Journal of Machine Learning Research 25 (2024) 1-40 Submitted 7/23; Revised 1/24; Published 1/24

A General Framework for the Analysis of Kernel-based Tests

Tamara Fernández TAMARA.FERNANDEZ@UAI.CL
Faculty of Engineering and Science
Universidad Adolfo Ibáñez, Chile

Nicolás Rivera NICOLAS.RIVERA@UV.CL
Facultad de Ciencias
Universidad de Valparaíso, Chile

Editor: Isabelle Guyon

Abstract

Kernel-based tests provide a simple yet effective framework that uses the theory of reproducing kernel Hilbert spaces to design non-parametric testing procedures. In this paper, we propose new theoretical tools that can be used to study the asymptotic behaviour of kernel-based tests in various data scenarios and in different testing problems. Unlike current approaches, our methods avoid working with U and V-statistics expansions that usually lead to lengthy and tedious computations and asymptotic approximations. Instead, we work directly with random functionals on the Hilbert space to analyse kernel-based tests. By harnessing the use of random functionals, our framework leads to much cleaner analyses, involving less tedious computations. Additionally, it offers the advantage of accommodating pre-existing knowledge regarding test-statistics as many of the random functionals considered in applications are known statistics that have been studied comprehensively. To demonstrate the efficacy of our approach, we thoroughly examine two categories of kernel tests, along with three specific examples of kernel tests, including a novel kernel test for conditional independence testing.

Keywords: kernel methods, hypothesis testing, reproducing kernel Hilbert space

1. Introduction

The aim of this paper is to introduce new general tools for the analysis of kernel-based methods in the context of hypothesis testing. For this purpose, consider the following general framework. Let $X_1, \ldots, X_n$ be a collection of data points and consider a null hypothesis of interest, which we denote by H_0. Suppose that to assess the validity of H_0 we have access to a test-statistic $S_n(\omega)$ that depends (implicitly) on our data and on a deterministic real-valued weight function $\omega : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is a space related to the observed data points. Furthermore, assume that $S_n(\omega)$ satisfies $S_n(\omega) \approx 0$ for any fixed ω under the null hypothesis. A testing procedure based on $S_n(\omega)$ works as follows. Choose a function ω and compute $S_n(\omega)$. If we observe $S_n(\omega) \approx 0$ then we do not reject the null hypothesis, but if we observe that $S_n(\omega)$ is far away from zero, then we use this as evidence to reject the null hypothesis. Of course, we might still have $S_n(\omega) \approx 0$ under the alternative hypothesis for some functions ω, and in such a case the test-statistic $S_n(\omega)$ will perform poorly.

In the previous setting, we need to carefully choose the weight function ω in order to have a robust test that is able to distinguish between null and alternative hypotheses.

Arguably, there are two approaches that can be followed to tackle the problem of choosing an appropriate weight function: we can learn the weight function ω from our data, i.e., an adaptive weight approach, or we can combine several weight functions into a single test-statistic, which is the path that motivates this work.

Following the idea of combining several weight functions into a single test, a simple proposal is to consider as test-statistic the random variable

$$\Psi_n = \sup_{\omega \in \mathcal{H} : \|\omega\|_{\mathcal{H}} = 1} S_n(\omega)^2, \tag{1}$$

where $\mathcal{H}$ is a separable reproducing kernel Hilbert space (RKHS). While in principle we could have taken supremum over an arbitrary space of functions, we choose the unit ball of an RKHSs since the favourable structural properties of RKHSs, coupled with certain regularity conditions imposed over the general test-statistic $S_n(\omega)$, will guarantee good properties of Ψ_n, which will be fundamental to construct rejection regions and thus the testing procedure. In particular, we will assume that $S_n(\omega)$ is linear in its argument ω, that is, $S_n(a\omega_1 + b\omega_2) = aS_n(\omega_1) + bS_n(\omega_2)$ for any $a, b \in \mathbb{R}$ and $\omega_1, \omega_2 \in \mathcal{H}$. Then it will follow that Ψ_n can be evaluated exactly via a closed-form expression, i.e., the optimisation problem can be solved explicitly, although the resulting expression may not be particularly pleasant. In the previous case, we say that $S_n(\omega)$ is a linear test-statistic, and we refer to Ψ_n to as the *kernelisation* of S_n. This simple idea is the basis for what is known as a kernel-based test (Gretton et al., 2006), and has been applied implicitly and explicitly in many contexts. In the literature, testing procedures based on test-statistics of the form of Eq. (1) are generally called kernel test-statistics, and the whole testing procedure derived from it is referred to as a kernel test.

Standard examples in the literature of kernelised tests, as defined in Eq. (1), include those based on the maximum mean discrepancy (**MMD**) introduced in the seminal work of Gretton et al. (2006). These tests are commonly used for the two-sample problem, which involves testing the null hypothesis $H_0 : F_0 = F_1$ using independent samples from F_0 and F_1. Since the development of the **MMD** test, extensive research has been conducted in the field of kernel-based tests. In the context of Goodness-of-Fit, a commonly employed approach is to kernelise a Stein's operator, resulting in tests known as Kernel Stein Discrepancy (KSD) tests. These tests have been proposed for various data domains such as $\mathbb{R}^d$ (Chwialkowski et al., 2016; Liu et al., 2016), point processes (Yang et al., 2019), and random graph models (Xu and Reinert, 2021), among others. Kernel methods have also been applied to address the problem of independence testing, where the main objective has been the analysis of the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005, 2007; Smola et al., 2007). Furthermore, kernel methods have found applications in other testing problems, including conditional independence (Zhang et al., 2011; Doran et al., 2014), composite goodness-of-fit (Key et al., 2021), and various testing problems in survival analysis (Fernández et al., 2020; Fernández and Rivera, 2021; Fernández et al., 2023; Ditzhaus et al., 2022). One of the advantages of kernel methods is their versatility in handling different types of data structures, such as graphs, strings, and sets. This enables the design of testing procedures for a wide range of data types. For an introductory review of kernel-based tests, we refer the reader to Chen and Markatou (2020). For a comprehensive overview of kernel methods and their applications in Statistics and Machine Learning, we recommend Hofmann et al. (2008) and Muandet et al. (2017).

Despite the vast literature on kernel-based methods, to the best of our knowledge, there are no works aiming to find a unified framework to analyse kernel-based tests. Thus, most of the existing tests and results are derived and analysed using first principles on a case-by-case basis, even though many similarities are present in the analyses. These similarities include: i) Most previous works focused on a random variable of the form Ψ_n, which arises implicitly or explicitly from a linear test statistic $S_n(\omega)$. ii) Most works base their analysis on expressing Ψ_n explicitly in the form of a V-statistic or a related object (e.g. a V-statistic with random kernel or dependent data). iii) The limit distribution of Ψ_n can be expressed as an (infinite) linear combination of independent chi-squared random variables, and is usually derived from limit theorems for V-statistics. iv) In most works the same resample schemes are used, with the wild bootstrap being a highly prevalent choice.

For certain specific but simple scenarios, kernel-based tests can be expressed as standard U and V-statistics, allowing for their analysis within this framework (see Section 4.1). However, this is not always the case and additional difficulties arise. In the latter scenario, the analyses of kernel-based tests require extra steps that often involve finding asymptotic approximations of the kernel test as a proper U or V-statistic (without random kernels or other type of complications), so the standard theory of V-statistics applies (for example, see Fernández et al. (2023) for an analysis following this path). Due to such approximations, the analysis becomes quite tedious and overly complicated; moreover, it is not guaranteed that such an approximation can be achieved in every case. The same can be said about the bootstrap versions of Ψ_n. In contrast, many of these computations are much simpler when we work directly with $S_n(\omega)$, and, furthermore, asymptotic results for $S_n(\omega)$ may already have been proven in the literature.

The goals of this work are to provide new ways to study kernel tests. We expect to i) avoid lengthy computations that usually arise by expressing the kernelised test statistic Ψ_n as an object that resembles a U or V-statistic and ii) be able to use already known results for $S_n(\omega)$ in the analysis of Ψ_n (which are usually much easier to obtain).

Our main idea to achieve our goals is to completely avoid writing Ψ_n as a U or V-statistic, and to work directly with S_n as a random functional on the Hilbert space $\mathcal{H}$, looking for conditions that allow us to extrapolate limiting results of $S_n(\omega)$, for fixed $\omega \in \mathcal{H}$, to Ψ_n. Working with random functionals is much simpler, and indeed our analysis is based on first principles of Hilbert space-valued random variables. At a high level we show that

i) Under the null hypothesis, some regularity conditions, and appropriate scaling it holds that if for all $\omega \in \mathcal{H}$, $S_n(\omega) \overset{\mathcal{D}}{\to} N(0, \sigma_\omega^2)$, then

$$\Psi_n \overset{\mathcal{D}}{\to} \sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

when the number of data points n tends to infinity. The variables Z_i are i.i.d. standard normal random variables and $\lambda_1, \lambda_2, \ldots$ are non-negative constants. More details are given in Theorems 1 and 2.

ii) Under the alternative hypothesis, some regularity conditions, and appropriate scaling it holds that if for all $\omega \in \mathcal{H}$, $S_n(\omega) \to c(\omega)$ almost surely, then almost surely we have that

$$\Psi_n \to c_*^2 := \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}=1} c(\omega)^2.$$

More details are given in Theorem 3.

We show that our conditions are not only sufficient but also necessary. We also provide additional sufficient conditions expressed in terms of integral conditions (see Propositions 4 and 6), which are not only more practical but also easier to verify. Consequently, these conditions might be of greater relevance to practitioners. We will also see that our conditions are useful not only to analyse test statistics but also bootstrap procedures such as wild bootstrap, allowing us to analyse the whole testing procedure.

To illustrate how to use our results, we analyse two general classes of tests: the first class considers statistics $S_n(\omega)$ that can be written as a sum of i.i.d. random variables, and the second one considers $S_n(\omega)$ as a U-statistic of degree 2 or more (i.e., we take supremum over U-statistics). These classes are rather important, as they include statistics such as the kernelised Stein discrepancy and the HSIC measure of independence, respectively. Additionally, due to their broad generality, these classes can also serve as readily applicable (or condensed) results for practitioners who need to kernelise a test-statistic falling within these categories. In these classes, we will see that our approach reduces the problem of showing the asymptotic correctness of the kernelised testing procedure to the verification of a few simple conditions (see Theorems 7 and 8), showing the effectiveness of our approach in practical problems.

Finally, we apply our ideas to specific testing problems. In particular, we analyse kernel tests for the problems of independence testing, the two-sample problem with right-censored data, and conditional independence testing. The last application is a novel test for conditional independence testing that kernelises the very recently proposed Weighted Generalised Covariance Measure (Scheidegger et al., 2022) which is a weighted generalisation of the Generalised Covariance Measure (Shah and Peters, 2020).

Notation. For the remainder of the paper, we adopt standard notation. In particular, we write $\overset{a.s}{\to}$, $\overset{\mathbb{P}}{\to}$, and $\overset{\mathcal{D}}{\to}$ to denote convergence almost surely, in probability, and in distribution (in law), respectively. All limits are taken when n, (usually the number of data points), tends to infinity, unless explicitly stated otherwise. We denote by $\mathbb{P}_X$ and $\mathbb{E}_X$ conditional probability and expectation on the random variable, event, or sigma-algebra X. For a positive integer k, we denote by $[k]$ the set $\{1, \ldots, k\}$.

2. Background

2.1 Hilbert Spaces and Random Linear Functionals

In this work, we are interested in Hilbert spaces of functions. The letters $\mathcal{H}$ and $\mathcal{G}$ usually denote spaces of functions from $\mathcal{X}$ to $\mathbb{R}$ and from $\mathcal{Y}$ to $\mathbb{R}$, respectively, with the respective inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{G}}$. We always assume that $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces (e.g. $\mathbb{R}^d$), and that Hilbert spaces are always separable.

In later applications to Hypothesis Testing, we will further consider Reproducing Kernel Hilbert Spaces (RKHS). We say that $\mathcal{H}$ is an RKHS if the evaluation functional E_x :

$\omega \to \omega(x) \in \mathbb{R}$ is bounded for each $x \in \mathcal{X}$. According to Riesz's representation theorem, there exists a unique $K_x \in \mathcal{H}$ such that $E_x\omega = \langle K_x, \omega\rangle_{\mathcal{H}}$. For $x, y \in \mathcal{X}$, we denote by $K(x, y) = \langle K_x, K_y\rangle_{\mathcal{H}}$ the so-called reproducing kernel of $\mathcal{H}$, which is a symmetric and positive definite function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The kernel function K characterises $\mathcal{H}$, and, in fact, given a symmetric positive definite function K there is a unique RKHS with such a function as reproducing kernel. In practice, we do not choose $\mathcal{H}$, but rather the kernel K. Standard kernel functions are the squared-exponential, the Ornstein–Uhlenbeck, and the rational quadratic kernels; see Chapter 2 in the work of Duvenaud (2014) for a compendium of kernel functions. Finally, since we are interested only in separable Hilbert spaces, it is worth mentioning that if $\mathcal{X}$ is separable and the kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous, then the RKHS $\mathcal{H}$ is separable.

We are interested in random variables taking values in $\mathcal{H}$ or in the space of bounded linear functionals $\mathcal{H}^*$ (with the standard operator norm $\|\cdot\|_{\mathcal{H}\to\mathbb{R}}$). We assume an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider random variables $\Omega \to \mathcal{H}$ or $\Omega \to \mathcal{H}^*$ that are measurable with respect to the corresponding Borel sigma-algebra. Other standard forms of measurability, such as the cylindrical sigma-algebra, are equivalent in the setting of separable Hilbert spaces. Due to the isometry between $\mathcal{H}$ and $\mathcal{H}^*$, we can define a random variable in one space and transfer its representation to the dual space without concerns about measurability issues. In particular, given a random bounded functional S, the random variable $\|S\|_{\mathcal{H}\to\mathbb{R}} = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S(\omega)$ is measurable. Similarly, for a random variable $\xi \in \mathcal{H}$, its norm $\|\xi\|_{\mathcal{H}}$ is also measurable.

Considering the aforementioned notation, our main interest is to study sequences $(S_n)_{n\in\mathbb{N}}$ of random functionals in $\mathcal{H}^*$. In particular, we are interested in the random variable

$$\Psi_n := \|S_n\|^2_{\mathcal{H}\to\mathbb{R}} = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n(\omega)^2. \tag{2}$$

In order to maintain the statistical motivation in our presentation, from now on we refer to random variables in $\mathcal{H}^*$ as bounded linear test-statistics.

2.2 Stable Convergence

In some settings, such as bootstrap or other resampling schemes, we want to extend the definition of convergence in distribution to allow convergence conditional on a sub-algebra, e.g. to obtain a limit distribution given the observed data. This idea is formalised by the concept of stable convergence (Häusler and Luschgy, 2015, Definition 3.15).

Consider a sub sigma-algebra $\mathcal{F}' \subseteq \mathcal{F}$. We say that the sequence of real random variables $(Z_n)_{n\geq 1}$ converges $\mathcal{F}'$-stably to Z, denoted $Z_n \overset{\mathcal{D}_{\mathcal{F}'}}{\to} Z$, if and only if

$$\mathbb{E}(X\mathbb{E}(f(Z_n)|\mathcal{F}')) \to \mathbb{E}(X\mathbb{E}(f(Z)|\mathcal{F}'))$$

for all $X \in L_1(\Omega, \mathbb{P}, \mathcal{F})$ and $f : \mathbb{R} \to \mathbb{R}$ continuous and bounded. Note that by taking $X = 1$, stable convergence implies the usual convergence in distribution. In this work, our limiting random variables will be independent of $\mathcal{F}'$, thus $\mathbb{E}(f(Z)|\mathcal{F}') = f(Z)$.

An equivalent definition is that $Z_n \overset{\mathcal{D}_{\mathcal{F}'}}{\to} Z$, if and only if Z_n converges to Z with respect to the conditional measure $\mathbb{P}(\cdot|F)$ for every $F \subseteq \mathcal{F}'$ with $\mathbb{P}(F) > 0$, i.e., $\mathbb{E}_F(f(Z_n)) \to \mathbb{E}_F(f(Z))$, for all bounded and continuous f.

In practice, most of the usual limit theorems, such as Linderberg's CLT, hold in the setting of stable convergence, but all conditions involving probabilities/expectation need to be replaced by the corresponding conditional probability/expectation on $\mathcal{F}$ (e.g. Theorem 6.1 of Häusler and Luschgy (2015)).

In data-driven problems involving bootstrap and resampling methods, we denote all data by D, and write $\mathcal{D}_D$ as a shortcut for $\mathcal{D}_{\sigma(D)}$ to represent $\sigma(D)$-stable convergence, where $\sigma(D)$ is the sigma-algebra represented by the data D.

3. Convergence of Kernelised Linear Test-statistics

As discussed in Section 1, in practical scenarios of hypothesis testing, $S_n(\omega)$ represents a test statistic that depends on the weight ω. Randomness is typically introduced by the observed data, which consists of n data points denoted as $X_1, \dots, X_n$. The subscript n corresponds to the number of data points. There are two cases of particular interest in this context. The first case occurs when $S_n(\omega)$ is properly scaled and converges to a normal distribution with a mean of 0 and a variance of $\sigma(\omega, \omega)$ for each function ω. This scenario is typically encountered under the null hypothesis. The other case is when $S_n(\omega)$ converges (assuming proper scaling) to a constant $c(\omega)$ that depends on ω (and hopefully $c(\omega) \neq 0$), which usually holds under the alternative hypothesis.

Let us start by analysing the first case, as it is the most interesting. In this case, we assume that the sequence of linear statistics $(S_n)_{n\geq 1}$ satisfies the following condition, referred to as Condition G_0, where G stands for *Gaussian* as in the Gaussian distribution.

Condition G_0 *For each $\omega \in \mathcal{H}$ we have $S_n(\omega) \xrightarrow{\mathcal{D}} S(\omega)$ with $S(\omega) \sim \mathcal{N}(0, \sigma(\omega, \omega))$ as n grows to infinity.*

Condition G_0 is a rather natural and common behaviour for test-statistics under the null hypothesis, so it can be seen as a framework rather than just a condition. Note that according to Condition G_0, $\sigma(\omega, \omega)$ is defined only on the diagonal. However, it can be extended to a bilinear operator by defining $\sigma(\omega, \omega') := \frac{1}{2}\left(\sigma(\omega + \omega', \omega + \omega') - \sigma(\omega, \omega) - \sigma(\omega', \omega')\right)$. The bilinear operator $\sigma : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ plays a fundamental role in our work, as it characterises the joint convergence of $S(\omega_1), \dots, S(\omega_m)$ with $\omega_1, \dots, \omega_m \in \mathcal{H}$ to a multivariate normal with mean 0 and covariance matrix $\Sigma_{ij} = \sigma(\omega_i, \omega_j)$ (see Lemma 13). Moreover, σ also characterises the potential limit (in law) of S_n. If we assume that σ is bounded, i.e., $\sup_{\omega \in \mathcal{H} : \|\omega\|_{\mathcal{H}} = 1} \sigma(\omega, \omega) < \infty$, then the Riesz representation theorem guarantees the existence of a unique linear transformation $T_\sigma : \mathcal{H} \to \mathcal{H}$ such that $\langle \omega, T_\sigma \omega' \rangle_{\mathcal{H}} = \sigma(\omega, \omega')$ (see Proposition 14 in Appendix A.1). Furthermore, due to the symmetry of σ, T_σ is self-adjoint.

