ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Joint Independence Testing

Master Thesis

Niklas A. Pfister

February 15, 2016
(revised version)

Advisors: Dr. Jonas Peters, Prof. Dr. Peter Bühlmann
Department of Mathematics, ETH Zürich

## Abstract

We investigate the problem of testing whether an arbitrary number of variables are jointly independent. Our method is an extension of the kernel-based two variable Hilbert-Schmidt independence criterion (HSIC) and allows for an arbitrary number of variables, which we denote by $d$-variable Hilbert-Schmidt independence criterion (dHSIC). In the population case, the value of dHSIC is zero if and only if the $d$ variables are jointly independent. Based on an empirical estimate of dHSIC, we define four different non-parametric hypothesis tests ($H_0$: joint independence); the permutation test, the bootstrap test, the gamma approximation based test and the eigenvalue based test. We prove that the permutation test achieves significance level and that the bootstrap test and the eigenvalue based test achieve pointwise asymptotic significance level as well as consistency (i.e. are able to detect any type of dependence in the large sample limit). Finally, we show that our tests can be applied to causal inference to determine the causal ordering on simulated and real data. We compare these results to an approach which combines different 2-variable HSIC tests using a Bonferroni correction.

## Acknowledgments

I am greatly indepted to Jonas Peters at the MPI Tübingen, for providing the topic of this thesis, insightful discussions, relentless proofreading and very helpful mentoring. Furthermore, I would also like to thank Peter Bühlmann for unproblematic mentoring at ETH Zürich and Bernhard Schölkopf for giving me the opportunity to write part of my thesis at the MPI Tübingen. Finally, I'm grateful to Mateo Rojas-Carulla, Sari De Martin and Vincent Lohmann for proofreading parts of the thesis and giving me many helpful comments.

# Contents

Chapter 1

# Introduction

## 1.1 Problem and motivation

We consider the problem whether in a sample $(\mathbf{X}_1, \ldots, \mathbf{X}_m)$, with $\mathbf{X}_i = (X_i^1, \ldots, X_i^d)$, the $d$ components $(X_1^1, \ldots, X_m^1), \ldots, (X_1^d, \ldots, X_m^d)$ are jointly independent. For example, a lot of statistical methods require the input variables to be jointly independent. It is therefore essential to be able to verify this assumption before applying these methods. Another example where this problem appears is causality, in particular in causal inference. This field deals with finding causal relationships between variables, e.g. which biological markers are causal for a certain disease. Causal statements over $d$ variables can be formalized using the concept of an SEM, for example. SEMs assume the existence of $d$ noise variables which are required to be jointly independent. As a consequence, many causal inference methods need to measure dependence of some sort; often this dependence can be non-linear, for example, when analyzing residuals after a linear regression. This motivates the need for widely applicable methods to test for and measure dependence.

The main difficulty when analyzing dependence between variables is that the dependence structure can be arbitrarily complicated. It is therefore not at all obvious how to appropriately measure dependence in a way that is sufficiently general to include all possibilities. For example, using correlation as a measure of dependence without additional assumptions is not sufficient because there exist random variables which are uncorrelated but dependent. Most classical methods for testing independence rely on assumptions on the underlying random variables. The Pearson's chi-squared independence test (e.g. Lehmann and Romano, 2005), for example, requires categorical data and Hoeffding's independence test (see Hoeffding, 1948b) requires a continuous density.

Several non-parametric approaches based on kernel methods have been developed at the beginning of the 21st century, most of which are based on the covariance operator in the reproducing kernel Hilbert space (RKHS). One of them is the Hilbert-Schmidt independence criterion (HSIC), which was introduced by Gretton et al. (2005) and is a measure of dependence for two random variables. It has the desirable property of being zero if and only if the two random variables are independent and additionally allows to be consistently estimated based on a finite sample. These two properties allow the detection of any type of dependence given a sufficiently large sample size.

This thesis introduces a direct extension of the two variable HSIC to a version allowing for an arbitrary number of variables, which we call the $d$-variable Hilbert-Schmidt independence criterion (dHSIC). Based on this criterion, we introduce four different hypothesis tests that can be used to test for statistically significant evidence of dependence.

## 1.2 Outline

In Section 2 we introduce all of the required background material. This involves a short summary of required tools from functional analysis, an introduction to kernel methods, an exposition of the theory of U-and V-statistics and finally a summary of the statistical framework used throughout this thesis. This section can be skipped and only used as reference whenever the notation or concepts become unclear in the remaining thesis. The main part starts in Section 3, where we introduce the $d$-variable Hilbert-Schmidt independence criterion and its empirical estimator. In Section 4 we construct four different hypothesis tests based on dHSIC and prove important properties of them. At the end of this section, we also give details on how to implement each of the four the hypothesis tests. Section 5 numerically assesses level, power and runtime of the hypothesis tests. Finally, in Section 6 we consider an application of dHSIC to causal inference.

## 1.3 Contributions

This thesis builds on the papers of Gretton et al. (2005), Gretton et al. (2007) and Smola et al. (2007) but extends the material in several aspects. First of all, the extension of the two variable HSIC to the generalized version dHSIC is new and has previously only been briefly mentioned by Sejdinovic et al. (2013), in the case of three variables. All results related to dHSIC are consequently also new, however most of them carry over directly from the existing results for HSIC. We prove new results about V-statistics which are required in the proofs of some of these statements. Previously, these results have only been stated without proof. The mathematical rigorous treatment of the permutation test and the bootstrap test is novel, as are the results about their level and consistency given in Proposition 4.5, Proposition 4.9 and Proposition 4.10. For the extension of the gamma approximation based test, we computed both the mean and variance in the general $d$-variable setting in Lemma 4.11 and Lemma 4.12, which turns out to be rather complex. The idea for the eigenvalue test comes from Gretton et al. (2009), in which this approach has been applied to the Maximum Mean Discrepancy (MMD). The MMD is directly related to the HSIC and hence we are able to extend the results to dHSIC. Finally, in order to make our tests accessible for everyone we have created an R-package, which will be available on CRAN. The following list summarizes the contributions:

(i) extension of HSIC to dHSIC

(ii) new results for V-statistics (have been used but not proved)

(iii) rigorous treatment of permutation and bootstrap test

(iv) gamma approximation for $d$-variables

(v) eigenvalue approach based test

(vi) R-package containing all tests (to be available on CRAN)

Chapter 2

# Background material

For simplicity we adopt the convention that any linear space is always a linear space over the field of the real numbers. The following list summarizes basic notation used throughout this work.

- Let $\mathcal{B}$ be a Banach space. Then we denote by

  $$L(\mathcal{B}) := \{A : \mathcal{B} \to \mathcal{B} \mid A \text{ bounded linear}\}$$

  the space of bounded linear operators.

- Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $\mathcal{B}$ a Banach space. Then we denote by

  $$\mathcal{L}^0(\mu, \|\cdot\|_{\mathcal{B}})$$

  the space of Borel-measurable functions from $\Omega$ to $\mathcal{B}$, and by

  $$\mathcal{L}^p(\mu, \|\cdot\|_{\mathcal{B}})$$

  the space of $p$-integrable functions from $\Omega$ to $\mathcal{B}$, and by

  $$L^p(\mu, \|\cdot\|_{\mathcal{B}})$$

  the space of equivalence classes of $p$-integrable functions from $\Omega$ to $\mathcal{B}$. For more details see Definition A.8 and Definition A.9.

- Let $\mathcal{X}$ be a separable metric space. We then denote by

  $$\mathcal{M}_f(\mathcal{X})$$

  the space of finite Borel measures, and by

  $$\mathcal{P}(\mathcal{X})$$

  the space of finite Borel probability measures, for more details see Definition 2.23.

- Let $\mathcal{H}$ be a Hilbert space. We then denote by
$$L_1(\mathcal{H})$$
the space of nuclear operators, and by
$$\mathrm{HS}(\mathcal{H})$$
the space of Hilbert-Schmidt operators, see Section 2.1.1 and Section 2.1.2 for more details.

- Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be a Hilbert spaces and let $A : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator. Then the adjoint of $A$ is the unique bounded linear operator $A^* : \mathcal{H}_2 \to \mathcal{H}_1$ satisfying for all $x \in \mathcal{H}_1$, $y \in \mathcal{H}_2$ that
$$\langle Ax, y\rangle_{\mathcal{H}_2} = \langle x, A^*y\rangle_{\mathcal{H}_1}.$$
Existence and uniqueness follow from the Riesz representation theorem.

- Let $\mathcal{H}$ be a Hilbert space and let $v, w \in \mathcal{H}$. We denote by $v \otimes w \in L(\mathcal{H})$ the function with the property that for all $x \in \mathcal{H}$ it holds that
$$(v \otimes w)(x) := \langle v, x\rangle_{\mathcal{H}} w. \tag{2.1}$$
Any operator of this form is referred to as rank one operator. While it will be clear from the context one has to be careful not to confuse this with the tensor product of functions or kernels introduced in Appendix A.2.

- Let $\mathcal{X}$ be a metric space with metric $d$ and let $U \subseteq \mathcal{X}$. We define the set
$$\overline{U} := \left\{ x \in \mathcal{X} \,\middle|\, \inf_{y \in U} d(x, y) = 0 \right\}$$
and call it the closure of $U$.

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{B}$ be a Banach space and let $X : \Omega \to \mathcal{B}$ be a random variable (i.e. a measurable map). Then we denote by $\mathbb{P}^X$ the Borel-measure on $\mathcal{B}$ with the property that for all Borel-measurable sets $A \subseteq \mathcal{B}$ it holds that
$$\mathbb{P}^X(A) = \mathbb{P}\left(X^{-1}(A)\right)$$
and call it the law of X (image measure of X).

- Let $(\Omega_1, \mathcal{F}_1, \mu_1)$,. . . ,$(\Omega_n, \mathcal{F}_n, \mu_n)$ be measure spaces then we denote by
$$(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n, \mu_1 \otimes \cdots \otimes \mu_n)$$
the product measure space. In particular, $\mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$ is the product $\sigma$-algebra, which corresponds to the smallest $\sigma$-algebra generated by sets of the form $A_1 \times \cdots \times A_n$ and $\mu_1 \otimes \cdots \otimes \mu_n$ is the product measure which satisfies
$$(\mu_1 \otimes \cdots \otimes \mu_n)(A_1 \times \cdots \times A_n) = \mu_1(A_1) \cdots \mu_n(A_n).$$
In order to simplify notation the $n$-fold product of the space $(\Omega, \mathcal{F}, \mu)$ is denoted by
$$(\Omega^n, \mathcal{F}^{\otimes n}, \mu^{\otimes n}).$$

## 2.1 Some functional analysis on Hilbert spaces

In this section we recall some reference material related to functional analysis on Hilbert spaces. Most of the material is from the lecture notes of Jentzen (2015). All the material, however, can also be found in Werner (2011) or Dunford and Schwartz (1963).

### 2.1.1 Nuclear operators

**Definition 2.1 (nuclear operators)**
*Let $\mathcal{H}$ be a Hilbert space, and let $A \in L(\mathcal{H})$ satisfy that there exist sequences $(v_n)_{n\in\mathbb{N}}, (w_n)_{n\in\mathbb{N}} \subseteq \mathcal{H}$ such that*

$$\sum_{n=1}^{\infty} \|v_n\|_{\mathcal{H}} \|w_n\|_{\mathcal{H}} < \infty$$

*and such that for all $x \in \mathcal{H}$ it holds that*

$$Ax = \sum_{n=1}^{m} \langle v_n, x\rangle_{\mathcal{H}} w_n.$$

*Then, $A$ is called a nuclear operator.*

Next, we define the space

$$L_1(\mathcal{H}) := \{A \in L(\mathcal{H}) \mid A \text{ is nuclear operator}\}$$

of nuclear operators and define a norm on $L_1(\mathcal{H})$ as follows

$$\|A\|_1 := \inf\Bigg\{a \in \mathbb{R} \;\Bigg|\; \exists (v_n)_{n\in\mathbb{N}}, (w_n)_{n\in\mathbb{N}} \subseteq \mathcal{H} \text{ s.t.} \\ \Big(a = \sum_{n=1}^{\infty} \|v_n\|_{\mathcal{H}} \|w_n\|_{\mathcal{H}} \text{ and } Ax = \sum_{n=1}^{m} \langle v_n, x\rangle_{\mathcal{H}} w_n \quad \forall x \in \mathcal{H}\Big)\Bigg\}.$$

The following theorem shows that the space $L_1(\mathcal{H})$ has a separable Banach space structure.

**Theorem 2.2 (structure of $L_1(\mathcal{H})$)**
*Let $\mathcal{H}$ be a separable Hilbert space, then the space $L_1(\mathcal{H})$ is a separable Banach space.*

**Proof** We only give a short outline of the proof. The fact that $L_1(\mathcal{H})$ forms a Banach space is a classical result and can be found in Werner (2011, Satz VI.5.3). To see separability, observe that the space of finite rank operators is dense in $L_1(\mathcal{H})$. It is therefore sufficient to show that the space of finite rank operators is separable. To this end let $(\varphi_n)_{n\in\mathbb{N}}$ be an orthonormal basis of $\mathcal{H}$. Then, the set given by

$$\mathcal{A} = \left\{ \sum_{i,j=1}^{m} a_{ij} \varphi_i \otimes \varphi_j \mid a_{ij} \in \mathbb{Q}, m \in \mathbb{N} \right\}$$

is a countable set. Furthermore, it can be shown that any finite rank operator can be approximated by elements in $\mathcal{A}$. This completes the outline of the proof of Theorem 2.2.□

Next, we introduce the trace operator.

