilde{\Sigma}_{n,r} = \frac{1}{n} \sum_{i=1}^{n} \tilde{u}_{i,r} \tilde{u}_{i,r}^T + \mu_n I_{d_f}
\]

\[
= \tilde{\Sigma}_n + \frac{1}{n} \sum_{i=1}^{n} (\tilde{u}_{i,r} - \tilde{u}_r) (\tilde{u}_{i,r} - \tilde{u}_r)^T + \frac{1}{n} \sum_{i=1}^{n} (\tilde{u}_{i,r} - \tilde{u}_r) (\tilde{u}_{i,r} - \tilde{u}_r)^T + \mu_n I_{d_f}
\]

By the law of large numbers, we get that under \( H_0 \): \( \tilde{\Sigma}_n \to \Sigma \). Moreover:

\[
\left[ \frac{1}{n} \sum_{i=1}^{n} (\tilde{u}_{i,r} - \tilde{u}_r)^T \right]_{kl} = \frac{1}{n} \sum_{i=1}^{n} \left( k_y(t_k^{(2)}, y_i) - E_Y \left[ k_y(t_k^{(2)}, Y) | Z = z_i \right] \right) \left( E_X \left[ k_X(t_k^{(1)}, X) | Z = z_i \right] - h_k^{(1)}_{H_{k,r,t_k^{(1)}}} (z_i) \right)
\]

which has been proven to converge in probability to 0 in the proof of Proposition 3. Then \( \frac{1}{n} \sum_{i=1}^{n} (\tilde{u}_{i,r} - \tilde{u}_r)^T \) converges in probability to 0. Similarly \( \frac{1}{n} \sum_{i=1}^{n} (\tilde{u}_{i,r} - \tilde{u}_r)^T \) and \( \frac{1}{n} \sum_{i=1}^{n} (\tilde{u}_{i,r} - \tilde{u}_r) (\tilde{u}_{i,r} - \tilde{u}_r)^T \) also converge in probability to 0. Then by Slutsky Lemma, \( \tilde{\Sigma}_{n,r} \) converges in probability to \( \Sigma \). By Slutsky’s lemma (again) and by Proposition 3, we have that: \( \Sigma_{n,r} \tilde{\Sigma}_{n,r} \) converges to a standard gaussian distribution \( N(0, I_{d_f}) \). The second part of the proposition is the same than the proof of Proposition 3.

B Additional Experiments

B.1 A note 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 \( \chi^2 \) distribution.

B.2 A first toy problem

Figure 4 compares the KS statistic and the AUPC obtained by different testing procedures in the exact same setting considered in Figure 2. Following that figure, one can see that our method outperforms the other tests both in terms of type-I error and power. In addition, Figure 5 presents the results obtained in a high dimensional regime, showing that our method tends to outperform the existing tests in most cases.

![Figure 5: 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). 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. These experiments show that our method outperforms the other tests both in term of KS and AUPC in most of the settings.](image)

B.3 A second toy problem

In this section, we provide an additional comparison of the KS statistic and the AUPC obtained by the different tests when the data is generated from the models defined in Eq. (8) and Eq. (9), respectively, focusing on a high
Figure 6: 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). Each point in the figures is obtained by repeating the experiment 100 independent trials. The panels present the KS (left) and AUPC (right) obtained by the tests when varying the dimension $d_z$ from 10 to 50 for a fixed number of samples $n = 1000$. These experiments show that our method outperforms other tests both in term of KS and AUPC in the high dimensional setting.

Figure 7: 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). Each point in the figures is obtained by repeating the experiment for 100 independent trials. The panels present the KS (left) and AUPC (right) obtained by the tests when varying the dimension $d_z$ from 10 to 50 for a fixed number of samples $n = 1000$. These experiments show that our method outperforms other tests both in terms of KS and AUPC in the high dimensional setting.