In the specific case where $\mathcal{H}$ is an RKHS with kernel K, then

$$(T_\sigma \omega)(x) = \sigma(\omega, K_x), \quad \forall x \in \mathcal{X}. \tag{3}$$

Recall that $K_x \in \mathcal{H}$ is the (unique) element associated with the evaluation functional E_x via the Riesz representation theorem. Assuming Condition G_1 below we show in Proposition 14 that T_σ is self-adjoint and trace class.

Condition G_1 *For some orthonormal basis $(\phi_i)_{i\geq 1}$ of $\mathcal{H}$ we have $\sum_{i\geq 1} \sigma(\phi_i, \phi_i) < \infty$.*

It is well known that if the condition above holds for one orthonormal basis, then it holds for every orthonormal basis (see, e.g., Corollary 18.2 of Conway (2000)).

By Condition G_1 and since T_σ is self-adjoint, there exists an orthonormal basis of $\mathcal{H}$ denoted as $(\phi_i)_{i\geq 1}$, satisfying $T_\sigma \phi_i = \lambda_i \phi_i$ for each $i \geq 1$. Additionally, the sum of the eigenvalues λ_i is bounded, and these eigenvalues are real since T_σ is self-adjoint and non-negative since σ is a variance (except for the potential eigenvalues 0). This will play a crucial role in defining the (potential) limit of S_n.

Finally, to ensure that S_n actually converges in distribution to a limiting functional, we require the following tightness condition:

Condition G_2 *For some orthonormal basis $(\phi_i)_{i\geq 1}$ of $\mathcal{H}$, and for any $\varepsilon > 0$, we have that*

$$\lim_{i\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sum_{j=i}^{\infty} S_n(\phi_j)^2 \geq \varepsilon \right) = 0.$$

Similarly to Condition G_1, we can show that if Condition G_2 holds for one basis, then it holds for any basis of $\mathcal{H}$.

We continue by presenting the main theorems that allow us to study kernel statistics through the random functional S_n. The first result allows us to derive a limit distribution for Ψ_n based on the fact that $S_n(\omega)$ converges in distribution for every $\omega \in \mathcal{H}$.

Theorem 1 *Let $(S_n)_{n\geq 1}$ be a sequence of bounded linear test-statistics satisfying Conditions G_0 to G_2. Define the random functional*

$$S(\cdot) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\, \langle \phi_i, \cdot \rangle_{\mathcal{H}}\, Z_i,$$

where $(\lambda_i, \phi_i)_{i\geq 1}$ are the eigenvalues and eigenvectors of the operator $T_\sigma : \mathcal{H} \to \mathcal{H}$ defined in Eq. (3), and $(Z_i)_{i\geq 1}$ are a collection of i.i.d. standard normal random variables.

Then S exists almost surely, i.e., the sum converges almost surely in $\mathcal{H}^*$, and

$$S_n \overset{\mathcal{D}}{\to} S, \qquad \text{and} \qquad \Psi_n = \|S_n\|^2_{\mathcal{H}\to\mathbb{R}} \overset{\mathcal{D}}{\to} \sum_{i=1}^{\infty} \lambda_i Z_i^2. \tag{4}$$

Random variables of the form $\sum_{i=1}^{\infty} \lambda_i Z_i^2$, where λ_i are the eigenvalues of an operator T_σ, frequently appear in our applications. Therefore, we introduce the notation $\chi^2(\sigma)$ to denote the distribution of the aforementioned series. We write

$$\Psi \sim \chi^2(\sigma) \tag{5}$$

to indicate that the random variable Ψ has the same distribution as $\sum_{i=1}^{\infty} \lambda_i Z_i^2$ associated with the bilinear form σ through the operator T_σ.

We also show that the conditions imposed in Theorem 1 are necessary. This result is stated in the following theorem.

Theorem 2 *Let $(S_n)_{n\geq 1}$ be a sequence of bounded linear test-statistics in $\mathcal{H}^*$, and define $S(\cdot) = \sum_{i=1}^{\infty}\sqrt{\lambda_i}Z_i \langle \phi_i, \cdot\rangle_{\mathcal{H}}$, where $(\lambda_i)_{i\geq 0}$ are positive constants, and $(\phi_i)_{i\geq 1}$ is an orthonormal basis of $\mathcal{H}$. Suppose that S converges almost surely in $\mathcal{H}^*$ (i.e., almost surely the truncated sums are a Cauchy sequence in $\mathcal{H}^*$). Then, if $S_n \overset{\mathcal{D}}{\to} S$, we have that Conditions G_0 to G_2 hold.*

Now, we present our second main tool to analyse kernel statistics, which pertains to the analysis of the second case of interest. This case arises when $S_n(\omega)$ converges almost surely or in probability to a constant value $c(\omega)$ for every $\omega \in \mathcal{H}$. Such a situation is often observed under the alternative hypothesis and can be described relatively straightforwardly. The following condition, which is essentially the almost surely version of Condition G_2, features our analysis.

Condition G_3 *For some orthonormal basis $(\phi_i)_{i\geq 1}$ of $\mathcal{H}$, it holds that*

$$\lim_{i\to\infty}\limsup_{n\to\infty}\sum_{j=i}^{\infty} S_n(\phi_j)^2 = 0, \quad a.s.$$

Theorem 3 *Consider a sequence $(S_n)_{n\geq 1}$ of linear test-statistics, and suppose that for each $\omega \in \mathcal{H}$ we have $S_n(\omega) \overset{a.s.}{\to} c(\omega)$, where $c : \mathcal{H} \to \mathbb{R}$ is a deterministic functional. Define $c_*^2 := \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} c(\omega)^2$, and suppose $c_* < \infty$. Then*

$$\lim_{n\to\infty} \Psi_n = c_*^2, \quad a.s.$$

if and only if Condition G_3 holds.

Furthermore, if for every $\omega \in \mathcal{H}$ we have $S_n(\omega) \overset{\mathbb{P}}{\to} c(\omega)$ as n grows to infinity, then $\Psi_n \overset{\mathbb{P}}{\to} c_*^2$ if and only if Condition G_2 holds.

The proofs of Theorems 1 to 3 are deferred to Appendix B.

3.1 Integrability Conditions

Verifying the conditions stated in Conditions G_1 and G_2 can be challenging in practical examples due to their algebraic nature. To address this, in this section, we introduce integrability conditions that imply Conditions G_1 and G_2 and that may be easier to verify in practical problems.

In the upcoming discussion, we consider Polish spaces $\mathcal{X}$ and $\mathcal{Y}$, along with the (separable) Hilbert Space $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ and a Hilbert Space $\mathcal{G}$ of functions from $\mathcal{Y}$ to $\mathbb{R}$, as well as the Borel measures μ and ν over $\mathcal{X}$ and $\mathcal{Y}$, respectively. Additionally, we assume that $\mathcal{H}$ and $\mathcal{G}$ are subspaces of $L_2(\mathcal{X},\mu)$ and $L_2(\mathcal{Y},\nu)$, respectively (considering the equivalence class of functions). It is worth noting that if $\mathcal{H}$ is a separable RKHS with kernel K, a sufficient condition for $\mathcal{H}$ to be a subset of $L_2(\mathcal{X},\mu)$ is that $\int_{\mathcal{X}} K(x,x)\mu(dx) < \infty$ (the analogous condition holds for $\mathcal{G}$). This condition holds because $\int_{\mathcal{X}} f(x)^2\mu(dx) \leq \|f\|_{\mathcal{H}}^2 \int_{\mathcal{X}} K(x,x)\mu(dx)$ by the Cauchy-Schwarz inequality.

The integrability conditions we are going to introduce are based on the existence of a linear transformation $Q : \mathcal{H} \to \mathcal{G} \subseteq L_2(\mathcal{Y},\nu)$ such that $\mathbb{E}(S_n(\omega)^2) \leq C\|Q\omega\|_{L_2(\nu)}^2$ for all sufficiently large n and for all $\omega \in \mathcal{H}$. By introducing Q and integrating with respect to ν, we can express Conditions G_1 and G_2 in a more natural form for practical applications.

Note that if $\mathcal{G}$ is an RKHS associated with the kernel L, we have $(Q\omega)(y) = \langle Q\omega, L_y\rangle_{\mathcal{G}} = \langle \omega, Q^* L_y\rangle_{\mathcal{H}}$ for any $y \in \mathcal{Y}$, where Q^* is the adjoint operator of Q. Then, for any $y, y' \in \mathcal{Y}$, we define the kernel $L^Q : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ as follows:

$$L^Q(y, y') = \langle Q^* L_y, Q^* L_{y'}\rangle_{\mathcal{H}}, \tag{6}$$

which is clearly symmetric and definite positive. Note that L^Q can be understood as the kernel $Q_1^* Q_2^* L$, where Q_i^* represents the application of Q^* to the i-th component of L. The kernel L^Q plays a significant role in various applications; however, its shape may not be very pleasant in some settings.

Proposition 4 *Suppose that there exists a linear transformation $Q : \mathcal{H} \to \mathcal{G} \subseteq L_2(\mathcal{Y}, \nu)$ and a constant $C > 0$, satisfying:*

$$\mathbb{E}(S_n(\omega)^2) \leq C \int_{\mathcal{Y}} (Q\omega)(y)^2 \nu(dy) \tag{7}$$

for all $\omega \in \mathcal{H}$ and for sufficiently large n. Moreover, consider the following conditions:

a) $\mathcal{H}$ is a RKHS with kernel K and it exists a measure μ on $\mathcal{X}$ such that

$$\left(\sup_{\omega \in \mathcal{H} : \|\omega\|_{L_2(\mu)} = 1} \int_{\mathcal{Y}} (Q\omega)(y)^2 \nu(dy)\right) \left(\int_{\mathcal{X}} K(x, x)\mu(dx)\right) < \infty, \tag{8}$$

b) $\mathcal{G}$ is a RKHS with kernel L, and

$$\int_{\mathcal{Y}} L^Q(y, y)\nu(dy) < \infty. \tag{9}$$

If a) or b) holds, then Condition G_2 holds. Furthermore, if Condition G_0 holds and a) or b) holds, then Condition G_1 holds.

Remark 5 *If both $\mathcal{H}$ and $\mathcal{G}$ are RKHSs, then Eq.* (8) *implies Eq.* (9)*.*

In a setting that involves bootstrap estimator or almost sure convergence of Ψ_n, we need to be able to use our convergence results to establish $\mathcal{F}'$-stable convergence for some appropriate sub-sigma $\mathcal{F}' \subseteq \mathcal{F}$ (recall the definition of stable convergence in Section 2.2). In this setting, we say that Condition G_0 holds $\mathcal{F}'$-stably if and only if for every $\omega \in \mathcal{H}$ there exists a random variable $S(\omega) \sim N(0, \sigma(\omega, \omega))$, independent of $\mathcal{F}'$, such that $S_n(\omega) \overset{\mathcal{D}_{\mathcal{F}'}}{\to} S(\omega)$. Also, we say that Condition G_2 holds $\mathcal{F}'$-stably if it holds when replacing the probability $\mathbb{P}$ by the conditional probability on $\mathbb{P}_F$ for all $F \subseteq \mathcal{F}'$ with $\mathbb{P}(F) > 0$. Note that other conditions do not involve probabilities. Therefore, Theorem 1 shows stable convergence if Conditions G_0 to G_2 hold $\mathcal{F}'$-stably (the same applies to Theorem 3). Within this context, the following extension of Proposition 4 helps us to prove $\mathcal{F}'$-stable convergence:

Proposition 6 *Let $\mathcal{F}'$ be a sub sigma-algebra of $\mathcal{F}$. Suppose there exists a linear operator $Q : \mathcal{H} \to \mathcal{G} \subseteq L_2(\mathcal{Y}, \nu)$, a constant $C > 0$, and a (potentially random) sequence of sigma-finite measures ν_n on $\mathcal{Y}$, such that for every $n \geq 1$, it holds that*

$$\mathbb{E}(S_n(\omega)^2|\mathcal{F}') \leq C \int_{\mathcal{Y}} (Q\omega)^2(y)\nu_n(dy), \text{ a.s.} \tag{10}$$

Moreover, suppose that for each $g \in L_1(\mathcal{Y}, \nu)$ we have $\int_{\mathcal{Y}} g(y)\nu_n(dy) \to \int_{\mathcal{Y}} g(y)\nu(dy)$ almost surely.

Consider the following conditions:

a) $\mathcal{H}$ is a RKHS, and it exists a measure μ on $\mathcal{X}$ such that

$$\left(\sup_{\omega \in \mathcal{H}: \|\omega\|_{L_2(\mu)}=1} \int_{\mathcal{Y}} (Q\omega)(y)^2 \nu(dy)\right) \left(\int_{\mathcal{X}} K(x,x)\mu(dx)\right) < \infty \tag{11}$$

b) $\mathcal{G}$ is a RKHS with kernel L, and $\int_{\mathcal{Y}} L^Q(y,y)\nu(dy) < \infty$,

If a) or b) is satisfied, then Condition G_2 holds $\mathcal{F}'$-stably. Additionally, if Condition G_0 holds $\mathcal{F}'$-stably and a) or b) is satisfied, then Condition G_1 holds $\mathcal{F}'$-stably. Finally, in the special case that $\mathcal{F}' = \mathcal{F}$ and a) or b) is satisfied, then Condition G_3 holds.

Applications of the previous results are found in the next section. The proofs of Propositions 4 and 6 are deferred to Appendix B.

4. Application to Hypothesis Testing

In this section, we provide applications of the previous results in the context of hypothesis testing.

4.1 $S_n(\omega)$ as a U-statistic of Degree 1

In this section, we investigate the behaviour of $S_n(\omega)$ and the kernelised estimator Ψ_n when $S_n(\omega)$ is a U-statistic of degree 1, meaning that $S_n(\omega)$ is the sum of i.i.d. random variables. This is arguably the simplest case of a kernelised test-statistic that we can consider, yet it gives rise to several important statistics that appear in practical applications, including the so-called maximum mean discrepancy and kernelised Stein discrepancy. Since $S_n(\omega)$ is very simple, we can find a straightforward expression for Ψ_n. In fact, it can be represented as a V-statistic of order 2. Although this type of representation is common in the literature of kernel tests, it is important to clarify that not all kernel test-statistics can be written as a standard V-statistic.

To fix our ideas, let us consider data $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} F$ taking values in a space $\mathcal{Y}$. Consider a separable Hilbert space $\mathcal{H}$ of functions from $\mathcal{X} \to \mathbb{R}$, and a subspace $\mathcal{G}$ of $L_2(\mathcal{Y}, F)$. In this context, $\mathcal{X}$ and $\mathcal{Y}$ may not necessarily be identical. Although they are often equal in various applications, our approach accommodates the possibility of them being distinct. This flexibility proves advantageous in certain practical scenarios; for example, in Section 4.3.3 we consider survival analysis data $(X_i, \Delta_, g_i)$ thus $\mathcal{Y} = \mathbb{R} \times \{0,1\} \times \{0,1\}$ but $\mathcal{X}$ will be chosen as $\mathbb{R}$ since the weight functions will only consider the time X_i.

Let $U : \mathcal{H} \to \mathcal{G}$ be a linear map and define the linear test-statistic S_n as:

$$S_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U\omega)(Y_i). \tag{12}$$

If $\mathcal{G}$ is an RKHS with kernel L, then $(U\omega)(y) = \langle U\omega, L_y \rangle_{\mathcal{G}} = \langle \omega, U^* L_y \rangle_{\mathcal{H}}$ where U^* is the adjoint operator of U, and recall the definition of L^U from Eq. (6). Then, we can express the kernelised statistic Ψ_n as a V-statistic with kernel L^U, indeed:

$$\Psi_n = \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}} = 1} S_n(\omega)^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} L^U(Y_i, Y_j) \tag{13}$$

which holds by noting that $S_n(\omega)^2 = \langle \omega, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U^* L_{Y_i} \rangle_{\mathcal{H}}^2$, thus the supremum over the unit ball equals the square of the norm of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U^* L_{Y_i}$ in $\mathcal{H}$.

We note that although the definition of L^U seems rather technical, it is not difficult to obtain this kernel in practice. In fact, a common approach in the literature, although informal, is to assume that $\mathcal{H}$ is an RKHS with kernel K, and then to obtain L^U we simply apply the transformation U to each coordinate of K.

The previous procedure can be formalised. Assume that $\omega \to (U\omega)(y)$ is a bounded linear functional for each $y \in \mathcal{Y}$, and let $N = \{\omega : (U\omega)(y) = 0 \text{ for all } y \in \mathcal{Y}\}$, and define $\mathcal{G} := \{U\omega : \omega \in N^\perp\} = \{U\omega : \omega \in \mathcal{H}\}$. Now, for every $g \in \mathcal{G}$, there exists $\omega \in N^\perp$ such that $g = U\omega$, then define the inner product in $\mathcal{G}$ by $\langle U\omega, U\omega' \rangle_{\mathcal{G}} = \langle \omega, \omega' \rangle_{\mathcal{H}}$ with $\omega, \omega' \in N^\perp$. Then U is an isometry between $N^\perp$ and $\mathcal{G}$. We claim that $\mathcal{G}$ with this inner product is an RKHS. Indeed, since $\omega \to (U\omega)(y)$ is a bounded functional in $N^\perp$, there exists $\xi_y \in N^\perp$ such that $\langle \omega, \xi_y \rangle_{\mathcal{H}} = (U\omega)(y)$ for $\omega \in N^\perp$, then we claim that $U\xi_x \in \mathcal{G}$ evaluates functions of $\mathcal{G}$, indeed for $g = (U\omega)$, we have $\langle U\xi_y, g \rangle_{\mathcal{G}} = \langle \xi_y, \omega \rangle_{\mathcal{H}} = (U\omega)(y) = g(y)$. Denote by $L_y = U\xi_y$ and the kernel $L(y, y') = \langle L_y, L_{y'} \rangle_{\mathcal{G}}$. Finally, since U is an isometry between $N^\perp$ and $\mathcal{G}$, UU^* is the identity in $\mathcal{G}$, and then

$$L^U(y, y') = \langle U^* L_y, U^* L_{y'} \rangle_{\mathcal{H}} = \langle L_y, UU^* L_{y'} \rangle_{\mathcal{G}} = L(y, y').$$

Note that the above argument also shows that $L^U(y, y') = \langle \xi_y, \xi_{y'} \rangle_{\mathcal{H}}$, therefore, we just need to find $\xi_y \in N^\perp$ such that $\langle \omega, \xi_y \rangle_{\mathcal{H}} = (U\omega)(y)$ for $\omega \in N^\perp$. Of course, finding $N^\perp$ may be difficult, but in many examples N is trivial (only contains the zero function), thus $N^\perp = \mathcal{H}$. In this case, note that $\langle K_x, \xi_y \rangle_{\mathcal{H}} = \xi_y(x) = (UK_x)(y)$. Hence, we can find $\xi_y(x)$ applying U to the second coordinate of $K(x, x')$ and then evaluating at y. Finally, since $L^U(y, y') = (U\xi_y)(y')$, we are just applying U to the function $x \to (UK_x)(y)$, explaining the heuristic idea that L^U is found by applying U to the first and second coordinate of K.