**Definition 2.3 (trace)**
*Let $\mathcal{H}$ be a Hilbert space, assume $\mathbb{B} \subseteq \mathcal{H}$ is an orthonormal basis of $\mathcal{H}$ and let $A \in L(\mathcal{H})$. Then, the trace of $A$ is given by*

$$\operatorname{trace}(A) := \sum_{b \in \mathbb{B}} \langle b, Ab \rangle_{\mathcal{H}}.$$

It can be shown (e.g. Werner, 2011, Section VI.5) that the trace is independent of the chosen orthonoromal basis. Furthermore, for an operator $A \in L(\mathcal{H})$, it holds that $\operatorname{trace}(|A|) < \infty$ if and only if $A \in L_1(\mathcal{H})$, where $|A| = (A^*A)^{\frac{1}{2}}$. This is the reason nuclear operators are often referred to as trace class operators. Moreover, it then holds that

$$\operatorname{trace}(|A|) = \|A\|_1.$$

### 2.1.2 Hilbert-Schmidt operators

**Definition 2.4 (Hilbert-Schmidt operator)**
*Let $\mathcal{H}$ be a Hilbert space, and let $A \in L(\mathcal{H})$ satisfy that there exists an orthonormal basis $\mathbb{B} \subseteq \mathcal{H}$ of $\mathcal{H}$ such that*

$$\sum_{b \in \mathbb{B}} \|Ab\|_{\mathcal{H}}^2 < \infty.$$

*Then, $A$ is called a Hilbert-Schmidt operator.*

Moreover, we define the space

$$\operatorname{HS}(\mathcal{H}) := \{A \in L(\mathcal{H}) \mid A \text{ is Hilbert-Schmidt operator}\}$$

of Hilbert-Schmidt operators.

Given operators $A, B \in \operatorname{HS}(\mathcal{H})$ and an orthonormal basis $\mathbb{B} \subseteq \mathcal{H}$, it can be shown (e.g. Werner, 2011, Satz VI.6.2) that the mapping defined by

$$\langle A, B \rangle_2 := \sum_{b \in \mathbb{B}} \langle Ab, Bb \rangle_{\mathcal{H}}.$$

is independent of the orthonormal basis $\mathbb{B}$ and thus a well defined inner product on $\operatorname{HS}(\mathcal{H})$. Furthermore, it induces the norm

$$\|A\|_2 = \left( \sum_{b \in \mathbb{B}} \|Ab\|_{\mathcal{H}}^2 \right)^{\frac{1}{2}}$$

and turns $\operatorname{HS}(\mathcal{H})$ into a separable Hilbert space.

### 2.1.3 Covariance operator

In this section we introduce the covariance operator. For our purposes the non-centered version is sufficient, but the same results also hold for the centered version.

**Definition 2.5 (non-centered covariance operator)**
*Let $\mathcal{H}$ be a Hilbert space and let $\mu \in \mathcal{P}(\mathcal{H})$ satisfy that for all $w \in \mathcal{H}$ it holds that $\int_{\mathcal{H}} |\langle w, v\rangle_{\mathcal{H}}|^2 \, \mu(dv) < \infty$. Then we denote by $\mathrm{CovOp}(\mu) \in L(\mathcal{H})$ the unique bounded linear operator such that for all $v, w \in \mathcal{H}$ it holds that*

$$\langle v, \mathrm{CovOp}(\mu)_{\mathcal{H}} w\rangle = \int_{\mathcal{H}} \langle v, x\rangle_{\mathcal{H}} \langle x, w\rangle_{\mathcal{H}} \, \mu(dx).$$

*Given an $\mathcal{H}$-valued random variable $X$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the property that for all $w \in \mathcal{H}$ it holds that $\mathbb{E}\left(|\langle X, w\rangle_{\mathcal{H}}|^2\right) < \infty$, we denote by $\mathrm{CovOp}(X) \in L(\mathcal{H})$ the operator defined by*

$$\mathrm{CovOp}(X) := \mathrm{CovOp}(\mathbb{P}^X).$$

The following example illustrates that the covariance operator is simply an extension of the covariance matrix in finite dimensions.

**Example 2.6 (covariance operator in finite dimensions)**
*Consider an $\mathbb{R}^n$-valued random variable $\mathbf{X}$ with law $\mathbb{P}^{\mathbf{X}}$. Then the non-centered covariance matrix is given by*

$$\Sigma := \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right),$$

*and therefore it holds for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ that*

$$\begin{aligned}
\langle \mathbf{v}, \Sigma\mathbf{w}\rangle_{\mathbb{R}^n} &= \mathbf{v}^\top \Sigma \mathbf{w} \\
&= \mathbf{v}^\top \mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right) \mathbf{w} \\
&= \mathbb{E}\left(\mathbf{v}^\top \mathbf{X}\mathbf{X}^\top \mathbf{w}\right) \\
&= \mathbb{E}\Big(\langle \mathbf{v}, \mathbf{X}\rangle_{\mathbb{R}^n} \langle \mathbf{X}, \mathbf{w}\rangle_{\mathbb{R}^n}\Big) \\
&= \int_{\mathbb{R}^n} \langle \mathbf{v}, \mathbf{x}\rangle_{\mathbb{R}^n} \langle \mathbf{x}, \mathbf{w}\rangle_{\mathbb{R}^n} \, \mathbb{P}^{\mathbf{X}}(d\mathbf{x}).
\end{aligned}$$

*This immediately implies that*

$$\mathrm{CovOp}(\mathbf{X}) = \Sigma.$$

The (non-centered) covariance operator satisfies the following properties (e.g. Da Prato and Zabczyk, 2014, Section 1.2).

**Lemma 2.7 (properties of the covariance operator)**
*Let $\mathcal{H}$ be a Hilbert space and $X \in \mathcal{L}^2(\mathbb{P}, \|\cdot\|_{\mathcal{H}})$. Then it holds that*

*(i)* $\mathrm{CovOp}(X)$ *is a symmetric, non-negative, nuclear operator,*

*(ii)* $\operatorname{CovOp}(X) = \mathbb{E}\left(X \otimes X\right)$ *and*

*(iii)* $\|\operatorname{CovOp}(X)\|_1 = \mathbb{E}\left(\|X\|_{\mathcal{H}}^2\right)$.

### 2.1.4 Spectral Theory

**Definition 2.8 (point spectrum)**
*Let $\mathcal{B}$ be a Banach space and let $A \in L(\mathcal{B})$, then we set*

$$\sigma_p\left(A\right) = \{\lambda \in \mathbb{C} \mid (\lambda - A) \text{ is not injective}\}$$

*and call this set the point spectrum of $A$.*

**Definition 2.9 (symmetric operator)**
*Let $\mathcal{H}$ be a Hilbert space and let $A \in L(\mathcal{H})$ satisfying for all $x, y \in \mathcal{H}$ that*

$$\langle x, Ay\rangle_{\mathcal{H}} = \langle Ax, y\rangle_{\mathcal{H}}.$$

*Then, $A$ is called symmetric (selfadjoint).*

It is straight forward to check that for a symmetric operator $A \in L(\mathcal{H})$ it holds that $\sigma_p(A) \subseteq \mathbb{R}$.

**Definition 2.10 (compact operator)**
*Let $\mathcal{B}$ be a Banach space and let $A \in L(\mathcal{B})$ satisfying for every boundend set $B \in \mathcal{B}$ that $A(B)$ is relatively compact in $\mathcal{B}$, i.e. $\overline{A(B)}$ is a compact subset of $\mathcal{B}$. Then, $A$ is called a compact operator.*

We denote by

$$\mathcal{K}(\mathcal{B}) := \{A \in L(\mathcal{B}) \mid A \text{ is a compact operator}\}$$

the space of compact operators. It is straightforward to see that every nuclear operator is a Hilbert-Schmidt operator. In Dunford and Schwartz (1963, Theorem 6, Section 6) it is furthermore shown that every Hilbert-Schmidt operator is a compact operator. This leads to the following relations,

$$L_1(\mathcal{B}) \subseteq \operatorname{HS}(\mathcal{B}) \subseteq \mathcal{K}(\mathcal{B}). \tag{2.2}$$

The following theorem is a version of the famous spectral theorem for compact operators. It is taken from Werner (2011, Theorem VI.3.2).

**Theorem 2.11 (spectral theorem for compact operators)**
*Let $\mathcal{H}$ be a Hilbert space and $A \in \mathcal{K}(\mathcal{H})$ a symmetric operator. Then there exists an at most countable index set $I \subseteq \mathbb{N}$, an orthonormal system $(e_i)_{i \in I} \subseteq \mathcal{H}$ and a set $(\lambda_i)_{i \in I} \subseteq \mathbb{R}\backslash\{0\}$, such that $\lim_{i\to\infty} \lambda_i = 0$ (if $|I| = \infty$) and*

$$\mathcal{H} = \ker(A) \oplus \overline{\operatorname{span}\{e_i \mid i \in I\}}.$$

*Moreover, for all $x \in \mathcal{H}$ it holds that*

$$Ax = \sum_{i \in I} \lambda_i \langle x, e_i \rangle_{\mathcal{H}} e_i.$$

*In particular, this means that $(\lambda_i)_{i \in I}$ are the eigenvalues of $A$ different from $0$ and $(e_i)_{i \in I}$ are the corresponding eigenfunctions.*

Let $\mathcal{H}$ a Hilbert space, $A \in L_1(\mathcal{H})$ and $B \in \mathrm{HS}(\mathcal{H})$. Then the spectral theorem together with (2.2) allows us to represent $A$ and $B$ in terms of their eigenvalues as

$$Ax = \sum_{i \in I} \lambda_i \langle x, e_i \rangle_{\mathcal{H}} e_i \quad \text{and} \quad Bx = \sum_{i \in I'} \nu_i \langle x, f_i \rangle_{\mathcal{H}} f_i.$$

Using the orthonormality of the eigenfunctions this implies that the norms can be expressed as

$$\|A\|_1 = \sum_{i \in I} |\lambda_i| \quad \text{and} \quad \|B\|_2 = \left( \sum_{i \in I'} |\nu_i|^2 \right)^{\frac{1}{2}}.$$

## 2.2 Kernel methods

The aim of this section is to give a short introduction to reproducing kernel Hilbert spaces and present the notation that is used in the following chapters. It is mostly based on Peters (2008) and Berlinet and Thomas-Agnan (2004).

### 2.2.1 Kernels

We begin by introducing kernels, which form the building block of reproducing kernel Hilbert spaces.

**Definition 2.12 (kernel)**
*Let $\mathcal{X}$ be a set, then a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is symmetric in its input arguments is called a (symmetric) kernel on $\mathcal{X}$.*

**Definition 2.13 (Gram matrix)**
*Let $\mathcal{X}$ be a set, let $m \in \mathbb{N}$, let $k$ be a kernel on $\mathcal{X}$ and $x_1, \dots, x_m \in \mathcal{X}$, then the matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ satisfying for all $i, j \in \{1, \dots, m\}$ that*

$$\mathbf{K}_{ij} = k(x_i, x_j)$$

*is called the Gram matrix of the kernel $k$ (given observations $x_1, \dots, x_m$).*

**Definition 2.14 (positive semi-definite matrix)**
*Let $\mathcal{X}$ be a set, let $m \in \mathbb{N}$, then a symmetric matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$ is called positive semi-definite if for all $\mathbf{z} \in \mathbb{R}^m$ it holds that*

$$\mathbf{z}^\top \mathbf{K} \mathbf{z} \geq 0.$$

**Definition 2.15 (positive semi-definite kernel)**
*Let $\mathcal{X}$ be a set, then a kernel $k$ on $\mathcal{X}$ is called positive semi-definite if for all $m \in \mathbb{N}$ and for all $x_1, \dots, x_m \in \mathcal{X}$ the Gram matrix $\mathbf{K}$ of the kernel $k$ given the observations $x_1, \dots, x_m$ is positive semi-definite.*

The following list of functions are common examples of positive semi-definite kernels on $\mathcal{X} = \mathbb{R}^n$.

- Gaussian kernel with bandwith $\sigma > 0$,
$$k(x, y) = \exp\left(-\frac{\|x - y\|_{\mathbb{R}^n}^2}{2\sigma^2}\right)$$
- polynomial kernel of degree $d \in \mathbb{N}$,
$$k(x, y) = \langle x, y \rangle_{\mathbb{R}^n}^d$$
- sigmoid kernel with $\kappa > 0$ and $\theta < 0$,
$$k(x, y) = \tanh\left(\kappa \langle x, y \rangle_{\mathbb{R}^n} + \theta\right).$$

### 2.2.2 Reproducing kernel Hilbert spaces

**Definition 2.16 (space of real-valued functions on $\mathcal{X}$)**
*Let $\mathcal{X}$ be a set. Then the space*

$$\mathcal{F}(\mathcal{X}) = \{f : \mathcal{X} \to \mathbb{R} \mid f \textit{ is a function}\}$$

*together with the standard scalar multiplication and summation defined for all $\lambda \in \mathbb{R}$, and for all $f, g \in \mathcal{F}(\mathcal{X})$ by*

$$\begin{aligned} (\lambda \cdot f)(x) &:= \lambda f(x) && \forall x \in \mathcal{X} \\ (f + g)(x) &:= f(x) + g(x) && \forall x \in \mathcal{X} \end{aligned}$$

*forms a linear space over $\mathbb{R}$. We call $\mathcal{F}(\mathcal{X})$ the space of real-valued functions on $\mathcal{X}$.*

Reproducing kernel Hilbert spaces on $\mathcal{X}$ are well-behaved subspaces of $\mathcal{F}(\mathcal{X})$. This is made precise in the following definition.