To illustrate the previous ideas, consider the following examples.

Example 1 (Maximum mean discrepancy for the goodness-of-fit) *In this example, we consider data $(Y_i)_{i=1}^n$ in $\mathbb{R}^d$ generated independently according to a distribution F. We want to test whether F is equal to a given distribution F_0 or not. Let $\mathcal{H}$ be an RKHS of functions $\mathbb{R}^d \to \mathbb{R}$, and define $(U\omega)(y) = \omega(y) - \mathbb{E}_0(\omega(Y))$, where $\mathbb{E}_0$ denotes expectation w.r.t. F_0. Then, informally, the statistic $\Psi_n = \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}} = 1} \sum_{i=1}^{n} \left(\frac{1}{\sqrt{n}} (U\omega)(Y_i) \right)^2$ can be* *written as the V-statistic $\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n L^U(Y_i,Y_j)$ where*

$$L^U(y,y') = (U^1U^2K)(y,y') = K(y,y') - \mathbb{E}_0(K(y,Y')) - \mathbb{E}_0(K(Y,y')) + \mathbb{E}_0(K(Y,Y')),$$

where U^i is the application of U to the i-th coordinate of K.

Example 2 (Kernel Stein Discrepancy) *In this example we want to test the null hypothesis $F = F_0$. We assume that F_0 has density p_0 in $\mathbb{R}^d$ with vanishing tails. Consider an RKHS $\mathcal{H}$ of real functions and let K be its kernel function. Define the RKHS $\mathcal{H}_d$ of functions $\mathbb{R}\times\{1,\dots,d\}\to\mathbb{R}$ in terms of its kernel K_d given by $K_d((x,i),(y,j)) = K(x,y)\delta_{ij}$, which is a kernel since it is the product of two kernels. Denote $g_i(x) = (\partial/\partial x_i)\log(p_0(x))$, and consider the operator U_i acting on $\mathcal{H}_d$, given by $(U_i\omega)(x) = g_i(x)\omega(x,i) + \frac{\partial}{\partial x_i}\omega(x,i)$, and $(U\omega)(x) = \sum_{i=1}^d (U_i\omega)(x)$. For practical purposes, suppose that $\lim_{h\to 0}\frac{K_{x+h}-K_x}{h}$ converges in $\mathcal{H}$. To ensure the existence of the previous limit, we shall assume that the kernel K has continuous second-order partial derivatives. Then, we apply U to the first and second coordinates of K to, informally, derive that*

$$L^U(x,y) = \sum_{i=1}^d g_i(x)g_i(x)K(x,y) + g_i(x)\frac{\partial}{\partial y_i}K(x,y) + g_i(y)\frac{\partial}{\partial x_i}K(x,y) + \frac{\partial}{\partial x_i}\frac{\partial}{\partial y_i}K(x,y).$$

In general, V-statistics of order 2 can be categorised into two general classes, degenerate and non-degenerate, which determine their asymptotic behaviour. A V-statistic is considered degenerate if the function $y \to \mathbb{E}(L^U(Y,y))$ is almost surely a constant with respect to y. On the other hand, if the function varies across different values of y, we say that the V-statistic is non-degenerate.

For hypothesis testing applications, U is chosen so that under the null hypothesis $\mathbb{E}((U\omega)(Y)) = 0$ for all $\omega\in\mathcal{H}$. Indeed, note that in our two previous examples this property holds. In this case, L^U is a degenerated V-statistic kernel since

$$\mathbb{E}(L^U(Y,y)) = \mathbb{E}(\langle U^*L_Y, U^*L_y\rangle_{\mathcal{H}}) = \mathbb{E}(\langle L_Y, UU^*L_y\rangle_{\mathcal{G}}) = \mathbb{E}((UU^*L_y)(Y)) = 0,$$

for all $y\in\mathcal{Y}$ since $U^*L_y\in\mathcal{H}$. On the other hand, under the alternative hypothesis, we expect that there exists $\omega\in\mathcal{H}$ such that $\mathbb{E}((U\omega)(Y))\neq 0$. If the latter holds, we find that L^U is not a degenerate kernel.

Since our main interest lies in applications in hypothesis testing, it becomes crucial to be able to bootstrap our test-statistics to approximate the rejection region. For this purpose, we propose a wild bootstrap approach that can be used within this setting: Consider a collection $(W_i)_{i=1}^n$ of i.i.d. Rademacher random variables, i.e., with equal probability they take value 1 or -1 (in general, random variables with mean 0 and variance 1 are enough). Then, the bootstrap statistics are defined as

$$S_n^W(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(U\omega)(Y_i), \quad \text{and} \quad \Psi_n^W = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^W(\omega)^2 = \frac{1}{n}\sum_{i,j=1}^n W_iW_jL^U(Y_i,Y_j).$$

The next result provides asymptotic results for Ψ_n and Ψ_n^W.

Theorem 7 *Let $U : \mathcal{H} \to \mathcal{G} \subseteq L_2(\mathcal{Y}, F)$, and suppose that $\mathcal{G}$ is an RKHS with kernel L and such that $\mathbb{E}(L^U(Y,Y)) < \infty$ for $Y \sim F$. Define the bilinear forms in $\mathcal{H}$,*

$$\sigma(\omega, \omega') = \mathbb{C}ov((U\omega)(Y), (U\omega')(Y)) \quad \text{and} \quad \sigma^W(\omega, \omega') = \mathbb{E}((U\omega)(Y)(U\omega')(Y)),$$

then

1. *If $\mathbb{E}((U\omega)(Y)) = 0$ for all $\omega \in \mathcal{H}$, then $\Psi_n \overset{\mathcal{D}}{\to} \Psi \sim \chi^2(\sigma)$.*
2. *$\Psi_n^W \overset{\mathcal{D}_{\mathcal{D}}}{\to} \Psi^W \sim \chi^2(\sigma^W)$. Moreover, if $\mathbb{E}((U\omega)(Y)) = 0$ for all $\omega \in \mathcal{H}$, then Ψ^W has the same distribution as Ψ.*
3. *Let $c_*^2 = \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}=1} \mathbb{E}((U\omega)(Y))^2$, then $n^{-1}\Psi_n \overset{a.s.}{\to} c_*^2$.*

We note that the previous result includes all the necessary elements that allow us to prove the correctness of the testing procedure based on Ψ_n and the wild bootstrap estimator Ψ_n^W to approximate the rejection region.

The proof of Theorem 7 can be carried out using results of the theory of U and V statistics. For example, Item 1. is a standard convergence result for degenerated V-statistics, e.g. Theorem 4.3.2 of Koroljuk and Borovskich (1994) establishes that Ψ_n converges in distribution to a weighted linear combination of independent chi-square random variables with one degree of freedom if $\mathbb{E}(L^U(Y,Y)) < \infty$ and $\mathbb{E}(L^U(Y,Y')^2) < \infty$, where Y' is an independent copy of Y. It is important to note that the condition $\mathbb{E}(L^U(Y,Y')^2) < \infty$ is implied by $\mathbb{E}(L^U(Y,Y)) < \infty$ since L^U is a positive-definite function. The wild bootstrap estimator Ψ_n^W was studied by Dehling and Mikosch (1994). Their Theorem 3.1 shows that $\Psi_n^W \overset{\mathcal{D}_{\mathcal{D}}}{\to} \Psi^W$ under the same moment conditions mentioned above, and that the limit distributions coincide if $\mathbb{E}(U\omega)(Y)) = 0$ for all $\omega \in \mathcal{H}$, leading to Item 2. Finally, Item 3 is the standard law of large numbers for V-statistics.

For completeness, we provide an alternative proof of Theorem 7 using our tools, ignoring all previous developments in the theory of U-statistics and wild bootstrap.

Proof Observe that $\mathbb{E}\left(\sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}=1} (U\omega)(Y)^2\right) < \infty$ since

$$\mathbb{E}\left(\sup_{\omega \in B_1(\mathcal{H})} (U\omega)(Y)^2\right) = \mathbb{E}\left(\sup_{\omega \in B_1(\mathcal{H})} \langle U\omega, L_Y\rangle_{\mathcal{G}}^2\right) = \mathbb{E}\left(\sup_{\omega \in B_1(\mathcal{H})} \langle \omega, U^* L_Y\rangle_{\mathcal{H}}^2\right)$$
$$= \mathbb{E}\left(\langle U^* L_Y, U^* L_Y\rangle_{\mathcal{H}}\right) = \mathbb{E}(L^U(Y,Y)) < \infty.$$

For item 1, we apply Theorem 1 to the operator S_n. Clearly $S_n(\omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (U\omega)(Y_i)$ converges to a centred normal with variance $\mathbb{V}ar((U\omega)(Y_i))$ by the CLT, thus Condition G_0 holds. To verify Conditions G_1 and G_2 we use Proposition 4. Note that since $\mathbb{E}((U\omega)(Y_i)) = 0$ for all ω, it holds that

$$\mathbb{E}\left(S_n(\omega)^2\right) \leq \int_{\mathcal{Y}} (U\omega)(y)^2 dF(y), \tag{14}$$

thus, Conditions G_1 and G_2 follows by taking $Q = U$, $C = 1$ and $\nu = F$ in Item b) of Proposition 4.

For item 2, we apply Theorem 1 to the operator S_n^W but conditional on the data $D = (Y_i)_{i\geq 1}$. Condition G_0 holds $\sigma(D)$-stably since $S_n^W(\omega)$ has conditional mean 0 and conditional variance given by $\sum_{i=1}^n \frac{1}{n}(U\omega)(Y_i)^2 \to \mathbb{E}((U\omega)(Y_i)^2)$ by the law of large numbers. An application of the Linderberg CLT (see Lemma 16 and Corollary 17 in Appendix A) shows that $S_n^W(\omega) \overset{\mathcal{D}_D}{\to} N(0, \sigma(\omega,\omega))$ with $\sigma(\omega,\omega') = \mathbb{E}((U\omega)(Y_1)(U\omega')(Y_1))$. To verify Conditions G_1 and G_2 we again use Proposition 6 with $Q = U$, $C = 1$, $\nu_n = \sum_{i=1}^n \delta_{Y_i}$, $\nu = F$, and $\mathcal{F}' = \sigma(D)$ the sigma algebra generated by the data. In this case, note that $\int_{\mathcal{Y}} g(y)\nu_n(dy) \to \int_{\mathcal{Y}} g(y)\nu(dy)$ by the Law of Large numbers whenever $g \in L_1(\mathcal{Y},\nu)$. Moreover, note that Eq. (10) holds since

$$\mathbb{E}_D\left(S_n^W(\omega)^2\right) = \frac{1}{n}\sum_{i=1}^n (U\omega)(Y_i)^2.$$

Therefore, since item b) of Proposition 6 holds, we have that Ψ_n^W converges in distribution. Furthermore, note that if $\mathbb{E}((U\omega)(Y_1)) = 0$ for all ω then the covariance operators of Ψ and Ψ^W coincide.

Finally, for item 3, we apply Theorem 3 to the operator $\frac{1}{\sqrt{n}}S_n$. Clearly, $\frac{1}{\sqrt{n}}S_n(\omega) = \frac{1}{n}\sum_{i=1}^n (U\omega)(Y_i)$ converges almost surely to a constant due to the law of large numbers. We only need to verify Condition G_3, for which we employ Proposition 6. A direct calculation yields $\left(\frac{1}{\sqrt{n}}S_n(\omega)\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n (U\omega)(Y_i)\right)^2 \leq \frac{1}{n}\sum_{i=1}^n (U\omega)(Y_i)^2$ so, in our application of Proposition 6 we take $Q = U$, $C = 1$, $\mathcal{F}' = \mathcal{F}$, $\nu_n = \sum_{i=1}^n \delta_{Y_i}$ and $\nu = F$. By our assumptions, condition b) of Proposition 6 holds, and thus Condition G_3 holds, that is, $\lim_{i\to\infty}\limsup_{n\to\infty}\frac{1}{n}\sum_{j=i}^\infty S_n(\phi_j)^2 = 0$ for some orthonormal basis ϕ_i of $\mathcal{H}$. ■

4.2 $S_n(\omega)$ as U-statistic of Degree 2 or More.

A much more interesting class of statistics arises when we move beyond simple sums and consider general U-statistics (with sums being U-statistics of degree 1). Let $\binom{[n]}{r}$ denote the family of subsets of $[n] = \{1,\dots,n\}$ with r distinct elements. The elements of $\binom{[n]}{r}$ represent a set of indices. For $A \in \binom{[n]}{r}$ we denote by $A_1,\dots,A_r$ the elements of the set A in increasing order. In this context, consider i.i.d. data $X_i \sim F$, and consider linear statistics S_n in the following form:

$$S_n(\omega) = \frac{\sqrt{n}}{r\binom{n}{r}}\sum_{A\in\binom{[n]}{r}} (U\omega)(\boldsymbol{X}_A), \tag{15}$$

where $\boldsymbol{X}_A = (X_{A_1},\dots,X_{A_r})$.

In this case, $S_n(\omega)$ represents a U-statistic of degree r (which is why we use the letter U to denote the transformation). In this context, it is convenient to assume that $\mathcal{H}$ is a space of functions that map $\mathcal{X}^d$ to $\mathbb{R}$. In particular, we can consider U as $U : \mathcal{H} \to L_2(\mathcal{X}^r, F^{\times r})$, with $F^{\times r}$ denoting the product measure. Note that d may be different from r, leading to a general enough setting for applications. It is important to assume that for every ω, the function $U\omega$ is symmetric in its input (i.e., the output does not depend on the order of the input), so the ordering of the index set A is not important. For the Wild Bootstrap version of $S_n(\omega)$, we propose the following:

$$S_n^W(\omega) = \frac{\sqrt{n}}{r\binom{n}{r}} \sum_{A \in \binom{[n]}{r}} \mathbf{W}_A (Uw)(\boldsymbol{X}_A),$$

where $\mathbf{W}_A = W_{A_1} + W_{A_2} + \cdots + W_{A_r}$, and $(W_i)_{i=1}^n$ are i.i.d. Rademacher random variables (however, other random variables with mean 0 and variance 1 can be considered). Note that when $r = 1$, we recover the previous setting discussed in Section 4.1. We proceed to analyse the kernel test-statistics Ψ_n and Ψ_n^W when $r \geq 2$. Our approach is not an extension of the case $r = 1$ since, while it is possible to expand Ψ_n and Ψ_n^W as V-statistics, the analyses of such expressions are rather intricate, involving the dependent random variables $(\boldsymbol{X}_A : A \in \binom{[n]}{r})$. Hence, the usual roadmap for analysing V-statistics becomes very complicated. We think that working with the random functional S_n is a better option in this setting, particularly condition a) of Propositions 4 and 6 is not hard to apply.

The following theorem gives a general tool to analyse statistics Ψ_n as described above (see Section 4.3.2 for an application).

Theorem 8 *Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F$ be random variables taking values in $\mathcal{X}$, and for $r \geq 1$ define the bilinear forms in the RKHS $\mathcal{H}$,*

$$\sigma(\omega, \omega') = \mathbb{C}ov((U\omega)(X_1, X_2, \ldots, X_r), (U\omega')(X_1, X_2', \ldots, X_r')) \tag{16}$$

and

$$\sigma^W(\omega, \omega') = \mathbb{E}((U\omega)(X_1, X_2, \ldots, X_r)(U\omega')(X_1, X_2', \ldots, X_r')) \tag{17}$$

where X_i' have the same distribution than X_i for $i \geq 2$, and $X_2', \ldots, X_r'$ are independent of $X_1, X_2, \ldots, X_r$ (for $r = 1$ we recover the covariances of Theorem 7).

Suppose that there exists a sigma-finite measure μ in $\mathcal{X}^d$ such that

$$\sup_{\omega \in \mathcal{H}: \|\omega\|_{L_2(\mu)}=1} \mathbb{E}\big((U\omega)(X_1, X_2, \ldots, X_r)^2\big) < \infty \text{ and } \int_{\mathcal{X}^d} K(\mathbf{x}, \mathbf{x})\mu(d\mathbf{x}) < \infty, \tag{18}$$

where K is the kernel of $\mathcal{H}$. Then the following statements hold

1. *If $\mathbb{E}((U\omega)(X_1, \ldots, X_r)) = 0$ for all $\omega \in \mathcal{H}$, then $\Psi_n \xrightarrow{\mathcal{D}} \Psi \sim \chi^2(\sigma)$.*

2. *$\Psi_n^W \xrightarrow{\mathcal{D}_{\mathcal{D}}} \Psi^W \sim \chi^2(\sigma^W)$, moreover, if $\mathbb{E}((U\omega)(X_1, \ldots, X_r)) = 0$ for all $\omega \in \mathcal{H}$, then Ψ^W and Ψ have the same distribution.*

3. *There exists a constant $c_* \geq 0$ such that $n^{-1}\Psi_n \xrightarrow{a.s.} c_*^2$.*

Proof We provide an independent proof for each item.

Item 1. For this item, we use Theorem 1. Condition G_0 holds immediately since $S_n(\omega)$ is a U-statistic, so according to Theorem A in Section 5.5.1 of Serfling (1980), $S_n(\omega) \xrightarrow{\mathcal{D}} N(0, \sigma(\omega, \omega))$ where $\sigma(\omega, \omega') = \mathbb{E}\left((U\omega)(X_1, X_2, \ldots, X_r)(U\omega')(X_1, X_2', \ldots, X_r')\right)$. To verify

Condition G_1 and Condition G_2, we use Proposition 4 with $Q = U$ and $\mu = \nu = F^{\times r}$. Note that Eq. (18) implies that the right-hand side of the inequality in Eq. (8) is finite, so the conclusion of Proposition 4 follows.

Item 2. We will verify the assumptions of Theorem 1, conditioned on the entire data $D = (X_i)_{i\geq 1}$. To use Theorem 1, we begin by claiming that Condition G_0 is valid. This follows from Linderberg's CLT, in particular, Corollary 17 (in Appendix A) states that for any U-statistic kernel $U\omega$ with finite second moment we have

$$S_n^W(\omega) = \frac{\sqrt{n}}{r\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \mathbf{W}_A \left(U\omega\right)\left(\mathbf{X}_A\right) \overset{\mathcal{D}_D}{\to} N(0, \sigma^W(\omega,\omega)). \tag{19}$$

To verify Conditions G_1 and G_2, we use Proposition 6 with $\mathcal{F}'$ as the sigma-algebra generated by the data D. We proceed to verify the conditions of Proposition 6, and, in particular, we will verify condition a) in this application. For $i \in [n]$, let I_i denote the family of subsets in $\binom{[n]}{r}$ such that $A \in I_i$ if and only if $i \in A$. Then, $S_n^W(\omega)$ can be written as follows:

$$S_n^W(\omega) = \frac{\sqrt{n}}{r\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \mathbf{W}_A \left(U\omega\right)\left(\mathbf{X}_A\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i \frac{1}{\binom{n-1}{r-1}} \sum_{A\in I_i} (U\omega)(\mathbf{X}_A). \tag{20}$$

We define $Y_{in} = \frac{1}{\binom{n-1}{r-1}} \sum_{A\in I_i} (U\omega)(\mathbf{X}_A)$. Consequently, we have $S_n^W(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i Y_{in}$. We observe that $\mathbb{E}_D(S_n^W(\omega)^2) = \frac{1}{n}\sum_{i=1}^{n} Y_{in}^2$, then the definition of Y_{in} coupled with Jensen's inequality yields

$$Y_{in}^2 = \left(\frac{1}{\binom{n-1}{r-1}} \sum_{A\in I_i} (U\omega)(\mathbf{X}_A)\right)^2 \leq \frac{1}{\binom{n-1}{r-1}} \sum_{A\in I_i} (U\omega)(\mathbf{X}_A)^2.$$

Thus, we obtain

$$\begin{aligned}\mathbb{E}_D\left(S_n^W(\omega)^2\right) \leq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\binom{n-1}{r-1}} \sum_{A\in I_i} (U\omega)(\mathbf{X}_A)^2 &= \frac{r}{n\binom{n-1}{r-1}} \sum_{A\in\binom{[n]}{r}} (U\omega)(\mathbf{X}_A)^2 \\ &= \frac{1}{\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \left((U\omega)(\mathbf{X}_A)\right)^2. \end{aligned} \tag{21}$$

This implies that Eq. (10) holds when we take $Q = U$ and ν_n as the empirical measure on $\mathbb{R}^r$ given by $\nu_n(S) = \frac{1}{\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \delta_S(\mathbf{X}_A)$, where $S \subseteq \mathbb{R}^r$ is a Borel set. Note that $\int_{\mathbb{R}^r} (U\omega)^2 \nu_n = \frac{1}{\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \left(U(\omega)(\mathbf{X}_A)\right)^2$. By the Law of Large Numbers for U-statistics, if $f : \mathbb{R}^r \to \mathbb{R}$ is symmetric and such that $E(|f(X_1,\ldots,X_r)|) < \infty$, then $\int f\mu_n \to E(f(X_1,\ldots,X_r))$, as required in the statement of Proposition 6. Finally, with our choice of Q and ν_n, we see that Eq. (8) is immediately verified by Eq. (18). We conclude that there exists a random variable Ψ^W such that $\Psi_n^W \overset{\mathcal{D}_D}{\to} \Psi$, and the covariance kernel is given by $\sigma^W(\omega,\omega))$ by Eq. (19). Finally, note that when $E((U\omega)(X_1,\ldots,X_r)) = 0$, then

σ^W and σ are the same covariance, and thus the corresponding limit distributions Ψ^W and Ψ are the same.

Item 3. Note that $\frac{1}{\sqrt{n}} S_n(\omega)$ converges almost surely to a constant $c(\omega)$ by the law of large numbers for U-statistics. Then, to invoke Theorem 3 we have to verify that Condition G_3 holds for $\frac{1}{\sqrt{n}} S_n(\omega)$. For this, we use Proposition 6 taking $\mathcal{F}' = \mathcal{F}$ (i.e. we do not take expectation in Eq. (10)). By Jensen's inequality and following Eq. (20) and Eq. (21) we get

$$\left(\frac{1}{\sqrt{n}} S_n(\omega)\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} W_i \frac{1}{\binom{n-1}{r-1}} \sum_{A \in I_i} (U\omega)(\mathbf{X}_A)\right)^2 \leq \frac{r}{n} \frac{1}{\binom{n-1}{r-1}} \sum_{A \in \binom{[n]}{r}} (U\omega)(\mathbf{X}_A)^2$$

and note that the latter expression is the same as the right-hand side of equation Eq. (21), hence by repeating the argument after Eq. (21) we get that the premises of Proposition 6 hold (particularly condition a)), implying that Condition G_3 holds. ■

4.3 Examples

In this section, we apply our tools to specific testing problems.

4.3.1 A Kernel Test for Conditional Independence

We present a new kernel test to test conditional independence that, asymptotically, fits within the framework of Section 4.1. Consider data points $(X_i, Y_i, Z_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$, where P is a probability measure on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ with $d \geq 1$. We are interested in testing whether X and Y are conditionally independent given Z, that is, we want to test $H_0 : X \perp Y | Z$ against $H_a : X \not\perp Y | Z$. In order to do this we use the Generalised Covariance Measure (**GCM**) that was introduced by Shah and Peters (2020):

$$\mathbf{GCM}(X, Y; Z) = \mathbb{E}(\epsilon_X(Z)\epsilon_Y(Z)),$$

where the error terms above are obtained from the following decomposition:

$$X = f(Z) + \epsilon_X(Z), \qquad \text{and} \qquad Y = g(Z) + \epsilon_Y(Z),$$

where $f(z) = \mathbb{E}(X|Z = z)$ and $g(z) = \mathbb{E}(Y|Z = z)$. Note that this decomposition always exists for integrable X and Y.

Under the null hypothesis $\mathbf{GCM}(X, Y; Z) = 0$, so the **GCM** can be used as a parameter to test the null hypothesis. In Scheidegger et al. (2022), a weighted generalisation of the **GCM**, denominated the weighted generalised covariance measure (**wGCM**), was introduced. Given a weight function $\omega : \mathbb{R}^d \to \mathbb{R}$, the **wGCM** is defined as

$$\mathbf{wGCM}(X, Y; Z) = \mathbb{E}(\omega(Z)\epsilon_X(Z)\epsilon_Y(Z)). \tag{22}$$

Again, under the null hypothesis we have that $\mathbf{wGCM}(X, Y; Z) = 0$ for any $\omega \in \mathcal{H}$. The motivation behind the weighted generalisation of the **GCM** is that under some alternatives, we may have $\mathbf{GCM}(X, Y; Z) = 0$. However, if we choose an appropriate weight, we will get $\mathbf{wGCM}(X, Y; Z) \neq 0$, and so the weighted version should be more robust.

In order to use the **GCM** and the **wGCM**, we need to estimate ϵ_X and ϵ_Y from the data. This can be achieved by estimating the conditional expectation of X and Y given Z using a regression estimator. Let $\widehat{f}$ and $\widehat{g}$ be the regression estimators of $\mathbb{E}(X|Z)$ and $\mathbb{E}(Y|Z)$, respectively. Here, $\widehat{f}$ is estimated using $(X_i, Z_i)_{i=1}^n$ while $\widehat{g}$ is estimated using $(Y_i, Z_i)_{i=1}^n$. Then define

$$\widehat{\epsilon}_{X_i}(Z_i) = X_i - \widehat{f}(Z_i) \quad \text{and} \quad \widehat{\epsilon}_{Y_i}(Z_i) = Y_i - \widehat{g}(Y_i), \tag{23}$$

and note that we can estimate $\mathbf{wGCM}(X, Y; Z)$ by computing $\frac{1}{n}\sum_{i=1}^n \omega(Z_i)\widehat{\epsilon}_{X_i}(Z_i)\widehat{\epsilon}_{Y_i}(Z_i)$, which should be close to 0 under the null hypothesis.

To kernelise the **wGCM**, we consider an RKHS $\mathcal{H}$ of functions $\mathbb{R}^d \to \mathbb{R}$. Then, with a convenient rescaling, we define the test-statistic $S_n(\omega)$ and its bootstrap version $S_n^W(\omega)$ by

$$S_n(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \omega(Z_i)\widehat{\epsilon}_{X_i}(Z_i)\widehat{\epsilon}_{Y_i}(Z_i), \quad \text{and} \quad S_n^W(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\omega(Z_i)\widehat{\epsilon}_{X_i}(Z_i)\widehat{\epsilon}_{Y_i}(Z_i),$$

respectively, where W_i are i.i.d. Rademacher random variables. Then, the kernelised versions of S_n and S_n^W are given by

$$\Psi_n = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n(\omega)^2 \quad \text{and} \quad \Psi_n^W = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} S_n^W(\omega)^2.$$

To analyse the kernelised **GCM** we need some conditions on the regression estimators $\widehat{f}$ and $\widehat{g}$, in order to ensure that the estimation is good enough, as well as some other regularity conditions.

Condition GCM *Consider the following quantities:*

$$A_f = \frac{1}{n}\sum_{i=1}^n (f(Z_i) - \widehat{f}(Z_i))^2, \qquad u_f(z, y) = \mathbb{E}(\epsilon_X(Z)^2|Z = z, Y = y)$$
$$A_g = \frac{1}{n}\sum_{i=1}^n (g(Z_i) - \widehat{g}(Z_i))^2, \qquad v_g(z, x) = \mathbb{E}(\epsilon_Y(Z)^2|Z = z, X = x).$$

We assume that the following conditions hold.

i. $A_f = o_p(n^{-1/2})$ *and* $A_g = o_p(n^{-1/2})$.

ii. $u_f(z, y)$ *and* $v_g(z, x)$ *are uniformly bounded.*

iii. $0 < \mathbb{E}(\epsilon_X^2(Z)\epsilon_Y^2(Z))$.

iv. There exists a constant $C > 0$ *such that* $|K(z, z')| \leq C$ *for all* $z, z' \in \mathbb{R}^d$.

Remark 9 *The conditions above are slightly stronger than the corresponding conditions of Scheidegger et al. (2022), but they allow us to avoid splitting the data as Scheidegger et al. (2022) (e.g. use half of the data to estimate f and g, and the other half in the testing* *procedure). Our conditions now require the conditional variances $u_f(z,y)$ and $v_g(z,x)$ to be uniformly bounded, which implies $\mathbb{E}(\epsilon_X^2\epsilon_Y^2) < \infty$. This is not restrictive at all, and it is a common assumption in practice. Moreover, the condition can easily be relaxed at the price of having less clear statements and longer proofs.*

We proceed to show that the kernelised testing procedure is correct. Furthermore, for completeness, we provide a small empirical evaluation of the testing procedure in Appendix C.

Corollary 10 *Define the bilinear covariance operators in $\mathcal{H}$, σ and σ^W, by*

$$\begin{aligned} \sigma(\omega,\omega') &= \mathbb{C}ov\left(\omega(Z)\epsilon_X(Z)\epsilon_Y(Z), \omega'(Z)\epsilon_X(Z)\epsilon_Y(Z)\right) \quad \textit{and} \\ \sigma^W(\omega,\omega') &= \mathbb{E}\left(\omega(Z)\omega'(Z)\epsilon_X(Z)^2\epsilon_Y(Z)^2\right). \end{aligned} \tag{24}$$

Then, under Condition GCM it holds that:

1. *Under the null hypothesis of conditional independence we have $\Psi_n \xrightarrow{\mathcal{D}} \Psi \sim \chi^2(\sigma)$.*
2. *Under the null or alternative, $\Psi_n^W \xrightarrow{\mathcal{D}_\mathcal{D}} \Psi^W \sim \chi^2(\sigma^W)$. Furthermore, under the null hypothesis, Ψ and Ψ^W have the same distribution.*
3. *Under the null or alternative, there exists $c_* \geq 0$ such that $\frac{1}{n}\Psi_n \xrightarrow{\mathbb{P}} c_*^2$.*

Proof Following the proof of Theorem 6 of Shah and Peters (2020), under Condition GCM, it holds that

$$S_n(\omega) = \widetilde{S}_n(\omega) + o_p(1), \qquad \text{where} \qquad \widetilde{S}_n(\omega) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_{X_i}(Z_i)\epsilon_{Y_i}(Z_i)\omega(Z_i),$$

where the $o_p(1)$ term does not depend on ω (the argument of Theorem 6 of Shah and Peters (2020) holds for ω as a constant function, but it can easily be adapted to bounded functions). Therefore, any limit theorem for $\sup_{\omega\in\mathcal{H}:\|\omega\|_\mathcal{H}=1}\widetilde{S}_n(\omega)^2$ applies for Ψ_n.

The linear test-statistic $\tilde{S}_n(\omega)$ falls within the framework of Section 4.1 by defining $U : \mathcal{H} \to L_2(\mathbb{R}\times\mathbb{R}\times\mathbb{R}^d, P)$ as $(U\omega)(x,y,z) = \epsilon_x(z)\epsilon_y(z)\omega(z)$. Therefore, we will invoke Theorem 7 to prove the result. This task requires us to identify the codomain of U, say $\mathcal{G}$, the kernel $L^U(Z,Z)$, and to prove $\mathbb{E}(L^U(Z,Z)) < \infty$.

Define the vector space $\mathcal{G} = \{U\omega : \omega \in \mathcal{H}\}$. Note that U is a bijection between $\mathcal{H}$ and $\mathcal{G}$. This bijection follows from noting that $\omega(z) = \frac{(U\omega)(x,y,z)}{\epsilon_x(z)\epsilon_y(z)}$ when $\epsilon_x(z)\epsilon_y(z) \neq 0$. Note that $x \to \epsilon_x(z)$ and $y \to \epsilon_y(z)$ take the value 0 if and only if $x = \mathbb{E}(X|Z=z)$ and $y = \mathbb{E}(Y|Z=z)$, respectively. Thus, we can choose any other value of x and y such that $\epsilon_x(z) \neq 0$ and $\epsilon_y(z) \neq 0$, and thus $\epsilon_x(z)\epsilon_y(z) \neq 0$.

Now, equip $\mathcal{G}$ with the inner product $\langle U\omega, U\omega'\rangle_\mathcal{G} = \langle \omega,\omega'\rangle_\mathcal{H}$, which is well-defined since U is a bijection. We then see U as a linear operator $\mathcal{H} \to \mathcal{G}$, and then $U^* = U^{-1}$ as U is unitary. We can easily verify that $L_{(x,y,z)}(\cdot) = \epsilon_x(z)\epsilon_y(z)(UK_z)(\cdot) \in \mathcal{G}$ evaluates the functions in $\mathcal{G}$ through the inner product. Indeed, $\langle U\omega, L_{(x,y,z)}\rangle_\mathcal{G} = \epsilon_x(z)\epsilon_y(z)\langle U\omega, UK_z\rangle_\mathcal{G} = \epsilon_x(z)\epsilon_y(z)\omega(z) = (U\omega)(x,y,z)$, thus $\mathcal{G}$ is an RKHS.

A simple computation allows us to find the kernel L^U. By using that U^*U equals the identity, it holds that

$$L^U((x,y,z),(x',y',z')) = \langle U^* L_{(x,y,z)}, U^* L_{(x',y',z')}\rangle_{\mathcal{H}} = \epsilon_x(z)\epsilon_y(z)\epsilon_{x'}(z')\epsilon_{y'}(z')K(z,z').$$

Finally, note that $L^U((X,Y,Z),(X,Y,Z)) < \infty$ by Condition GCM, therefore, we can apply Theorem 7. To finish the proof, note that under the null hypothesis of conditional independence, it holds $\mathbb{E}((U\omega)(X,Y,Z)) = 0$ for all $\omega \in \mathcal{H}$ since $\mathbb{E}(\epsilon_X(Z)\epsilon_Y(Z)\omega(Z)) = 0$ (which is required by items 1 and 2 of Theorem 7). ■

We recall that Items 1 and 2 are enough to show that our testing procedure using wild bootstrap will be calibrated (correct type-1 error); however, note that the value c_*^2 in item 3 is not necessarily different from 0, and thus we cannot ensure power under any alternative. Showing that c_*^2 is different from 0 usually requires $\mathbb{E}(S_n(\omega)) > 0$ for at least one function $\omega \in \mathcal{H}$, thus having a rich enough RKHS $\mathcal{H}$, e.g. universal, ensures that $c_*^2 > 0$. However, there are cases where $\mathbb{E}(S_n(\omega)) = 0$ for all weights ω. For example, this happens when we consider the **wGCM** together with the following way of generating data (X,Y,Z): Let (U,V) be uniformly sampled on the unit disk of $\mathbb{R}^2$, which means that $U^2+V^2 \leq 1$. Note $\mathbb{E}(UV) = 0$ as they are uncorrelated with a symmetric distribution around 0. Also, consider a non-negative random variable Z independent of (U,V). Define $X = ZU$ and $Y = ZV$. It is clear that given Z, X and Y are dependent. In this case $\mathbb{E}(X|Z) = \mathbb{E}(Y|Z) = 0$, therefore $\varepsilon_X = X$ and $\varepsilon_Y = Y$. Then, we have $\mathbb{E}(\varepsilon_X\varepsilon_Y|Z) = \mathbb{E}(XY|Z) = Z^2\mathbb{E}(UV|Z) = 0$, since $\mathbb{E}(UV|Z) = \mathbb{E}(UV) = 0$. Therefore, for any weight ω, it holds $\mathbb{E}(S_n(\omega)) = \mathbb{E}(\omega(Z)\varepsilon_X\varepsilon_Y) = 0$. This example shows that the weighted Generalised Covariance Measure will fail to recognise that X and Y are not conditional independent given Z. Since this holds for any weight, kernelising the statistic will not help to improve the performance of the **wGCM**.

4.3.2 The HSIC Statistic and Independence Testing

One of the most popular kernelised estimators that can be expressed as the supremum of U-statistics (as discussed in Section 4.2), is the HSIC. Consider n i.i.d. points $D_i = (X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}^q$. The HSIC estimator (Gretton et al., 2007) measures the independence between X_i and Y_i by comparing the joint distribution $P_{X,Y}$ with the product measure of the marginals $P_X \times P_Y$. This comparison is carried out by embedding the difference between these measures in a RKHS $\mathcal{H}$ of functions from $\mathbb{R}^p \times \mathbb{R}^q$ to $\mathbb{R}$. The HSIC estimator is then given by

$$\mathbf{HSIC}(X,Y) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1}\left(\frac{1}{n}\sum_{i=1}^n \omega(X_i,Y_i) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \omega(X_i,Y_j)\right)^2.$$

Define $U : \mathcal{H} \to L_2((\mathbb{R}^p\times\mathbb{R}^q)^2, P_{X,Y}\times P_{X,Y})$ by

$$(U\omega)(D_i,D_j) = \omega(X_i,Y_i) + \omega(X_j,Y_j) - \omega(X_i,Y_j) - \omega(X_j,Y_i).$$

Then, the test-statistic Ψ_n defined as the **HSIC** scaled by n can be written as

$$\Psi_n = n\mathbf{HSIC}(X,Y) = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1}\left(\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(U\omega)(D_i,D_j)\right)^2$$

which has the same form of Eq. (15) and thus the right scaling limit. A wild bootstrap version is given by

$$\Psi_n^W = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1}\left(\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(W_i+W_j)(U\omega)(D_i,D_j)\right)^2,$$

where $(W_i)_{i=1}^n$ are i.i.d. Rademacher random variables.