**Definition 2.17 (Reproducing kernel Hilbert space)**
*Let $\mathcal{X}$ be a set, let $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X})$ be a Hilbert space. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) if there exists a kernel $k$ on $\mathcal{X}$ satisfying*

*(i)* $\forall x \in \mathcal{X}$: $k(x, \cdot) \in \mathcal{H}$ *and*

*(ii)* $\forall f \in \mathcal{H}$, $\forall x \in \mathcal{X}$: $\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$.

*Moreover, we call $k$ a reproducing kernel of $\mathcal{H}$.*

In order to get a better understanding of this definition we go over several consequences and properties of reproducing kernel Hilbert spaces.

The next proposition shows that reproducing kernels are unique.

**Proposition 2.18 (uniqueness of the kernel)**
*Let $\mathcal{X}$ be a set and let $\mathcal{H}$ be an RKHS on $\mathcal{X}$. Assume both $k$ and $\tilde{k}$ are reproducing kernels of $\mathcal{H}$. Then $k = \tilde{k}$.*

**Proof** Observe that by the properties of a reproducing kernel it holds for all $x, y \in \mathcal{X}$ that

$$k(x, y) = \langle k(x, \cdot), \tilde{k}(y, \cdot)\rangle_{\mathcal{H}} = \langle \tilde{k}(y, \cdot), k(x, \cdot)\rangle_{\mathcal{H}} = \tilde{k}(y, x) = \tilde{k}(x, y)$$

which completes the proof of Proposition 2.18. □

The following theorem gives an alternative characterization of RKHS (e.g. Berlinet and Thomas-Agnan, 2004, Theorem 1).

**Theorem 2.19 (alternative characterization)**
*Let $\mathcal{X}$ be a set and for all $x \in \mathcal{X}$ let $\delta_x : \mathcal{F}(\mathcal{X}) \to \mathbb{R}$ be the function with the property that for all $f \in \mathcal{F}(\mathcal{X})$ it holds that $\delta_x(f) = f(x)$. Then, a Hilbert space $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X})$ is a reproducing kernel Hilbert space if and only if for each $x \in \mathcal{X}$ the function $\delta_x$ is continuous on $\mathcal{H}$.*

**Proof** Assume $\mathcal{H}$ is an RKHS with reproducing kernel $k$, then for all $f \in \mathcal{H}$ and all $x \in \mathcal{X}$ it holds that

$$\delta_x(f) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}$$

which implies that $\delta_x$ is a linear. Together with the Cauchy-Schwarz inequality we also have

$$|\delta_x(f)| = |\langle f, k(x, \cdot)\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|k(x, \cdot)\|_{\mathcal{H}} = \|f\|_{\mathcal{H}} \left[k(x, x)\right]^{\frac{1}{2}}.$$

Furthermore notice that for $f = k(x, \cdot)$ the upper bound is achieved which implies that

$$\|\delta_x\| = \sup_{\|f\|_{\mathcal{H}} \neq 0} \frac{|\delta_x(f)|}{\|f\|_{\mathcal{H}}} = \left[k(x, x)\right]^{\frac{1}{2}}.$$

Therefore $\delta_x$ is continuous.

Conversely assume $\delta_x$ is continuous for all $x \in \mathcal{X}$. Then for fixed $x \in \mathcal{X}$ by Riesz's representation theorem there exists a function $\Phi_x \in \mathcal{H}$ such that for all $f \in \mathcal{H}$ it holds that

$$\langle f, \Phi_x\rangle_{\mathcal{H}} = \delta_x(f) = f(x).$$

Since this holds for all $x \in \mathcal{X}$ we can set for all $x, y \in \mathcal{X}$, $k(x, y) = \Phi_x(y)$ which is a reproducing kernel on $\mathcal{H}$, which implies that $\mathcal{H}$ is an RKHS. This completes the proof of Theorem 2.19. □

**Proposition 2.20 (reproducing kernels are positive semi-definite)**
*Let $\mathcal{X}$ be a set and let $\mathcal{H}$ be an RKHS with reproducing kernel $k$. Then $k$ is positive semi-definite.*

**Proof** Let $x_1, \dots, x_m \in \mathcal{X}$ and let $\mathbf{z} \in \mathbb{R}^m$, then

$$\begin{aligned}
\mathbf{z}^\top \mathbf{K} \mathbf{z} &= \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) z_i z_j \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \langle z_i k(x_i, \cdot), z_j k(x_j, \cdot) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{m} z_i k(x_i, \cdot), \sum_{j=1}^{m} z_j k(x_j, \cdot) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^{m} z_i k(x_i, \cdot) \right\|_{\mathcal{H}} \\
&\geq 0
\end{aligned}$$

which completes the proof of Proposition 2.20. □

The next theorem shows that the reverse of the above proposition is in fact also true, which in particular proves the existence of non-trivial reproducing kernel Hilbert spaces. The proof is very constructive and illustrates how an RKHS can be constructed from a positive semi-definite kernel (e.g. Peters, 2008, Proposition 3.10).

**Theorem 2.21 (positive semi-definite kernels induce RKHS)**
*Let $\mathcal{X}$ be a set and $k$ a positive semi-definite kernel on $\mathcal{X}$, then there exists an RKHS on $\mathcal{X}$ with reproducing kernel $k$.*

**Proof** Define the space

$$\mathcal{H}^0 = \left\{ f : \mathcal{X} \to \mathbb{R} \mid f = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot) \text{ for some } m \text{ and some } \alpha_i \in \mathbb{R} \right\}$$

Then for $f = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$ and $g = \sum_{j=1}^{n} \beta_j k(y_j, \cdot)$ define the function $\langle \cdot, \cdot \rangle : \mathcal{H}^0 \times \mathcal{H}^0 \to \mathbb{R}$ by

$$\begin{aligned}
\langle f, g \rangle &= \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j k(x_i, y_j) \\
&= \sum_{i=1}^{m} \alpha_i g(x_i) \\
&= \sum_{j=1}^{n} \beta_j f(y_j).
\end{aligned}$$

The last two identities show that the expression does not depend on the expansion of f or g. Therefore, $\langle \cdot, \cdot \rangle$ is well-defined. Furthermore it is symmetric and bilinear. It is positive semi-definite, since

$$\langle f, f \rangle = \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

Using a generalized version of the Cauchy-Schwarz in equality for symmetric positive semi-definite bilinear forms, it follows for all $x \in \mathcal{X}$ that

$$f(x)^2 = \langle f, k(\cdot, x)\rangle^2 \leq \langle f, f\rangle k(x, x)$$

and hence $\langle f, f\rangle = 0$ implies that $f = 0$. So we have shown that $\langle \cdot, \cdot \rangle$ is an inner product.