Corollary 11 *Let σ and σ^W be defined as in Eq. (16) and Eq. (17) (writing D_i instead of X_i) respectively, and suppose that*

$$\int K((x,y),(x,y))P_{X,Y}(dx,dy) < \infty, \qquad \textit{and} \qquad \int K((x,y),(x,y))P_X(dx)P_Y(dy) < \infty.$$

Then

1. *If $P_{X,Y} = P_X \times P_Y$ (i.e. under the null), then $\Psi_n \overset{\mathcal{D}}{\to} \Psi \sim \chi^2(\sigma)$, where σ is as defined in Eq. (16).*
2. *Under null or alternative, $\Psi_n^W \overset{\mathcal{D}_{\mathcal{D}}}{\to} \Psi^W \sim \chi^2(\sigma^W)$, where σ^W is defined in Eq. (17). Moreover, under the null hypothesis it holds that Ψ^W and Ψ have the same distribution.*
3. *Under null or alternative, there exists a constant $c_* \geq 0$ such that $\frac{1}{n}\Psi_n \overset{a.s.}{\to} c_*^2$.*

Proof Recall that the domain of the functions ω is $\mathbb{R}^p \times \mathbb{R}^q$, while the domain of $U\omega$ is $(\mathbb{R}^p\times\mathbb{R}^q)^2$. Let μ be a measure on $\mathbb{R}^p\times\mathbb{R}^q$ defined as $\mu = P_{X,Y} + P_XP_Y$. We just need to verify that μ satisfies Eq. (18) (we set $\mathcal{X} = \mathbb{R}^p\times\mathbb{R}^q$ in the setup of 4.2)

For $\omega\in\mathcal{H}$ with $\int_{\mathbb{R}^p\times\mathbb{R}^q}\omega(x,y)^2d\mu(x,y) = 1$, it holds by symmetry that

$$\mathbb{E}((U\omega)(D_1,D_2)^2) \leq 8\mathbb{E}(\omega(X_1,Y_1)^2 + \omega(X_1,Y_2)^2) = 8\int_{\mathcal{X}\times\mathcal{Y}}\omega(x,y)^2\mu(dx,dy) = 8$$

and note that $\int_{(\mathcal{X}\times\mathcal{Y})} K((x,y),(x,y))\mu(dx,dy) < \infty$, by the assumptions in the statement about the kernel K. Then Eq. (18) is satisfied, and thus the conclusion of each item of Theorem 8 holds. Note that for items 1 and 2 of Theorem 8 we have $\mathbb{E}((U\omega)(D_i,D_j)|D_j) = 0$ for all $\omega\in\mathcal{H}$ since, under the null hypothesis, $P_{X,Y} = P_XP_Y$, and thus (X_i,Y_i), (X_j,Y_j), (X_i,Y_j), and (X_j,Y_i) have the same distribution. ■

We remark that the results of Corollary 11 are already known. For example, parts 1 and 3 were proven by Gretton et al. (2007), while the wild bootstrap resampling scheme was studied by Chwialkowski et al. (2014). However, we highlight that, after our development, the new analysis is very concise.

4.3.3 Kernel Log-rank Test for the Two-sample Problem with Right-censored Data

In our last example, we explore right-censored data in the context of Survival Analysis. In the two-sample setting, we have i.i.d. triples $(X_i, \Delta_i, g_i)_{i=1}^n$, where X_i represents the observed time obtained as $X_i = \min(T_i, C_i)$. Here, T_i is the time of interest, such as the death time of patients in a clinical trial, and C_i is a nuisance time, for example, the time when the patient leaves the study or the study ends. Our primary focus is on T_i, but we cannot always observe this time due to the censoring time C_i. The censoring indicator, $\Delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$, takes the value 1 when we observe T_i, 0 otherwise. The variable $g_i \in \{0,1\}$ denotes the group label (e.g., one group of patients receives a new drug and the other a placebo).

Our goal is to compare the distributions F_0 and F_1, which generate the death times of interest T_i for each group (0 and 1). However, a standard two-sample test is not suitable here since we do not observe directly T_i, but rather X_i. Consequently, comparing F_0 and F_1 using observed data X_i without including the information of the censoring indicator will likely lead to false rejection of the null hypothesis. The challenge of incorporating censored information led to the development and study of the so-called weighted log-rank test, which is one of the pillars of survival analysis.

Within this context, we focus on the kernel log-rank test statistic Ψ_n, which is the kernelisation of the weighted log-rank test, and it is defined as

$$\Psi_n = \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}=1} S_n(\omega)^2,$$

where

$$S_n(\omega) = \sqrt{\frac{n}{n_0 n_1}} \int_0^\infty \omega(x) \frac{Y_0(x) Y_1(x)}{Y(x)} \left(\frac{dN_0(x)}{Y_0(x)} - \frac{dN_1(x)}{Y_1(x)} \right).$$

Here, $S_n(\omega)$ is the so-called weighted log-rank estimator with weight function ω, and $\mathcal{H}$ is a RKHS of functions $\mathbb{R} \to \mathbb{R}$. To understand $S_n(\omega)$, we need to introduce some standard notation used in Survival Analysis. For $\ell \in \{0,1\}$, n_ℓ is the sample size of group ℓ, $N_\ell(x) = \sum_{i=1}^n \Delta_i \mathbb{1}_{\{X_i \leq x, g_i = \ell\}}$ is a counting process (which counts the number of observed events in group ℓ up to time x), $Y_\ell(x) = \sum_{i=1}^n \mathbb{1}_{\{X_i \geq x, g_i = \ell\}}$ counts the number of patients in group ℓ who are still in the study by time x, $Y(x) = Y_0(x) + Y_1(x) \leq n$. Note then that integration in the definition of $S_n(\omega)$ is with respect to a counting process, i.e. it is just a sum.

We also consider a wild bootstrap version of $S_n(\omega)$ and Ψ_n. For that, consider W_i as Rademacher i.i.d. random variables (or any random variable with mean 0 and variance 1), and let $\ell \in \{0,1\}$ define the bootstrap counting processes $N_\ell^W(x) = \sum_{i=1}^n W_i \Delta_i \mathbb{1}_{\{X_i \leq x, g_i = \ell\}}$. Also, define S_n^W as S_n but replacing N_ℓ by N_ℓ^W, i.e.

$$S_n^W(\omega) = \sqrt{\frac{n}{n_0 n_1}} \int_0^\infty \omega(x) \frac{Y_0(x) Y_1(x)}{Y(x)} \left(\frac{dN_0^W(x)}{Y_0(x)} - \frac{dN_1^W(x)}{Y_1(x)} \right)$$

In this case, the kernelised version of $S_n^W(\omega)$ is $\Psi_n^W = \sup_{\omega \in \mathcal{H}: \|\omega\|_{\mathcal{H}}=1} S_n^W(\omega)^2$.

Note also that the processes $Y_\ell(t)$ depend on all the data points, so it is not possible to write $S_n(\omega)$ nor $S_n^W(\omega)$ as a sum of i.i.d. terms, and the usual way to deal with log-rank tests is by the use of the theory of stochastic integration of continuous-time martingales, especially for limiting results (see, for example, Fernández and Rivera (2021)). Despite this technical difficulty, our results can still be applied to understand the asymptotic distribution of Ψ_n and Ψ_n^W under the null and alternative hypotheses. This is possible because all the necessary components are already well-known results in Survival Analysis. As a result, we do not need to delve into martingale theory or related concepts.

For our analysis, let's define the covariance operators σ and σ^W in $\mathcal{H}$ by

$$\sigma(\omega,\omega') = \int_0^\infty \omega(x)\omega'(x)\frac{y_0(x)y_1(x)}{\rho_0 y_0(x)+\rho_1 y_1(x)}\frac{dF_0(x)}{1-F_0(x)},$$

and

$$\sigma^W(\omega,\omega') = \int \omega(x)\omega'(x)\frac{y_0(x)y_1(x)}{(\rho_0 y_0(x)+\rho_1 y_1(x))^2}\left(\rho_1 y_1(x)\frac{dF_0(x)}{1-F_0(x)} + \rho_0 y_0(x)\frac{dF_1(x)}{1-F_1(x)}\right),$$

where $y_\ell(x) = (1-G_\ell(x))(1-F_\ell(x))$ for $\ell \in \{0,1\}$. Note that under the null hypothesis $\sigma^W = \sigma$. We remark that these covariance operators are well known in the theory of log-rank tests in survival analysis, as they characterise joint convergence of log-rank statistics with different weight functions.

During our analysis, we will make use of the following identities $\mathbb{E}(Y_0(x)) = n_0(1-F_0(x))(1-G_0(x))$ and $\mathbb{E}(Y_1(x)) = n_1(1-F_1(x))(1-G_1(x))$. Also, we will use the fact that the law of large numbers applies, i.e., $\lim_{n\to\infty}\frac{1}{n_\ell}\int_0^\infty f(x)dN_\ell(x) = \int_0^\infty f(x)(1-G_\ell)dF_\ell(x)\ a.s.$ Finally, for theoretical and practical reasons it is necessary to assume that $n_0/n \to \rho_0 > 0$ and $n_1/n \to \rho_1 > 0$, i.e., that the proportions of both groups do not vanish.

Corollary 12 *Suppose that $\int_{\mathcal{X}} K(x,x)(dF_0(x)+dF_1(x)) < \infty$. Then*

1. *Under the null hypothesis (if $F_0 = F_1$). Then $\Psi_n \overset{\mathcal{D}}{\to} \Psi \sim \chi^2(\sigma)$.*
2. *Under null or alternative, $\Psi_n^W \overset{\mathcal{D}_\mathcal{D}}{\to} \Psi^W \sim \chi^2(\sigma^W)$. Moreover, under the null hypothesis it holds that Ψ^W and Ψ have the same distribution.*
3. *Under null or alternative, there exists a constant $c_* \geq 0$ such that $\frac{1}{n}\Psi_n \overset{a.s.}{\to} c_*^2$.*

Proof We shall apply Theorem 1 in conjunction with Proposition 4 and Proposition 6.

Item 1. Under the null hypothesis that $F_0 = F_1$, we verify that S_n satisfies Conditions G_0 to G_2 in order to apply Theorem 1. For Condition G_0, we have $S_n(\omega) \overset{\mathcal{D}}{\to} N(0,\sigma(\omega,\omega))$, which is the standard convergence to a normal limit of the log-rank estimator (see, for example, Lemma 1 of Brendel et al. (2014)). The other two conditions follow from Proposition 4, by setting Q as the identity and μ and ν as F_0, and by noting that

$$\begin{aligned}\mathbb{E}(S_n(\omega)^2) &= \mathbb{E}\left(\frac{n}{n_0 n_1}\int_0^\infty \omega(x)^2\frac{Y_0(x)Y_1(x)}{Y(x)}\frac{dF_0(x)}{1-F_0(x)}\right)\\ &\leq \frac{n}{n_0 n_1}\int_0^\infty \omega(x)^2\mathbb{E}\left(Y_0(x)\right)\frac{dF_0(x)}{1-F_0(x)}\\ &= \frac{n}{n_1}\int_0^\infty \omega(x)^2(1-G_0(x))dF_0(x) \leq C\int_0^\infty \omega(x)^2 dF_0(x) \end{aligned} \qquad (25)$$

where the first equality is from Lemma 4.1.2 of Gill (1980), then we use that for every $x \geq 0$ it holds $Y_0(x) + Y_1(x) = Y(x)$ thus $Y_i(x)/Y(x) \leq 1$, that $\mathbb{E}(Y_0) = n_0(1 - G_0)(1 - F_0)$, and that n/n_1 converges to a constant, so it is bounded for large enough n. Finally, by the assumptions in the statement, Eq. (8) holds and thus the conclusion of Proposition 4 follows.

Item 2. We can assume that either the null hypothesis or the alternative hypothesis holds. We verify the conditions of Theorem 1 for Ψ_n^W conditioned on $\mathcal{D}$. Condition G_0 follows from the proofs of Theorems 5 and 6 of Ditzhaus and Pauly (2019) where it is shown that $S_n^W(\omega) \overset{\mathcal{D}_{\mathcal{D}}}{\to} N(0, \sigma^W(\omega, \omega))$.

To verify Conditions G_1 and G_2 we use Proposition 6. For that we set $\mathcal{F}'$ as the sigma-algebra generated by the data $D = (X_i, \Delta_i, g_i)_{i\geq 1}$. Then by Theorems 5 and 6 of Ditzhaus and Pauly (2019) (or rather the proof of) we have

$$\mathbb{E}_D(S_n^W(\omega)^2) = \frac{n}{n_0 n_1} \int_0^\infty \omega(x)^2 \left(\frac{Y_0(x)Y_1(x)}{Y(x)}\right)^2 \left(\frac{dN_0(x)}{Y_0(x)^2} + \frac{dN_1(x)}{Y_1(x)^2}\right). \tag{26}$$

and thus by the definition of Y_0, Y_1 and Y, as well as the relation between n_0, n_1, and n, from Eq. (26) we deduce that the following inequality holds for large enough n:

$$\mathbb{E}_D(S_n^W(\omega)^2) \leq C \int_0^\infty \omega(x)^2 \left(\frac{dN_0(x)}{n_0} + \frac{dN_1(x)}{n_1}\right), \tag{27}$$

for some $C > 0$.

Now, for our application of Proposition 6 we set $\nu_n(dx) = \left(\frac{dN_0(x)}{n_0} + \frac{dN_1(x)}{n_1}\right)$, $\mu(dx) = \nu(dx) = ((1 - G_0(x)dF_0(x) + (1 - G_1(x))dF_1(x))$, and the operator $Q : \mathcal{H} \to L_2(\mu)$ as the identity matrix. Recall that $dN_\ell(x)/n_\ell$ satisfies the law of large numbers; thus $\int f\nu_n \to \int f\nu$ for every f such that the latter integral exists. Finally, note that condition a) of Proposition 6 is satisfied, since the first term on the left-hand side term of Eq. (11) is trivially finite since $\mu = \nu$ and Q is the identity, the other term is $\int_0^\infty K(x, x)\mu(dx)$ which is finite, as per the assumptions stated in the statement.

We conclude that all the conditions for invoking Proposition 6 (with condition a) are satisfied, and thus the convergence of Ψ_n^W to $\Psi^W \sim \chi^2(\sigma^W)$ follows. Furthermore, note that under the null hypothesis the covariance σ^W is equal to σ, showing that Ψ^W and Ψ have the same distribution when $F_0 = F_1$.

Item 3. We apply Theorem 3 to $\sqrt{\frac{n}{n_0 n_1}} S_n(\omega)$. First, we have (Fleming and Harrington, 1991, Section 7.3) that

$$\sqrt{\frac{n}{n_0 n_1}} S_n(\omega) \overset{a.s.}{\to} c(\omega) = \int_0^\infty \omega(x) \frac{y_0(x)y_1(x)}{\rho_0 y_0(x) + \rho_1 y_1(x)} \left(\frac{dF_0(x)}{1 - F_0(x)} - \frac{dF_1(x)}{1 - F_1(x)}\right),$$

where $y_\ell(x) = (1 - G_\ell(x))(1 - F_\ell(x))$ for $\ell \in \{0, 1\}$. It only remains to verify Condition G_3 for which we invoke Proposition 6 with $\mathcal{F}' = \mathcal{F}$ (i.e., we do not take expectation at all). First, we claim that for n large enough there exists $C > 0$ such that

$$\left|\sqrt{n/(n_0 n_1)} S_n(\omega)\right|^2 \leq C \int_0^\infty \omega(x)^2 \left(\frac{dN_0(x)}{n_0} + \frac{dN_1(x)}{n_1}\right). \tag{28}$$

so we set $\nu_n(dx) = \frac{dN_0(x)}{n_0} + \frac{dN_1(x)}{n_1}$ and Q as the identity in the application of Proposition 6. Equation (28) follows by using that

$$\left|\sqrt{\frac{n}{n_0 n_1}} S_n(\omega)\right| \leq \frac{n}{n_0 n_1} \int_0^\infty |\omega(x)| \frac{Y_0(x)Y_1(x)}{Y(x)} \nu_n(dx),$$

and by noting that $Y_0(x)/n_0 = \sum_{i=1}^n \mathbb{1}_{\{X_i \geq x, g_i = 0\}}/n_0 \leq 1$ and $Y_1(x)/Y(x) = Y_1(x)/(Y_0(x) + Y_1(x)) \leq 1$, we have $\frac{n}{n_0 n_1} \int_0^\infty |\omega(x)| \frac{Y_0(x)Y_1(x)}{Y(x)} \nu_n(dx) \leq \frac{n}{n_1} \int_0^\infty |\omega(x)| \nu_n(dx)$. Then, since n/n_1 converges to a non-zero quantity, we conclude that for large enough n,

$$\left|\sqrt{\frac{n}{n_0 n_1}} S_n(\omega)\right| \leq \frac{C'}{n} \int_0^\infty |\omega(x)| \nu_n(dx),$$

for some constant $C' > 0$. By squaring both sides of the previous equation, and by using that $\int_0^\infty dN_\ell(x)/n_\ell \leq 1$ together with Jensen's inequality we get Eq. (28).

Now, the random measure $\nu_n(dx) = \left(\frac{dN_0(x)}{n_0} + \frac{dN_1(x)}{n_1}\right)$ satisfies the law of large numbers; indeed, $\int f(x)\nu_n(dx)$ converges a.s. to $\int_0^\infty f(x)\left((1 - G_0(x)dF_0(x) + (1 - G_1(x))dF_1(x)\right)$ if the latter integral exists. We now set $\mu(dx) = \nu(dx) = (1 - G_0(x))dF_0(x) + (1 - G_1(x))dF_1(x)$ in our application of Proposition 6. Finally, to use Proposition 6 we verify item a) of its statement, and since Q is the identity, we just need to verify that $\int_0^\infty K(x,x)\mu(dx) < \infty$, but this holds by the assumptions stated. ■

5. Conclusion

We have introduced new tools to analyse the asymptotic behaviour of kernel-based tests. These tools give us necessary and sufficient conditions to extend asymptotic results for standard weighted test-statistics to kernelised test-statistics, making the analysis of the kernel tests much simpler, cleaner, and shorter. The latter is a direct consequence of the fact that our analysis is carried out directly on random functionals on the Hilbert space, avoiding the intricate expansions that usually appear in the literature of kernel tests. Also, we provide additional sufficient conditions that can be easy to apply in practice, e.g. Propositions 4 and 6, replacing the algebraic nature of Conditions G_2 and G_3 by integrability conditions.

To exemplify the wide range of application of our results and to offer readily applicable results for practitioners, we analyse two general classes of kernel test: when the linear test statistic is a sum of i.i.d. random variables and when the test-statistic is a U-statistic of order $r \geq 2$ (Theorems 7 and 8). These classes are very general and contain important test statistics, such as the kernelised Stein discrepancy, the MMD for goodness-of-fit problems, and the HSIC measure. For a concrete example, we present a simple analysis of an independence test based on the HSIC measure.

Beyond the analysis of general classes, we additionally study two specific kernelised testing procedures: the kernel log-rank test for the two-sample problem with right-censored data, and a new kernel test for conditional independence (testing whether X and Y are independent given Z), showing that our techniques are effective in obtaining concise and clean analysis of the testing procedures. The novel test for conditional independence is based on the recently introduced generalised covariance measure. For this test, we present an asymptotic analysis using our developments and a few experimental results to show how kernelisation is able to make simple tests more robust to different alternatives, having better performance than tests based on the generalised covariance measure and its weighted generalisation. We leave for future work a more detailed study of this new testing procedure, especially in settings where the dimensions of X and Y are greater than 1.

Acknowledgments

The authors would like to thank Rolando Rebolledo for his comments, suggestions, and stimulating discussions. Tamara Fernández the Data Observatory Foundation - ANID Technology Center No DO210001, ANID FONDECYT grant No 11221143, and MATH-AMSUD grant No 220041. Nicolás Rivera was supported by ANID FONDECYT grant No 3210805 and ANID SIA grant No 85220033, and ANID EXPLORACION grant No 13220168.

Appendix A. Auxiliary Results

We will now present essential supporting statements that will be used in the proofs of our main results in Appendix B.

A.1 Auxiliary Results for Theorems 1 to 3

The next result allows us to deduce the joint convergence of $S_n(\omega_1), \ldots, S_n(\omega_n)$ from Condition G_0. This result is used in the proof of Theorem 1.

Lemma 13 *Suppose that Condition G_0 holds. Then the function $\sigma : \omega \to \sigma(\omega, \omega)$ can be extended to unique symmetric bilinear form $\sigma : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ such that for any $m \in \mathbb{N}$, the linear test-statistic $(S_n(\omega_1), \ldots, S_n(\omega_m))$ converges in distribution to a multivariate normal random variable with mean 0 and covariance matrix given by $\Sigma_{ij} = \sigma(\omega_i, \omega_j)$.*

The following result establishes important properties of the operator T_σ derived from the bilinear form σ defined after Condition G_0.

Proposition 14 *Let $\mathcal{H}$ be a separable Hilbert Space, and let $\sigma : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be a continuous bilinear form satisfying Condition G_1. Then, it exists a unique self-adjoint trace-class linear operator T_σ such that $\langle T_\sigma f, \omega \rangle_{\mathcal{H}} = \sigma(f, \omega)$ for any $f, \omega \in \mathcal{H}$.*

The next result is used in the proofs of Theorems 1 and 2.

Lemma 15 *Let $(\lambda_i)_{i \geq 1}$ be a sequence of non-negative real numbers and let $(Z_i)_{i \geq 1}$ be a collection of i.i.d. standard normal random variables. Then $\sum_{i \geq 1} \lambda_i < \infty$ if and only if $\sum_{i \geq 1} \lambda_i Z_i^2$ converges almost surely to a random variable.*

A.2 Auxiliary Results for Theorems 7 and 8

The following results are an application of Linderberg's Central Limit theorem to settings related to wild bootstrap estimators. They feature in the proofs of Theorems 7 and 8

Lemma 16 *Consider a triangular array of real random variables $D = (Y_{in} : i \in [n], n \geq 1)$ such that $(Y_{in}^2 : i \in [n], n \geq 1)$ is uniformly integrable and $\frac{1}{n}\sum_{i=1}^n Y_{in}^2 \overset{a.s.}{\to} c^2$ for some $c > 0$. Let (W_i) be i.i.d. Rademacher random variables. Then, we have*

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i Y_{in} \overset{\mathcal{D}_D}{\to} N(0, c^2).$$

The next result establishes the asymptotic normality of the wild bootstrap version of a U-statistic, conditioned on the data.

Corollary 17 *Let $(U\omega)$ be a U-statistic kernel of degree $r \geq 1$ on data $D = (X_i)_{i \geq 1}$, and assume that $\mathbb{E}((U\omega)(X_1, \dots, X_r)^2) < \infty$. Let $(W_i)_{i=1}^n$ be i.i.d. Rademacher random variables, and independent of the data D, then*

$$\frac{\sqrt{n}}{r\binom{n}{r}} \sum_{A \in \binom{[n]}{r}} (W_{A_1} + W_{A_2} + \ldots + W_{A_r})\,(U\omega)\,(X_{A_1}, X_{A_2}, \dots, X_{A_r}) \overset{\mathcal{D}_D}{\to} N(0, \sigma^W(\omega, \omega)),$$

where $\sigma^W(\omega, \omega) = \mathbb{E}(\omega(X_1, X_2, \dots, X_r)\omega(X_1, X_2', \dots, X_r'))$, and $X_2', \dots X_r'$ are independent copies of $X_2, \dots, X_r$, and are independent of everything else.

A.3 Proofs of Auxiliary results

Proof of Lemma 13. We start by claiming that if $\omega_1, \omega_2, \omega_3 \in \mathcal{H}$, then the random vector $\mathbf{S}_n = (S_n(\omega_1), S_n(\omega_2), S_n(\omega_3))$ converges in distribution. To see this, observe that $S_n(\omega_i)$ converges in distribution for each $i \in \{1, 2, 3\}$ and thus $\mathbf{S}_n$ is a tight random variable in $\mathbb{R}^3$, therefore a subsequence of it converges in distribution to a random vector $\boldsymbol{Z}$. We shall verify that the entire sequence $\mathbf{S}_n$ converges in distribution to $\boldsymbol{Z}$. For that, consider $\boldsymbol{a} = (a_1, a_2, a_3) \in \mathbb{R}^3$, and define $\omega_{\boldsymbol{a}} = a_1\omega_1 + a_2\omega_2 + a_3\omega_3 \in \mathcal{H}$. Then, $\boldsymbol{a}^\top \mathbf{S}_n = S_n(\omega_{\boldsymbol{a}})$ converges in distribution to a Normal random variable with mean 0 and variance $\sigma(\omega_{\boldsymbol{a}}, \omega_{\boldsymbol{a}})$ due to Condition G_0. Note that such a limit coincides with $\boldsymbol{a}^\top \boldsymbol{Z}$, and thus $\mathbf{S}_n$ converges to $\boldsymbol{Z}$, proving the claim. Note that we also conclude that $\sigma(\omega_{\boldsymbol{a}}, \omega_{\boldsymbol{a}}) = \mathbb{E}((\boldsymbol{a}^\top \boldsymbol{Z})^2)$ for any $\boldsymbol{a} \in \mathbb{R}^3$.

We continue with the proof of the lemma. We first extend σ to $\sigma : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ by:

$$2\sigma(\omega, \omega') = \sigma(\omega + \omega', \omega + \omega') - \sigma(\omega, \omega) - \sigma(\omega', \omega').$$

Let $\boldsymbol{b} \in \mathbb{R}^3$, then by the definition of the extension of σ, it holds

$$2\sigma(\omega_{\boldsymbol{a}}, \omega_{\boldsymbol{b}}) = \mathbb{E}\left(\left(\boldsymbol{a}^\top \boldsymbol{Z} + \boldsymbol{b}^\top \boldsymbol{Z}\right)^2\right) - \mathbb{E}\left(\left(\boldsymbol{a}^\top \boldsymbol{Z}\right)^2\right) - \mathbb{E}\left(\left(\boldsymbol{b}^\top \boldsymbol{Z}\right)^2\right) = 2\mathbb{E}\left((\boldsymbol{a}^\top \boldsymbol{Z})(\boldsymbol{b}^\top \boldsymbol{Z})\right),$$

showing that σ is bilinear in the subspace generated by $\omega_1, \omega_2, \omega_3$ as $(\boldsymbol{a}, \boldsymbol{b}) \to \mathbb{E}((\boldsymbol{a}^\top \boldsymbol{Z})(\boldsymbol{b}^\top \boldsymbol{Z}))$ is a bilinear form, and since the ω_i's are arbitrary, σ is bilinear in $\mathcal{H}$. Recall that a symmetric bilinear form is characterised by its diagonal values; therefore, the symmetric bilinear extension of σ is unique.

Finally, given $\omega_1, \dots, \omega_m \in \mathcal{H}$ and $a_1, \dots, a_m \in \mathbb{R}$, we have, by hypothesis, that $S_n(\sum_{i=1}^m a_i\omega_i)$ converges in distribution to a Normal random variable. Since σ is bilinear, the variance of the previous limit is

$$\sigma\left(\sum_{i=1}^m a_i\omega_i, \sum_{i=1}^m a_i\omega_i\right) = \sum_{i=1}^m \sum_{j=1}^m a_i a_j \sigma(\omega_i, \omega_j),$$

showing that $(S_n(\omega_1), \dots, S_n(\omega_m))$ converges jointly to a multivariate normal with mean 0 and covariance matrix $\Sigma_{ij} = \sigma(\omega_i, \omega_j)$. ■

Proof of Proposition 14. Let's show that T_σ exists. For any $f \in \mathcal{H}$, define $A_f : \mathcal{H} \to \mathbb{R}$ by $A_f\omega = \sigma(f, \omega)$ for all $\omega \in \mathcal{H}$. Clearly A_f is linear since σ is bilinear. Moreover, by the continuity of σ, we have that $\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} \sigma(\omega, \omega) < \infty$, which implies that A_f is bounded by an application of the Cauchy-Schwarz inequality, indeed:

$$\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} |A_f\omega|^2 = \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} |\sigma(f, \omega)|^2 \leq \sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1} \sigma(\omega, \omega)\sigma(f, f) < \infty.$$

Then, by Riesz's representation theorem, for all $f \in \mathcal{H}$ there exists a unique element $\xi_f \in \mathcal{H}$ such that for all $\omega \in \mathcal{H}$, $A_f\omega = \langle \xi_f, \omega\rangle_{\mathcal{H}}$. We define $T_\sigma : \mathcal{H} \to \mathcal{H}$ as $T_\sigma f = \xi_f$, which we can easily verify is linear and bounded. Also, note that

$$\langle T_\sigma f, \omega\rangle_{\mathcal{H}} = \langle \xi_f, \omega\rangle_{\mathcal{H}} = A_f(\omega) = \sigma(f, \omega). \tag{29}$$

The uniqueness of T_σ follows from the linearity of σ. To see that T_σ is self-adjoint note that

$$\langle T_\sigma f, \omega\rangle_{\mathcal{H}} = \sigma(f, \omega) = \sigma(\omega, f) = \langle T_\sigma \omega, f\rangle_{\mathcal{H}} = \langle f, T_\sigma\omega\rangle_{\mathcal{H}}.$$

Finally, we claim that T_σ is trace class. For that consider an orthonormal basis $(\phi_i)_{i\geq 1}$ of $\mathcal{H}$, and observe that $\text{Trace}(T_\sigma) = \sum_{i=1}^\infty \langle \phi_i, T_\sigma\phi_i\rangle_{\mathcal{H}} = \sum_{i=1}^\infty \sigma(\phi_i, \phi_i) < \infty$, where the inequality is due to Condition G_1. ■

Proof of Lemma 15. ($\Longrightarrow$) We will prove that $\sum_{i=1} \lambda_i < \infty$ implies that the series $\sum_{i=1}^n \lambda_i Z_i^2$ converges almost surely. To do this, we verify the two conditions of Kolmogorov's two-series theorem. We first need to verify that $\sum_{i=1}^\infty \mathbb{E}(\lambda_i Z_i^2) < \infty$, but this follows immediately since $\mathbb{E}(Z_i^2) = 1$. We also need to verify that $\sum_{i=1}^\infty \mathbb{V}ar(\lambda_i Z_i^2) < \infty$, which holds since $\mathbb{V}ar(Z_i^2) = 2$, and since $\sum_{i=1}^\infty \lambda_i < \infty$, it implies that $\sum_{i=1}^\infty \lambda_i^2 < \infty$.

($\Longleftarrow$) We proceed to prove that if $\sum_{i=1}^\infty \lambda_i Z_i^2$ converges almost surely, then $\sum_{i=1}^\infty \lambda_i < \infty$. Note that since $\sum_{i=1}^\infty \lambda_i Z_i^2$ converges almost surely, Kolmogorov's three series theorem yields that for any $A > 0$ it holds that

$$\text{i)} \quad \sum_{i=1}^\infty \mathbb{P}(\lambda_i Z_i^2 \geq A) < \infty, \qquad \text{and} \qquad \text{ii)} \quad \sum_{i=1}^\infty \lambda_i \mathbb{E}\left(Z_i^2 \mathbb{1}_{\{\lambda_i Z_i^2 \leq A\}}\right) < \infty.$$

We use i) to deduce that the sequence $(\lambda_i)_{i=1}^\infty$ is bounded: Consider $A = 1$, and suppose, for a contradiction, that there exists a subsequence $(\lambda_{n_k})_{k=1}^\infty$ such that $\lambda_{n_k} \to \infty$ as $k \to \infty$.

Then, there exists $N \in \mathbb{N}$ large enough such that $\mathbb{P}(\lambda_{n_k} Z_{n_k}^2 \geq 1) \geq 1/2$ for all $k \geq N$, and thus $\sum_{k=1}^{\infty} \mathbb{P}(\lambda_{n_k} Z_{n_k}^2 \geq 1) \to \infty$ which contradicts i). We conclude that $\sup_k \lambda_k$ is bounded by some constant.

Let $C = \mathbb{E}\left(Z_i^2 \mathbb{1}_{\{(\sup_k \lambda_k) Z_i^2 \leq 1\}}\right) > 0$ which is independent of i, then by ii) with $A = 1$, we get

$$\sum_{i=1}^{\infty} \lambda_i = \frac{1}{C} \sum_{i=1}^{\infty} \lambda_i \mathbb{E}\left(Z_i^2 \mathbb{1}_{\{(\sup_k \lambda_k) Z_i^2 \leq 1\}}\right) \leq \frac{1}{C} \sum_{i=1}^{\infty} \lambda_i \mathbb{E}\left(Z_i^2 \mathbb{1}_{\{\lambda_i Z_i^2 \leq 1\}}\right) < \infty.$$

■

Proof of Lemma 16. Since $W_i^2 = 1$, the result is equivalent to prove that $\sum_{i=1}^m W_i Y_{im}/s_m$ converges $\sigma(D)$-stably to a standard normal when m tends to infinity, where $s_m^2 = \sum_{i=1}^m Y_{im}^2$. Denote $Z_{im} = \frac{W_i Y_{im}}{s_m}$ which has mean 0 conditioned on D. By Theorem 6.1 of Häusler and Luschgy (2015), which extends Linderberg's CLT to stable convergence, the result follows by proving that

$$\mathbb{E}\left(\sum_{i=1}^m Z_{im}^2 \middle| D\right) \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \mathbb{E}\left(\sum_{i=1}^m Z_{im}^2 \mathbb{1}_{\{Z_{im}^2 \geq \varepsilon^2\}} \middle| D\right) \xrightarrow{\mathbb{P}} 0 \quad \text{for all } \varepsilon > 0.$$

The first limit is trivial since $\sum_{i=1}^m Z_{im}^2 = 1$. For the second one, let $\varepsilon > 0$, and note that

$$\mathbb{E}\left(\sum_{i=1}^m Z_{im}^2 \mathbb{1}_{\{Z_{im}^2 \geq \varepsilon^2\}} \middle| D\right) = \frac{\sum_{i=1}^m Y_{im}^2 \mathbb{1}_{\{Y_{im}^2 \geq \varepsilon s_m^2\}}}{s_m^2}.$$

Now, recall that $s_m^2/m \to c^2$, then by considering the cases $s_m^2/m \geq c^2/2$ and $s_m^2/m < c^2/2$, we get

$$\begin{aligned}
\frac{\sum_{i=1}^m Y_{im}^2 \mathbb{1}_{\{Y_{im}^2 \geq \varepsilon s_m^2\}}}{s_m^2} &= \sum_{i=1}^m \frac{Y_{im}^2}{s_m^2} \mathbb{1}_{\{Y_{im}^2 \geq s_m^2 \varepsilon^2\}} \mathbb{1}_{\{s_m^2/m \geq c^2/2\}} + \sum_{i=1}^m \frac{Y_{im}^2}{s_m^2} \mathbb{1}_{\{Y_{im}^2 \geq s_m^2 \varepsilon^2\}} \mathbb{1}_{\{s_m^2/m < c^2/2\}} \\
&\leq \sum_{i=1}^m 2 \frac{Y_{im}^2}{c^2 m} \mathbb{1}_{\{Y_{im}^2 \geq m c^2 \varepsilon^2/2\}} + \mathbb{1}_{\{s_m^2/m < c^2/2\}} \sum_{i=1}^m \frac{Y_{im}^2}{s_m^2} \\
&= \frac{2}{c^2} \frac{\sum_{i=1}^m Y_{im}^2 \mathbb{1}_{\{Y_{im}^2 \geq m c \varepsilon^2/2\}}}{m} + \mathbb{1}_{\{s_m^2/m < c^2/2\}} := Q_m.
\end{aligned}$$

Then, it is enough to show that $Q_m \xrightarrow{\mathbb{P}} 0$. We prove such a limit in expectation, and thus in probability. Clearly $\mathbb{P}(s_m^2/m < c^2/2) \to 0$ since $s_m^2/m \to c^2$. For the other term, recall that D is a uniformly integrable set of random variables, for every $\delta > 0$ there exists $K > 0$ such that $\mathbb{E}(Y_{im}^2 \mathbb{1}_{\{Y_{im}^2 > K\}}) \leq c^2\delta/2$. Therefore, for large enough m, we have $\mathbb{E}(Q_m) \leq \delta$, and since δ is arbitrary, the result holds. ■

Proof of Corollary 17. For $A \in \binom{[n]}{r}$, denote $\alpha_A = (U\omega)(\boldsymbol{X}_A)$. For $i \in [n]$, define $I_i = \{A \in \binom{[n]}{r} : i \in A\}$, and define $Y_{in} = \frac{1}{\binom{n-1}{r-1}} \sum_{A \in I_i} \alpha_A$. Then

$$\frac{\sqrt{n}}{r\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \mathbf{W}_A \cdot (U\omega)(\boldsymbol{X}_A) = \frac{\sqrt{n}}{r\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \sum_{k=1}^{r} W_{A_k}\alpha_A = \frac{\sqrt{n}}{r\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \sum_{k=1}^{r}\sum_{i=1}^{n} W_i \mathbb{1}_{\{A_k=i\}}\alpha_A$$

$$= \frac{\sqrt{n}}{r\binom{n}{r}} \sum_{i=1}^{n} W_i \sum_{A\in\binom{[n]}{r}} \sum_{k=1}^{r} \mathbb{1}_{\{A_k=i\}}\alpha_A = \frac{\sqrt{n}}{r\binom{n}{r}} \sum_{i=1}^{n} W_i \sum_{A\in\binom{[n]}{r}} \mathbb{1}_{\{i\in A\}}\alpha_A = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i Y_{in}$$

The convergence in distribution follows from Lemma 16 which requires the following properties:

1. $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n Y_{in}^2 \to \sigma^W(\omega,\omega)$ almost surely.
2. $(Y_{in}^2 : i \in [n], n \geq 1)$ is uniformly integrable.

We begin by verifying property 1. Start by noting that

$$\frac{1}{n}\sum_{i=1}^{n} Y_{in}^2 = \frac{1}{n\binom{n-1}{r-1}^2} \sum_{i=1}^{n} \sum_{A\in I_i} \sum_{B \in I_i} \alpha_A\alpha_B = \frac{1}{n\binom{n-1}{r-1}^2} \sum_{A\in\binom{[n]}{r}} \sum_{B\in\binom{[n]}{r}} |A\cap B|\alpha_A\alpha_B.$$

Let $J_k \subseteq \binom{[n]}{r} \times \binom{[n]}{r}$ be such that $(A,B) \in J_k$ if and only if $|A\cap B| = k$. Then

$$\frac{1}{n}\sum_{i=1} Y_{in}^2 = \frac{1}{n\binom{n-1}{r-1}^2} \sum_{(A,B)\in J_1} \alpha_A\alpha_B + \frac{1}{n\binom{n-1}{r-1}^2} \sum_{k=2}^{r} k \sum_{(A,B)\in J_k} \alpha_A\alpha_B.$$

We shall verify that the second term converges to 0 (assuming that the degree r is greater than or equal to 2, otherwise, there is nothing to prove). For $k \geq 2$, we have

$$\frac{1}{n\binom{n-1}{r-1}^2} \left| \sum_{(A,B)\in J_k} \alpha_A\alpha_B \right| \leq \frac{1}{n\binom{n-1}{r-1}^2} \sum_{(A,B)\in J_k} \frac{\alpha_A^2 + \alpha_B^2}{2} = \frac{1}{n\binom{n-1}{r-1}^2} \sum_{(A,B)\in J_k} \alpha_A^2$$

$$= \frac{1}{n\binom{n-1}{r-1}^2} \sum_{A\in\binom{[n]}{r}} \alpha_A^2 \binom{r}{k}\binom{n-r}{r-k} \leq \frac{c}{n^{-k+1}\binom{n}{r}} \sum_{A\in\binom{[n]}{r}} \alpha_A^2$$

where the last inequality follows from the standard bound $(n/\ell)^\ell \leq \binom{n}{\ell} \leq (en/\ell)^\ell$ for any $\ell \in [n]$, and note that $c > 0$ does not depend on n.