Finally let $\mathcal{H}$ be the completion of $\mathcal{H}^0$, i.e. $\mathcal{H}$ is a Hilbert space containing a dense subspace which is isometric with $\mathcal{H}^0$. Now observe that given a sequence $(f_i)_{i\in\mathbb{N}} \subseteq \mathcal{H}^0$ converging to $f \in \mathcal{H}$ it holds for all $x \in \mathcal{X}$ and all $i, j \in \mathbb{N}$ that

$$|f_i(x) - f_j(x)| = |\langle k(x, \cdot), f_i - f_j\rangle| \leq \sqrt{k(x,x)}\|f_i - f_j\|.$$

This implies that norm convergence in $\mathcal{H}^0$ implies pointwise convergence. Therefore we get

$$\begin{aligned}\langle f, k(x, \cdot)\rangle &= \langle \lim_{i\to\infty} f_i, k(x, \cdot)\rangle \\ &= \lim_{i\to\infty} \langle f_i, k(x, \cdot)\rangle \\ &= \lim_{i\to\infty} f_i(x) \\ &= f(x).\end{aligned}$$

This implies that the reproducing property also holds for $\mathcal{H}$ which completes the proof of Theorem 2.21. □

Given an RKHS, it is of interest whether one can achieve more regularity of the RKHS by making further assumptions on the space $\mathcal{X}$ and on the kernel $k$. The next theorem gives one such result (e.g. Berlinet and Thomas-Agnan, 2004, Corollary 4).

**Theorem 2.22 (separability and continuity)**
*Let $\mathcal{X}$ be a separable metric space, let $k$ be a continuous, bounded and positive semi-definite kernel on $\mathcal{X}$ and let $\mathcal{H}$ be the RKHS with reproducing kernel $k$. Then, $\mathcal{H}$ is separable Hilbert space consisting only of continuous functions. Furthermore, given an orthonormal basis $(\varphi_n)_{n\in\mathbb{N}}$ of $\mathcal{H}$ it holds for all $x, y \in \mathcal{X}$ that*

$$k(x, y) = \sum_{n=1}^{\infty} \varphi_n(x)\varphi_n(y).$$

The Gaussian kernel on $\mathbb{R}^n$ satisfies all these conditions, therefore the associated RKHS is a separable Hilbert space consisting only of continuous functions.

### 2.2.3 Tensor products of RKHS

In this section we collect some basic facts about tensor products of RKHS and introduce the notation that is used later in this work. A short introduction to tensor products is given in Appendix A.2.

For $j \in \{1, \ldots, d\}$, let $\mathcal{X}^j$ be a separable metric space and denote by $\boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^d$ the product space. Moreover for $j \in \{1, \ldots, d\}$, let $k^j$ be a continuous bounded positive semi-definite kernel on $\mathcal{X}^j$ and denote by $\mathcal{H}^j$ the corresponding RKHS. Let $\mathbf{k} = k^1 \otimes \cdots \otimes k^d$ be the tensor product of the kernels $k^j$ and $\boldsymbol{\mathcal{H}} = \mathcal{H}^1 \otimes \cdots \otimes \mathcal{H}^d$ the tensor product of the RKHS $\mathcal{H}^j$. Then, by Theorem A.5 it holds that $\boldsymbol{\mathcal{H}}$ is an RKHS on $\boldsymbol{\mathcal{X}}$ with reproducing kernel $\mathbf{k}$.

The following properties will be heavily used in later calculations. They are immediate from the definitions,

(i) for all $\mathbf{x}, \mathbf{y} \in \boldsymbol{\mathcal{X}}$ it holds that

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = k^1(x^1, y^1) \cdots k^d(x^d, y^d)$$

(ii) for all $\mathbf{f} = f^1 \otimes \cdots \otimes f^d \in \boldsymbol{\mathcal{H}}$ and for all $\mathbf{g} = g^1 \otimes \cdots \otimes g^d \in \boldsymbol{\mathcal{H}}$ it holds that

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\boldsymbol{\mathcal{H}}} = \langle f^1, g^1 \rangle_{\mathcal{H}^1} \cdots \langle f^d, g^d \rangle_{\mathcal{H}^d}$$

(iii) for all $\mathbf{f} = f^1 \otimes \cdots \otimes f^d \in \boldsymbol{\mathcal{H}}$ it holds that

$$\|\mathbf{f}\|_{\boldsymbol{\mathcal{H}}} = \|f^1\|_{\mathcal{H}^1} \cdots \|f^d\|_{\mathcal{H}^d}.$$

Moreover, (i) immediately implies that $\mathbf{k}$ inherits the boundedness and continuity of the kernels $k^1, \ldots, k^d$.

### 2.2.4 Embedding of distributions

One of the strengths of RKHS is that we can embed complicated objects into them and use the Hilbert space structure to analyze them. Being able to express inner products as function evaluations via the reproducing property, additionally simplifies computation within an RKHS. In this work, we use this embedding technique to analyze distributions.

**Definition 2.23 (space of finite Borel measures)**
*Let $\mathcal{X}$ be a separable metric space, then define*

$$\mathcal{M}_f(\mathcal{X}) := \{\mu \mid \mu \text{ is a finite Borel measure on } \mathcal{X}\}.$$

Using the Bochner integral (see Appendix A.3) we can define an embedding of $\mathcal{M}_f(\mathcal{X})$ into an RKHS.

**Definition 2.24 (mean embedding function)**
*Let $\mathcal{X}$ be a separable metric space, let $k$ be a continuous bounded positive semi-definite kernel and let $\mathcal{H}$ be the RKHS with reproducing kernel $k$. Then, let $\Pi : \mathcal{M}_f(\mathcal{X}) \to \mathcal{H}$ be the function with the property that for all $\mu \in \mathcal{M}_f(\mathcal{X})$ it holds that*

$$\Pi(\mu) = \int_{\mathcal{X}} k(x, \cdot)\, \mu(dx).$$

*We call $\Pi$ the mean embedding (associated to $k$).*

Observe that the Bochner integral in this definition is well-defined since $k(x,\cdot)$ is continuous and bounded and therefore in particular in $\mathcal{L}^1\left(\mu, \|\cdot\|_{\mathcal{H}}\right)$.

In order to infer that two distributions are equal given that their embeddings coincide, it is necessary that the mean embedding is injective. Similar to Sriperumbudur et al. (2008) we make the following definition.

**Definition 2.25 (characteristic kernel)**
*Let $\mathcal{X}$ be a separable metric space, let $k$ be a continuous bounded positive semi-definite kernel, let $\mathcal{H}$ be the RKHS with reproducing kernel $k$ and let $\Pi : \mathcal{M}_f(\mathcal{X}) \to \mathcal{H}$ be the mean embedding. We say that $k$ is* characteristic *if $\Pi$ is injective.*

The Gaussian kernel on $\mathbb{R}^n$ is a commonly used example of a characteristic kernel.

## 2.3 U-statistics and V-statistics

U-and V-statistics play a crucial role in constructing hypothesis tests based on the Hilbert-Schmidt independence criterion. This section aims at providing a complete overview of all results necessary from this theory. Most of this section in particular everything related to U-statistics, is based on Serfling (1980). The corresponding results about V-statistics are extensions of the U-statistic results.