By the law of large numbers for U-statistics (Theorem A in Section 5.4 of Serfling (1980)), we have that $\binom{n}{r}^{-1} \sum_{A\in\binom{[n]}{r}} \alpha_A^2 \overset{a.s.}{\to} \mathbb{E}((U\omega)(X_1,\ldots,X_r)^2) < \infty$, and thus we conclude that for $k \geq 2$, $\lim_{n\to\infty} \frac{1}{n\binom{n-1}{r-1}^2} \left|\sum_{(A,B)\in J_k} \alpha_A\alpha_B\right| = 0$, almost surely, and thus

$$\frac{1}{n}\sum_{i=1}^{n} Y_{in}^2 = \frac{1}{n\binom{n-1}{r-1}^2} \left(\sum_{(A,B)\in J_1} \alpha_A\alpha_B \right) + o(1).$$

To deal with the last summation, we partition J_1. Let $C \in \binom{[n]}{2r-1}$ and define $J_1^C = \{(A,B) \in J_1 : A \cup B = C\}$. Then

$$\frac{1}{n\binom{n-1}{r-1}^2} \sum_{(A,B)\in J_1} \alpha_A \alpha_B = \frac{1}{n\binom{n-1}{r-1}^2} \sum_{C \in \binom{[n]}{2r-1}} \sum_{(A,B)\in J_1^C} \alpha_A \alpha_B = \frac{1}{n\binom{n-1}{r-1}^2} \sum_{C \in \binom{[n]}{2r-1}} \beta_C,$$

where β_C is defined as $\sum_{(A,B)\in J_1^C} \alpha_A \alpha_B$. Note $\binom{n}{2r-1}^{-1} \sum_{C \in \binom{[n]}{2r-1}} \beta_C \to \mathbb{E}(\beta_{[2r-1]})$ by the Law of Large Numbers for U-statistics. Recall that $[2r-1] = \{1, \ldots, 2r-1\}$ and note that $\mathbb{E}(\beta_C)$ is not dependent on $C \in \binom{[n]}{2r-1}$. By noting that $\lim_{n\to\infty} \frac{\binom{n}{2r-1}}{n\binom{n-1}{r-1}^2} = \frac{1}{(2r-1)}\binom{2r-2}{r-1}^{-1}$ and that $\mathbb{E}(\beta_{[2r-1]}) = (2r-1)\binom{2r-2}{r-1}\sigma^W(\omega,\omega)$, we conclude

$$\frac{1}{n\binom{n-1}{r-1}^2} \sum_{C \in \binom{[n]}{2r-1}} \beta_C \overset{a.s.}{\to} \sigma^W(\omega,\omega).$$

For the second property, first note that the variables $(Y_{in})_{i=1}^n$ have the same distribution for a fixed value n, thus it is enough to verify that $(Y_{1n}^2 : n \geq 1)$ is uniformly integrable. We will prove that $\lim_{n\to\infty} Y_{1n}$ exists in L_2, which implies that $(Y_{1n} : n \geq 1)$ is uniformly integrable. We claim that as $n \to \infty$, $\mathbb{V}ar\left(Y_{1n} - \mathbb{E}(\alpha_{[k]}|X_1)\right) \to 0$, i.e. Y_{1n} converges in L_2 to $\mathbb{E}(\alpha_{[k]}|X_1)$. To prove this, recall that by definition $Y_{1n} = \binom{n-1}{r-1}^{-1} \sum_{A \in I_1} \alpha_A$ is a U-statistic of order $r-1$ on variables $X_2, \ldots, X_n$ conditioned on X_1. Then $\mathbb{E}(Y_{1n}|X_1) = \mathbb{E}(\alpha_{[r]}|X_1)$ and $\mathbb{V}ar(Y_{1n}|X_1) \leq \frac{r-1}{n-1}\mathbb{V}ar(\alpha_{[r]}|X_1)$ by Lemma A of Section 5.2.1 of Serfling (1980). Now, the claim follows by noting that $\mathbb{E}(\mathbb{V}ar(Y_{1n}|X_1)) \to 0$ since $\mathbb{E}(\mathbb{V}ar(\alpha_{[r]}|X_1)) \leq \mathbb{V}ar(\alpha_{[r]}) < \infty$. ■

Appendix B. Proofs of Convergence Results

In this section, we prove Theorems 1 to 3, as well as Propositions 4 and 6 that give alternative conditions to apply the convergence results.

Proof of Theorem 1. By Proposition 14 we have that $T_\sigma : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ defined in Eq. (3) is self-adjoint and trace class in $\mathcal{H}$. Then, by the spectral theorem, there exists an orthonormal basis $(\phi_i)_{i\geq 1}$ of $\mathcal{H}$ such that $T_\sigma \phi_i = \lambda_i \phi_i$ for all $i \geq 0$. Recall that the basis is countable since $\mathcal{H}$ is separable, and that the eigenvalues are all non-negative since $\lambda_i = \langle T_\sigma \phi_i, \phi_i \rangle_{\mathcal{H}} = \sigma(\phi_i, \phi_i) \geq 0$, by Proposition 14 (recall that $\sigma(\phi_i,\phi_i)$ is a variance, so it is non-negative). Also, note that $\sum_{i=1}^\infty \lambda_i < \infty$ as T_σ is a trace class operator.

Note that $S_n(\omega) = \sum_{i=1}^\infty \langle \phi_i, \omega \rangle_{\mathcal{H}} S_n(\phi_i)$, and recall that the functional S is defined by $S(\omega) = \sum_{i=1}^\infty \langle \phi_i, \omega \rangle_{\mathcal{H}} \sqrt{\lambda_i} Z_i$, where Z_i are i.i.d. standard normal random variables. Observe that S is well-defined, since $\sum_{i=1}^\infty \lambda_i Z_i^2$ converges a.s. due to Lemma 15 (in Appendix A) since we have $\sum_{i=1}^\infty \lambda_i < \infty$. Our proof also requires the definition of a partial version of S_n and S, given by

$$S_n^m := \sum_{i=1}^m \langle \phi_i, \cdot \rangle_{\mathcal{H}} S_n(\phi_i) \quad \text{and} \quad S^m(\omega) = \sum_{i=1}^m \langle \phi_i, \omega \rangle_{\mathcal{H}} \sqrt{\lambda_i} Z_i, \quad m \geq 1$$

We will show that $S_n \overset{\mathcal{D}}{\to} S$ via an application of Theorem 3.2 of Billingsley (2013), which requires the following properties:

i. $S_n^m \overset{\mathcal{D}}{\to} S^m$ as $n \to \infty$

ii. $S^m \overset{\mathcal{D}}{\to} S$ as $m \to \infty$

iii. for all $\varepsilon > 0$, $\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\|S_n - S_n^m\|_{\mathcal{H}\to\mathbb{R}} > \varepsilon\right) = 0$.

For the first property, Condition G_0 together with Lemma 13 tells us that the random vector $\boldsymbol{S}_n^m = (S_n(\phi_1), \ldots, S_n(\phi_m))$ in $\mathbb{R}^m$ is such that $\boldsymbol{S}_n^m \overset{\mathcal{D}}{\to} N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{ij} = \sigma(\phi_i, \phi_j)$ for any $i, j \in [m]$. Moreover, by Proposition 14, it holds that

$$\boldsymbol{\Sigma}_{ij} = \sigma(\phi_i, \phi_j) = \langle T_\sigma \phi_i, \phi_j \rangle_{\mathcal{H}} = \lambda_i \langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \lambda_i \delta_{ij}$$

where $\delta_{ij} = 1$ if $i = j$, otherwise it is 0. Then $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$, and thus $\boldsymbol{S}_n^m$ converges in distribution to the vector $(\sqrt{\lambda_1} Z_1, \ldots, \sqrt{\lambda_m} Z_m)$, where $Z_1, \ldots, Z_m$ are independent and identically distributed standard normal random variables. Consequently, the continuous mapping theorem and the continuity of the transformation $\mathbb{R}^m \to \mathcal{H}^*$ given by $\mathbf{x} \to \sum_{i=1}^m x_i \langle \phi_i, \cdot \rangle_{\mathcal{H}}$ yield property i. The second property follows immediately from the definition of S^m. Finally, note that $\|S_n - S_n^m\|_{\mathcal{H}}^2 = \sum_{i=m+1}^\infty S_n(\phi_i)^2$, thus property iii. follows directly from Condition G_2.

We can now apply Theorem 3.2 in Billingsley (2013) to deduce that $S_n \overset{\mathcal{D}}{\to} S$, and by the continuous mapping theorem we get $\|S_n\|_{\mathcal{H}\to\mathbb{R}}^2 \overset{\mathcal{D}}{\to} \|S\|_{\mathcal{H}\to\mathbb{R}}^2 = \sum_{i=1}^\infty \lambda_i Z_i^2$. ■

Proof of Theorem 2. Since S converges almost surely in $\mathcal{H}^*$, we get that $\|S\|_{\mathcal{H}\to\mathbb{R}}^2 = \sum_{i\geq 1} \lambda_i Z_i^2$ converges a.s., in $\mathbb{R}$, and by Lemma 15 we have $\sum_{i\geq 1} \lambda_i < \infty$.

Let's verify Condition G_0. Let $\omega \in \mathcal{H}$ and note that the map $\mathcal{H}^* \to \mathbb{R}$ given by $S \to S(\omega)$ is continuous in $\mathcal{H}^*$. Then, since $S_n \overset{\mathcal{D}}{\to} S$, the continuous mapping theorem yields

$$S_n(\omega) \overset{\mathcal{D}}{\to} \sum_{i=1}^\infty \sqrt{\lambda_i} \langle \phi_i, \omega \rangle_{\mathcal{H}} Z_i.$$

The sum on the right-hand side term converges almost surely by Doob's martingale convergence theorem: indeed, for any $\omega \in \mathcal{H}$, $M_t = \sum_{i=1}^t \sqrt{\lambda_i} Z_i \langle \phi_i, \omega \rangle_{\mathcal{H}}$ is a 0 mean martingale with second moment $\sum_{i=1}^t \lambda_i \langle \phi_i, \omega \rangle_{\mathcal{H}}^2 \leq \|\omega\|^2 \sum_{i\geq 1} \lambda_i$, i.e. uniformly bounded second moment, so the sum converges almost surely. Since Z_i has standard Normal distribution, $S_n(\omega)$ converges to a Normal with mean 0 and variance $\sigma(\omega, \omega)$ given by $\sum_{i=1}^\infty \lambda_i \langle \phi_i, \omega \rangle_{\mathcal{H}}^2$.

To verify Condition G_1 just observe that $\sigma(\phi_i, \phi_i) = \lambda_i$, and recall that $\sum_{i\geq 1} \lambda_i < \infty$. Finally, to prove Condition G_2 note that for any subspace V of $\mathcal{H}$, the transformation $\mathcal{H}^* \to \mathbb{R}$ given by $U \to \|U \circ P_V\|_{\mathcal{H}\to\mathbb{R}}$ is continuous, where P_V is the orthogonal projection onto V. Now, let V_i be the span of $\phi_1, \ldots, \phi_i$, then since $S_n \overset{\mathcal{D}}{\to} S$, the continuous mapping theorem gives $\|S_n \circ P_{V_i^\perp}\|_{\mathcal{H}\to\mathbb{R}} \overset{\mathcal{D}}{\to} \|S \circ P_{V_i^\perp}\|_{\mathcal{H}\to\mathbb{R}}$. This implies that for any $\varepsilon > 0$ we have $\limsup_{n\to\infty} \mathbb{P}\left(\|S_n \circ P_{V_i^\perp}\|_{\mathcal{H}\to\mathbb{R}} \geq \varepsilon\right) = \mathbb{P}\left(\|S \circ P_{V_i^\perp}\|_{\mathcal{H}\to\mathbb{R}} \geq \varepsilon\right)$. Finally, since $\|S \circ P_{V_i^\perp}\|^2 = \sum_{j=i+1}^\infty \lambda_j Z_j^2 \to 0$ when $i \to \infty$, we deduce that Condition G_2 holds. ■

Proof of Theorem 3. Let's assume that $\lim_{u\to\infty}\limsup_{n\to\infty}\sum_{i=u+1}^{\infty}S_n(\psi_i)^2=0$ holds for some basis $(\psi_i)_{i\geq 1}$, we will prove that $\Psi_n\to c_*^2$ a.s.

Start by noting that the limiting functional $c:\mathcal{H}\to\mathbb{R}$ given by $c(\omega)=\lim_{n\to\infty}S_n(\omega)$ is linear since S_n is linear, and by the hypothesis it is also bounded. Then, by the Riesz representation theorem, there exists $\xi\in\mathcal{H}$ such that $c(\omega)=\langle\xi,\omega\rangle_{\mathcal{H}}$. Let $\phi_1=\xi/\|\xi\|_{\mathcal{H}}$, and complete an orthonormal basis $\phi_1,\phi_2,\ldots$ of $\mathcal{H}$. Then note that

$$c(\phi_1)^2=\|\xi\|_{\mathcal{H}}^2=c_*^2 \qquad\text{and}\qquad c(\phi_i)=\|\xi\|_{\mathcal{H}}\langle\phi_1,\phi_i\rangle_{\mathcal{H}}=0,\quad\text{for all } i\geq 2.$$

Now, since $S_n:\mathcal{H}\to\mathbb{R}$ is linear and bounded, there exists $\xi_n\in\mathcal{H}$ such that

$$\sup_{\omega\in\mathcal{H}:\|\omega\|_{\mathcal{H}}=1}S_n(\omega)^2=\|\xi_n\|_{\mathcal{H}}^2=\sum_{i=1}^{\infty}\langle\xi_n,\phi_i\rangle_{\mathcal{H}}^2=\sum_{i=1}^{\infty}S_n(\phi_i)^2.$$

Note that $S_n(\phi_1)^2\to c(\phi_1)^2$, and that $S_n(\phi_i)^2\to c(\phi_i)^2=0$ a.s. for all $i\geq 2$. Then, we only need to show that $\lim_{n\to\infty}\sum_{i=2}^{\infty}S_n(\phi_i)^2=0, a.s.$ The latter follows from a simple computation. Indeed, let u be a positive integer and write $\sum_{i=2}^{\infty}S_n(\phi_i)^2=\sum_{i=2}^{u}S_n(\phi_i)^2+\sum_{i=u+1}^{\infty}S_n(\phi_i)^2$, so

$$\limsup_{n\to\infty}\sum_{i=2}^{\infty}S_n(\phi_i)^2=\limsup_{n\to\infty}\sum_{i=2}^{u}S_n(\phi_i)^2+\limsup_{n\to\infty}\sum_{i=u+1}^{\infty}S_n(\phi_i)^2=0+\limsup_{n\to\infty}\sum_{i=u+1}^{\infty}S_n(\phi_i)^2.$$

The previous holds for arbitrary u, then by taking the limit when u tends to infinity, we get $\limsup_{n\to\infty}\sum_{i=2}^{\infty}S_n(\phi_i)^2=0$ since we assumed that $\lim_{u\to\infty}\limsup_{n\to\infty}\sum_{i=u+1}^{\infty}S_n(\psi_i)^2=0$.

Now, we prove the converse. Suppose that $\Psi_n\to c_*^2$, almost surely. Consider the orthonormal basis $(\phi_i)_{i\geq 1}$ as above. Then, note that since almost surely $\lim_{n\to\infty}\sum_{i=1}^{\infty}S_n(\phi_i)^2=c_*^2$ and $\lim_{n\to\infty}S_n(\phi_1)^2=c_*^2$, it holds $\lim_{n\to\infty}\sum_{i=2}^{\infty}S_n(\phi_i)^2\to 0$, *a.s.*, and thus we get that $\lim_{u\to\infty}\lim_{n\to\infty}\sum_{i=u+1}^{\infty}S_n(\phi_i)^2=0$ a.s.

The result for convergence in probability follows by using the same arguments. ■

Proof of Proposition 4. Let $(\phi_k)_{k\geq 1}$ be an orthonormal base of $\mathcal{H}$. We start by claiming that $f_i(y):=\sum_{k\geq i}Q\phi_k(y)^2\in L_1(\mathcal{Y},\nu)$ for all $i\geq 1$. To verify such a claim, it is clear that $0\leq f_i(y)\leq f_1(y)$ for all $y\in\mathcal{Y}$, so we simply prove the claim for f_1. First, suppose that item a) holds. Then

$$\begin{aligned}\int_{\mathcal{Y}}|f_1(y)|\nu(dy)&=\sum_{k\geq 1}\int_{\mathcal{Y}}(Q\phi_k)(y)^2\nu(dy)=\sum_{k\geq 1}\left(\int_{\mathcal{Y}}\frac{(Q\phi_k)(y)^2}{\|\phi_k\|_{L_2(\mu)}^2}\nu(dy)\right)\|\phi_k\|_{L_2(\mu)}^2\\&\leq\sup_{\omega\in\mathcal{H}:\|\omega\|_{L_2(\mu)}=1}\int_{\mathcal{Y}}\left(Q\omega\right)^2(y)\nu(dy)\int K(x,x)\mu(dx)<\infty.\end{aligned}$$

where the first inequality holds since $\sum_{k\geq 1}\|\phi_k\|_{L_2(\mu)}^2=\int_{\mathcal{X}}K(x,x)\mu(dx)$, and the last inequality holds by a). Now, if item b) holds, i.e. $\mathcal{G}$ is an RKHS with kernel L and $\int_{\mathcal{Y}}L^Q(y,y)\nu(dy)<\infty$, then $\int_{\mathcal{Y}}\sum_{k\geq 1}(Q\phi_k)(y)^2\nu(dy)=\sum_{k\geq 1}\langle\phi_k,Q^*L_y\rangle_{\mathcal{H}}^2=\|Q^*L_y\|_{\mathcal{H}}^2=L^Q(y,y)$, and thus $f_1\in L_1(\mathcal{Y},\nu)$.

In either case, we have $f_i \in L_1(\nu)$ for all $i \geq 1$. Note that the previous computations also show $\lim_{i\to\infty} \|f_i\|_{L_1(\nu)} = 0$, which will be used later in the proof.