Throughout this section we will heavily use the following setting.

**Setting 2.26 (U-and V-statistic)**
*Let $m \in \mathbb{N}$, $q \in \{1,\dots,m\}$, $\mathcal{X}$ a metric space, $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $X : \Omega \to \mathcal{X}$ a random variable with law $\mathbb{P}^X$ and $(X_i)_{i\in\mathbb{N}}$ a sequence of iid copies of $X$, i.e., $(X_i)_{i\in\mathbb{N}} \overset{iid}{\sim} \mathbb{P}^X$.*

The sequence $(X_i)_{i\in\mathbb{N}}$ should be seen as the generating process of observations. Thus we can interpret a realization $(x_i)_{i\in\mathbb{N}} \subseteq \mathcal{X}$ of $(X_i)_{i\in\mathbb{N}}$ as one particular experimental outcome.

Furthermore, define the sets

- $\mathbf{C}_q(m) \coloneqq \{(i_1,\dots,i_q) \in \{1,\dots,m\} : i_1 < \dots < i_q\}$ (all combinations),
- $\mathbf{P}_q(m) \coloneqq \{(i_1,\dots,i_q) \in \{1,\dots,m\} : i_1,\dots,i_q \text{ distinct}\}$ (all permutations) and
- $\mathbf{M}_q(m) \coloneqq \{1,\dots,m\}^q$ (all mappings).

Observe that $|\mathbf{C}_q(m)| = \binom{m}{q}$, $|\mathbf{P}_q(m)| = (m)_q \coloneqq \frac{m!}{(m-q)!}$ and $|\mathbf{M}_q(m)| = m^q$.

Consider a measurable symmetric (i.e. invariant under any permutation of its input arguments) function $g : \mathcal{X}^q \to \mathbb{R}$. Suppose we are interested in the statistical functional

$$\theta_g \coloneqq \theta_g\left(\mathbb{P}^X\right) \coloneqq \mathbb{E}\left(g(X_1,\dots,X_q)\right). \tag{2.3}$$

To this end we define three estimators: The U-statistic

$$U_m(g) \coloneqq \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} g(X_{i_1},\dots,X_{i_q}), \tag{2.4}$$

the alternative U-statistic

$$U_m^*(g) := \frac{1}{(m)_q} \sum_{\mathbf{P}_q(m)} g(X_{i_1}, \dots, X_{i_q}), \tag{2.5}$$

and the V-statistic

$$V_m(g) := \frac{1}{m^q} \sum_{\mathbf{M}_q(m)} g(X_{i_1}, \dots, X_{i_q}). \tag{2.6}$$

Due to the symmetry of $g$ both $U_m(g)$ and $U_m^*(g)$ are unbiased estimators of the statistical functional $\theta_g$ (hence the name U-statistic). In particular it holds that $U_m(g) = U_m^*(g)$. The V-statistic $V_m(g)$ on the other hand has a bias due to the occurrence of equal indices in $\mathbf{M}_q(m)$. The function $g$ is commonly referred to as a kernel function. In order to avoid confusion, we refer to $g$ as a core function.

Both U-and V-statistics are important. For practical applications, it is generally easier to work with V-statistics because the sets $\mathbf{C}_q(m)$ and $\mathbf{P}_q(m)$ become quite complicated for $q > 2$. Whenever deriving theoretical results, however, it turns out to be easier to first consider U-statistics and then deduce the corresponding result for V-statistics.

To illustrate these definitions consider the following example.

**Example 2.27 (sample variance)**
*Let $\mathcal{X} = \mathbb{R}$ and consider a sequence of iid real-valued random variables $(X_i)_{i\in\mathbb{N}}$ and the core function $g : \mathbb{R}^2 \to \mathbb{R}$ defined for all $x, y \in \mathbb{R}$ by*

$$g(x, y) := \frac{(x+y)^2}{2} = \frac{x^2+y^2}{2} - xy.$$

*The V-statistic corresponding to $g$ satisfies*

$$\begin{aligned} V_m(g) &= \frac{1}{m^2} \sum_{i,j=1}^{m} \frac{X_i^2 + X_j^2}{2} - \frac{1}{m^2} \sum_{i,j=1}^{m} X_i X_j \\ &= \frac{1}{m} \sum_{i=1}^{m} X_i^2 - \left( \frac{1}{m} \sum_{i=1}^{m} X_i \right)^2 \\ &= \frac{1}{m} \sum_{i=1}^{m} \left( X_i - \frac{1}{m} \sum_{j=1}^{m} X_j \right)^2 . \end{aligned}$$

*Hence, the V-statistic $V_m(g)$ corresponds to the biased sample variance. Performing a* *similar calculation for the alternative U-statistic corresponding to $g$ leads to*

$$\begin{aligned}
U_m^*(g) &= \frac{1}{m(m-1)} \sum_{i \neq j} \frac{X_i^2 + X_j^2}{2} - \frac{1}{m(m-1)} \sum_{i \neq j} X_i X_j \\
&= \frac{1}{m(m-1)} \sum_{i,j=1}^{m} \frac{X_i^2 + X_j^2}{2} - \frac{1}{m(m-1)} \sum_{i,j=1}^{m} X_i X_j \\
&= \frac{1}{m-1} \sum_{i=1}^{m} X_i^2 - \frac{m}{m-1} \left( \frac{1}{m} \sum_{i=1}^{m} X_i X_j \right)^2 \\
&= \frac{1}{m-1} \sum_{i=1}^{m} \left( X_i - \frac{1}{m} \sum_{j=1}^{m} X_j \right)^2 .
\end{aligned}$$

*Thus, the alternative U-statistic $U_m^*(g)$ and consequently also the standard U-statistic $U_m(g)$ correspond to the classical unbiased sample variance.*

Given a non-symmetric core function $g$ one can always construct the symmetrized version

$$\widehat{g}(x_1, \ldots, x_q) = \frac{1}{q!} \sum_{\pi \in S_q} g(x_{\pi(1)}, \ldots, x_{\pi(q)}),$$

where $S_q$ is the set of permutations on $\{1, \ldots, q\}$.

In the following subsections we derive some useful properties of these statistics, which we will exploit later when deriving asymptotic properties of our mutual independence tests.

- Section 2.3.1: Variance of U-statistics
- Section 2.3.2: Method to analyze asymptotic properties of U-statistics
- Section 2.3.3: Consistency of U-statistics
- Section 2.3.4: Asymptotic distribution of U-statistics
- Section 2.3.5: Connection between U-statistics and V-statistics
- Section 2.3.6: Consistency of V-statistics
- Section 2.3.7: Variance of V-statistics
- Section 2.3.8: Bias of V-statistics
- Section 2.3.9: Asymptotic distribution of V-statistics
- Section 2.3.10: Resampling results for U-statistics and V-statistics

### 2.3.1 Variance of U-statistics

Following Serfling (1980) we introduce the following notation, which becomes useful when proving results about U-statistics.