We proceed to verify Condition G_1 assuming that Condition G_0 is true. In such a case we have for any $\omega \in \mathcal{H}$ that $S_n(\omega) \overset{\mathcal{D}}{\to} S(\omega) \sim N(0, \sigma(\omega,\omega))$. Now, by considering the function $x \to x^2 \wedge M$, we have $\mathbb{E}(S^2 \wedge M) = \liminf_{n\to\infty} \mathbb{E}(S_n^2 \wedge M) \leq \liminf_{n\to\infty} \mathbb{E}(S_n^2)$, and by the dominated convergence theorem we have $\lim_{M\to\infty} \mathbb{E}(S^2 \wedge M) = \mathbb{E}(S^2)$. Then,

$$\sum_{i\geq 1} \sigma(\phi_i,\phi_i) \leq \sum_{i\geq 1} \liminf_{n\to\infty} \mathbb{E}(S_n(\phi_i)^2) \leq \sum_{i\geq 1} C \int_{\mathcal{Y}} (Q\phi_i)^2(y)\nu(dy) = \int_{\mathcal{Y}} f_1(y)\nu(dy) < \infty$$

Now, let us prove Condition G_2. For that, we need to verify that for any $\varepsilon > 0$, $\lim_{i\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sum_{k>i} S_n(\phi_k)^2 \geq \varepsilon\right) = 0$. By Markov's inequality and Eq. (7)

$$\begin{aligned}\lim_{i\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sum_{k>i} S_n(\phi_k)^2 \geq \varepsilon\right) &\leq \lim_{i\to\infty} \limsup_{n\to\infty} \frac{1}{\varepsilon} \sum_{k>i} \mathbb{E}\left(S_n(\phi_k)^2\right) \\ &\leq \frac{C}{\varepsilon} \lim_{i\to\infty} \sum_{k>i} \int_{\mathcal{Y}} (Q\phi_k)^2(y)\nu(dy) = \frac{C}{\varepsilon} \lim_{i\to\infty} \int_{\mathcal{Y}} f_i(y)\nu(dy) = 0,\end{aligned}$$

where the last equality holds since $\lim_{i\to\infty} \|f_i\|_{L_1(\nu)} = 0$. ■

Proof of Proposition 6. Following the same argument as in the proof of Proposition 4, we consider an orthonormal basis $(\phi_k)_{k\geq 1}$ of $\mathcal{G}$, and $f_i := \sum_{k\geq i} Q\phi_k^2 \in L_1(\nu)$ for all $i \geq 1$ and, moreover, $\lim_{i\to\infty} \|f_i\|_{L_1(\nu)} = 0$.

Let us verify Condition G_1. Since for each $\omega \in \mathcal{H}$ it holds that $S_n(\omega) \overset{\mathcal{D}_{\mathcal{F}'}}{\to} S(\omega) \sim N\left(0, \sigma(\omega,\omega)\right)$, then by definition of $\mathcal{F}'$-stable convergence, $\mathbb{E}(S_n(\omega)^2 \wedge M)$ converges to $\mathbb{E}(S(\omega)^2 \wedge M)$ for $M > 0$, then by taking M tending to infinity, we have

$$\begin{aligned}\sigma(\omega,\omega) = \mathbb{E}(\mathbb{E}(S(\omega)^2|\mathcal{F}')) &\leq \mathbb{E}\left(\liminf_{n\to\infty} \mathbb{E}(S_n(\omega)^2|\mathcal{F}')\right) \\ &\leq \mathbb{E}\left(\liminf_{n\to\infty} C \int_{\mathcal{Y}} f_1(y)\nu_n(dy)\right) = C \int_{\mathcal{Y}} f_1(y)\nu(dy) < \infty,\end{aligned}$$

where the first inequality is due to Fatou's Lemma, the second inequality follows from Eq. (10), and the limit holds by our assumptions in ν_n and ν and the fact that $f_1 \in L_1(\nu)$.

We continue by showing that Condition G_2 holds $\mathcal{F}'$-stably. Let $Y_{in} = \sum_{k\geq i} S_n(\phi_k)^2$, then by the Markov inequality,

$$\mathbb{P}\left(Y_{in} \geq \varepsilon \middle| \mathcal{F}'\right) \leq \frac{1}{\varepsilon}\mathbb{E}\left(Y_{in}|\mathcal{F}'\right) \leq \frac{1}{\varepsilon} \sum_{k\geq i} \int_{\mathcal{Y}} (Q\phi_k)^2(y)\nu_n(dy) = \frac{1}{\varepsilon} \int_{\mathcal{Y}} f_i(y)\nu_n(dy).$$

Therefore, $\limsup_{n\to\infty} \mathbb{P}\left(Y_{in} \geq \varepsilon|\mathcal{F}'\right) \leq \limsup_{n\to\infty} \int_{\mathcal{Y}} f_i(y)\nu_n(dy) = \int_{\mathcal{Y}} f_i(y)\nu(dy)$ a.s. By combining this inequality with the reverse Fatou's inequality, we have for every $F \in \mathcal{F}'$ with $\mathbb{P}(F) > 0$ that $\limsup_{n\to\infty} \mathbb{P}(Y_{in} \geq \varepsilon|F) = \mathbb{E}\left(\limsup_{n\to\infty} \mathbb{P}\left(Y_{in} \geq \varepsilon|\mathcal{F}'\right)|F\right) \leq \frac{1}{\varepsilon} \int_{\mathcal{Y}} f_i(y)\nu(dy)$. Finally, following the same argument as in the proof of Proposition 4, we have $\lim_{i\to\infty} \frac{1}{\varepsilon} \int_{\mathcal{Y}} f_i(y)\nu(dy) = 0$, concluding that Condition G_2 holds $\mathcal{F}'$-stably.

Appendix C. Experiments

For completeness, we include two experiments with simulated data to evaluate the empirical performance of the kernelised **GCM**, hereafter referred to as **KGCM**, as introduced in Section 4.3.1. For comparison purposes, we also implement tests based on the **wGCM** and the **GCM**.

Regarding the details of implementation, for the **KGCM** we choose the kernels as the squared exponential kernel $K_\ell(z,z') : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which is given by $K_\ell(z,z') = \exp\{-\frac{1}{\ell^2}\|z-z'\|^2\}$, where $\ell > 0$ is the length-scale parameter. The length-scale parameter ℓ controls the fluctuations of the functions of $\mathcal{H}$. A larger length-scale parameter is associated with flatter curves, whilst a smaller one is associated with functions with more fluctuations. Thus, a smaller length-scale parameter should be preferred for problems involving nonlinear structures. In our experiments, we take $\ell \in \{0.1, 0.5, 1, mh\}$ where mh denotes the median heuristic, a well-known heuristic for choosing the length-scale parameter, which takes ℓ as the median of all pairwise differences $\|Z_i - Z_j\|$ for $i, j \in [n]$. Although we do not pursue the goal of finding the best length-scale in our experiments, we remark that the problem of choosing an appropriate length-scale is currently a very active research topic in Statistics and Machine Learning (for recent results, see Albert et al. (2022); Schrab et al. (2023, 2022a,b)). To find rejection regions, we use wild bootstrap with Rademacher weights and obtain $M = 1000$ bootstrap samples to approximate the rejection regions. In our experiments, we choose $\alpha = 0.05$ as the level of the test. Since the test-statistic Ψ_n^2 is non-negative, the rejection region for **KGCM** is chosen as $(q_{1-\alpha}^M, \infty)$ where $q_{1-\alpha}^M$ is the value at position $(1-\alpha)M$ of the M bootstrap samples (when sorted in increasing order). To obtain power estimations, we repeat the experiment 1000 times. In our results below, we write **KGCM**(ℓ) to represent the kernel test with length-scale parameter ℓ.

We also implement the weighted generalised covariance measure **wGCM**. We follow the setup of Scheidegger et al. (2022) and consider fixed weight functions of the form $\omega_a(z) = \text{sign}(z-a)$ where $a \in \mathbb{R}$ (a similar construction is used for $z \in \mathbb{R}^d$). Our implementation combines k weight functions of the form ω_a with the constant weight function $\omega(z) = 1$ in a single test-statistic. We choose $k \in \{0, 1, 4, 7\}$, and we remark that when $k = 0$ we recover the **GCM**. In our results, we denote by **wGCM**(k) the test with $k+1$ weights functions. We use the same number of weight functions as in the experiments of Scheidegger et al. (2022), so the setups of the experiments are the same. We refer the reader to Scheidegger et al. (2022), particularly Section 2.3, for more details. In our experiments, we use the code provided by the authors[1], which automatically chooses the k weight functions ω_a, and performs the test.

Recall that the **GCM** and variations require the estimation of conditional means $\mathbb{E}(X|Z)$ and $\mathbb{E}(Y|Z)$. For this task, we employ polynomial regression with a degree of 3 in our first experiment and a degree of 1 in the second, as it performs well enough on our data. This allows us to focus on the testing part of the problem rather than on the estimation part of the problem. However, we highlight that **GCM**, **wGCM** and **KGCM** rely on selecting a good regression procedure satisfying Condition GCM, and thus for complex data sets, we might require more sophisticated regression methods.

We describe our data sets and the results obtained.

[1] `https://cran.r-project.org/web/packages/weightedGCM/index.html`

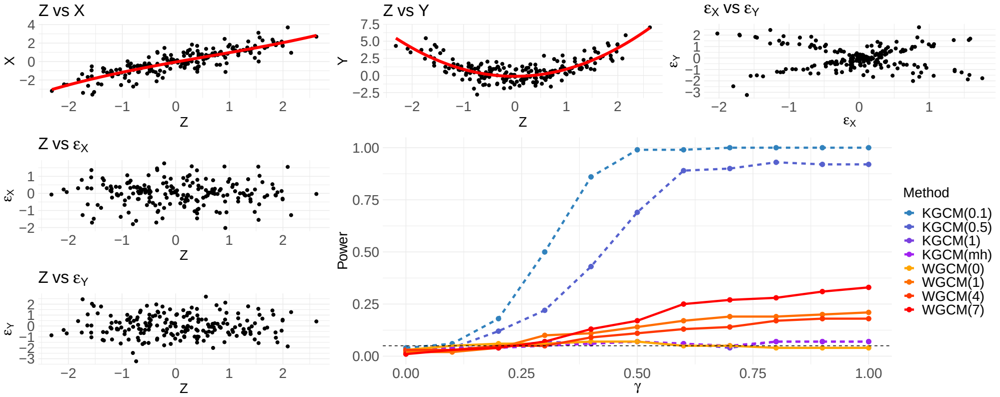

Figure 1: Scatter plots for $\gamma = 1$. The bottom-right figure shows the power of the different tests for different values of γ. The null hypothesis is recovered when $\gamma = 0$.

Data 1. Let $U_1 \sim N(0, 1)$ and $U_2 \sim N(0, 1)$ be independent. Given a parameter $\gamma \in [0, 1]$, we generate data as $Z \sim N(0, 1)$, $X = Z + U_1 \sin(5Z)$ and $Y = Z^2 + \gamma U_1 + (1 - \gamma)U_2$.

In this experiment, we vary the parameter γ, so we compute rejection rates for each of them (in a grid). Our experiments consider $n = 100$ data points, and we repeat the experiment 1000 times to estimate the rejection rate. The results are shown in Figure 1.

On the one hand, it is not difficult to see that if $\gamma = 0$, then the null hypothesis holds. Thus, in this case, the rejection rate should not be greater than the level of the test given by $\alpha = 0.05$. In Figure 1 (bottom right), we observe that all tests show a rejection rate close to $\alpha = 0.05$ (black dashed line) when $\gamma = 0$, which shows a correct Type-I error. On the other hand, when γ increases, the conditional dependence of X and Y given Z starts to be more noticeable. In fact, we expect that the rejection rate (power) starts to increase as γ approaches 1 for all tests. However, note that this will not happen for the **GCM** (**wGCM**(0) in our experiments) as for any value of $\gamma \in [0, 1]$, we have $\mathbb{E}(\epsilon_X(Z)\epsilon_Y(Z)) = \mathbb{E}((X - \mathbb{E}(X|Z))(Y - \mathbb{E}(Y|Z))) = \gamma\mathbb{E}(\sin(5Z)) = 0$, because $\mathbb{E}(X|Z) = Z$ and $\mathbb{E}(Y|Z) = Z^2$. As a consequence of the previous result, we expect that the **GCM** fails to reject the null hypothesis when the null is false. This behaviour can be observed in Figure 1.

Our experiments show promising results for the **KGCM** when the length-scale parameter is small. This good result can be explained by the fact that a smaller length-scale parameter is associated with functions with more fluctuations and thus it can be a good candidate as our data are generated by the function $z \to \sin(5z)$. Finally, we observe that the **wGCM**(k) with $k \geq 1$ is able to detect some dependence, but the results are not optimal.

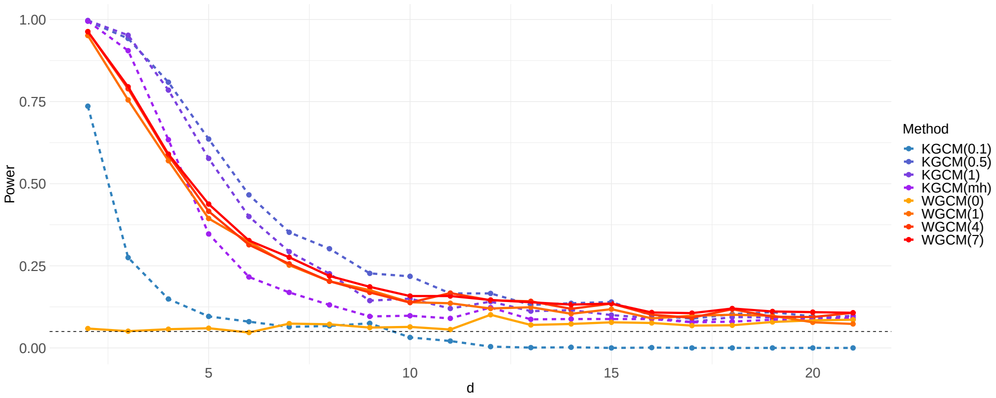

Figure 2: Test power attained by the tests for different values of the dimension d. The null hypothesis is recovered when $d \to \infty$.

Data 2. Let $d \geq 2$, and let I_d be the $d \times d$ identity matrix. Then we generate data

$$\boldsymbol{Z} = (Z_1, \dots, Z_d) \sim N(0, I_d), \quad X = Z_1 + \frac{1}{\sqrt{d}} \sum_{i=1}^{d} U_i Z_i, \quad \text{and} \quad Y = Z_2 + \frac{1}{\sqrt{d}} \sum_{i=1}^{d} U_i,$$

where $\boldsymbol{U} = (U_1, \dots, U_d) \sim N(0, I_d)$ is independent of $\boldsymbol{Z}$.

In this experiment, we now consider a multivariate $\boldsymbol{Z}$ having d dimensions (we use bold letters to remark the fact that we have a vector in $\mathbb{R}^d$). The goal of this experiment is to test whether X and Y are independent given the random vector $\boldsymbol{Z}$. Note that in this case, X is not independent of Y given $\boldsymbol{Z}$ as both X and Y depend on the vector $\boldsymbol{U}$. Also, observe that $\mathbb{E}(X|\boldsymbol{Z}) = Z_1$ and $\mathbb{E}(Y|\boldsymbol{Z}) = Z_2$, from which it can easily be deduced that $\mathbb{E}(\epsilon_X \epsilon_Y) = 0$. Lastly, observe that by the Central Limit Theorem, it holds that $\epsilon_X = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} U_i Z_i \xrightarrow{\mathcal{D}} N(0,1)$ and $\epsilon_Y = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} U_i \xrightarrow{\mathcal{D}} N(0,1)$, when d grows to infinity. Then, since $\mathbb{E}(\epsilon_X(\boldsymbol{Z})\epsilon_Y(\boldsymbol{Z})) = 0$, we can deduce that (ϵ_X, ϵ_Y) actually converges in distribution to a pair of independent standard normal random variables. Thus, we expect to observe a loss of power for all of our tests as the parameter d increases.

The results obtained by the implemented tests are shown in Figure 2. As expected, it shows that the power of all tests decreases as the dimension d increases. Also, we can observe that the **GCM** - denoted by **wGCM**(0) in Figure 2 - fails to reject the alternative for any $d \geq 2$, which is justified by the fact that $\mathbb{E}(\epsilon_X(\boldsymbol{Z})\epsilon_Y(\boldsymbol{Z})) = 0$ for any $d \geq 2$. In this example both the **wGCM** and its kernelised version perform rather well, and they do not tend to lose power very fast when d increases, except when the length-scale is 0.1, in which case the kernel test is rather weak.

References

Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2), 2022.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Michael Brendel, Arnold Janssen, Claus-Dieter Mayer, and Markus Pauly. Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41(3), 2014.

Yang Chen and Marianthi Markatou. Kernel tests for one, two, and k-sample goodness-of-fit: state of the art and implementation considerations. *Statistical Modeling in Biomedical Research: Contemporary Topics and Voices in the Field*, 2020.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International conference on machine learning*. PMLR, 2016.

Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014.

John B. Conway. *A course in operator theory*, volume 21 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2000.

Herold Dehling and Thomas Mikosch. Random quadratic forms and the bootstrap for u -statistics. *Journal of Multivariate Analysis*, 51, 1994.

Marc Ditzhaus and Markus Pauly. Wild bootstrap logrank tests with broader power functions for testing superiority. *Computational Statistics & Data Analysis*, 136, 2019.

Marc Ditzhaus, Tamara Fernández, and Nicolás Rivera. A multiple kernel testing procedure for non-proportional hazards in factorial designs. *arXiv preprint arXiv:2206.07239*, 2022.

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2014.

David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

Tamara Fernández and Nicolás Rivera. A reproducing kernel hilbert space log-rank test for the two-sample problem. *Scandinavian Journal of Statistics*, 48(4), 2021.

Tamara Fernández, Nicolás Rivera, Wenkai Xu, and Arthur Gretton. Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*. PMLR, 2020.

Tamara Fernández, Arthur Gretton, David Rindt, and Dino Sejdinovic. A kernel log-rank test of independence for right-censored data. *Journal of the American Statistical Association*, 118(542), 2023.

Thomas R. Fleming and David P. Harrington. *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991.

Richard D Gill. Censoring and stochastic integrals. *Statistica Neerlandica*, 34(2), 1980.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*. Springer, 2005.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*. MIT Press, 2006.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2007.

Erich Häusler and Harald Luschgy. *Stable convergence and stable limit theorems*, volume 74. Springer, 2015.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, 36(3), 2008.

Oscar Key, Tamara Fernandez, Arthur Gretton, and François-Xavier Briol. Composite goodness-of-fit tests with kernels. *arXiv preprint arXiv:2111.10275*, 2021.

V. S. Koroljuk and Yu. V. Borovskich. *Theory of U-statistics*, volume 273 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1994.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, 2016.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2), 2017.

Cyrill Scheidegger, Julia Hörrmann, and Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273), 2022.

Antonin Schrab, Benjamin Guedj, and Arthur Gretton. Ksd aggregated goodness-of-fit test. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022a.

Antonin Schrab, Ilmun Kim, Benjamin Guedj, and Arthur Gretton. Efficient aggregated kernel tests using incomplete u-statistics. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022b.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. Mmd aggregated two-sample test. *Journal of Machine Learning Research*, 24(194), 2023.

Robert J. Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1980.

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

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*. Springer, 2007.

Wenkai Xu and Gesine Reinert. A stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

Jiasen Yang, Vinayak Rao, and Jennifer Neville. A stein–papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2011.