Assume $g \in \mathcal{L}^1((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ is a symmetric core function. We then define for every $c \in \{1, \dots, q-1\}$ the function $g_c : \mathcal{X}^c \to \mathbb{R}$ by

$$g_c(x_1, \dots, x_c) \coloneqq \mathbb{E}\left(g(x_1, \dots, x_c, X_{c+1}, \dots, X_q)\right)$$

and $g_q \equiv g$. $g_c$ is again a symmetric core function such that for every $c \in \{1, \dots, q-1\}$, it holds that

$$g_c(x_1, \dots, x_c) = \mathbb{E}\left(g_{c+1}(x_1, \dots, x_c, X_{c+1})\right)$$

and

$$\mathbb{E}\left(g_c(X_1, \dots, X_c)\right) = \mathbb{E}\left(g(X_1, \dots, X_q)\right) = \theta_g.$$

Further define $\tilde{g} \equiv g - \theta_g$ and for all $c \in \{1, \dots, q\}$ define $\tilde{g}_c \equiv g_c - \theta_g$ to be the centered versions of the core functions.

Define for every $c \in \{1, \dots, q\}$,

$$\xi_c \coloneqq \mathrm{Var}\left(g_c(X_1, \dots, X_c)\right) = \mathbb{E}\left(\tilde{g}_c(X_1, \dots, X_c)^2\right). \tag{2.7}$$

We sometimes write $\xi_c(g)$ to make clear which core function we are talking about.

We can now state the main theorem of this section (see Serfling, 1980, Lemma A, Section 5.2.1) which was originally proved by Hoeffding (1948a).

**Theorem 2.28 (variance of a U-statistic)**
*Assume Setting 2.26 and let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a symmetric core function. The variance of $U_m(g)$ is given by*

$$\mathrm{Var}\left(U_m(g)\right) = \binom{m}{q}^{-1} \sum_{c=1}^{q} \binom{q}{c}\binom{m-q}{q-c}\xi_c. \tag{2.8}$$

**Proof** Let $(i_1, \dots, i_q), (j_1, \dots, j_q) \in \mathbf{C}_q(m)$ such that the two sequences have exactly $c$ common values. Denote by $(a_1, \dots, a_q)$ the reordering of $(i_1, \dots, i_q)$ and by $(b_1, \dots, b_q)$ the reordering of $(j_1, \dots, j_q)$ such that for all $k \in \{1, \dots, c\}$ it holds that $a_k = b_k$. Then using symmetry of $\tilde{g}$ and independence of $X_1, \dots, X_m$ it holds that

$$\begin{aligned}
&\mathbb{E}\left(\tilde{g}\left(X_{i_1}, \dots, X_{i_q}\right)\tilde{g}\left(X_{j_1}, \dots, X_{j_q}\right)\right) \\
&\quad = \mathbb{E}\left(\tilde{g}\left(X_{a_1}, \dots, X_{a_q}\right)\tilde{g}\left(X_{b_1}, \dots, X_{b_q}\right)\right) \\
&\quad = \mathbb{E}\left(\tilde{g}_c\left(X_{a_1}, \dots, X_{a_c}\right)\tilde{g}_c\left(X_{b_1}, \dots, X_{b_c}\right)\right) \\
&\quad = \mathbb{E}\left(\tilde{g}_c\left(X_{a_1}, \dots, X_{a_c}\right)^2\right) \\
&\quad = \xi_c
\end{aligned}$$

Now observe that the number of distinct ways two such sequences $(i_1, \dots, i_q)$ and $(j_1, \dots, j_q)$ can be chosen is $\binom{m}{q}\binom{q}{c}\binom{m-q}{q-c}$. Hence using

$$U_m(g) - \theta_g = \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} \tilde{g}(X_{i_1}, \dots, X_{i_q})$$

we get

$$\begin{aligned}\operatorname{Var}\left(U_m(g)\right) &= \mathbb{E}\left(\left(U_m(g)-\theta_g\right)^2\right)\\ &= \binom{m}{q}^{-2}\sum_{\mathbf{C}_q(m)}\sum_{\mathbf{C}_q(m)}\mathbb{E}\left(\tilde{g}(X_{i_1},\dots,X_{i_q})\tilde{g}(X_{j_1},\dots,X_{j_q})\right)\\ &= \binom{m}{q}^{-1}\sum_{c=1}^{q}\binom{q}{c}\binom{m-q}{q-c}\xi_c\end{aligned}$$

which completes the proof of Theorem 2.28. □

### 2.3.2 Projection of U-statistics

This section introduces the projection of a U-statistic, which is a technique to determine asymptotic results about U-statistics. Given a U-statistic $U_m(g)$ we define its projection by

$$\widehat{U}_m(g) := \sum_{j=1}^{m}\mathbb{E}\left(U_m(g)\mid X_j\right)-(m-1)\theta_g. \tag{2.9}$$

Observe that $\mathbb{E}\left(g(X_{i_1},\dots,X_{i_q})\mid X_j\right)$ is equal to $\theta_g$ if $j\notin\{i_1,\dots,i_q\}$ and it is equal to $g_1(X_j)$ otherwise. Hence we get,

$$\begin{aligned}\mathbb{E}\left(U_m(g)\mid X_j\right) &= \binom{m}{q}^{-1}\sum_{\mathbf{C}_q(m)}\mathbb{E}\left(g(X_{i_1},\dots,X_{i_q})\mid X_j\right)\\ &= \binom{m}{q}^{-1}\binom{m-1}{q}\theta_g+\binom{m}{q}^{-1}\binom{m-1}{q-1}g_1(X_j)\\ &= \frac{m-q}{m}\theta_g+\frac{q}{m}g_1(X_j).\end{aligned}$$

This results in

$$\widehat{U}_m(g)-\theta_g = \frac{q}{m}\sum_{j=1}^{m}\tilde{g}_1(X_j). \tag{2.10}$$

Therefore, we have shown that the projection of a U-statistic is a sum of iid random variables. This fact allows us to apply results such as the central limit theorem or the law of large numbers to these projections. In order to connect such results back to the original U-statistic we use the following theorem (for a stronger version see Serfling, 1980, Theorem, Section 5.3.2).

**Theorem 2.29 (second moment of $U_m(g)-\widehat{U}_m(g)$)**
*Assume Setting 2.26 and let $g\in\mathcal{L}^2((\mathbb{P}^X)^{\otimes q},|\cdot|_{\mathbb{R}})$ be a symmetric core function. Then*

$$\mathbb{E}\left[\left(U_m(g)-\widehat{U}_m(g)\right)^2\right]=\mathcal{O}\left(m^{-2}\right)$$

*as $m\to\infty$.*

**Proof** Set

$$w(x_1, \ldots, x_q) = g(x_1, \ldots, x_q) - \tilde{g}_1(x_1) - \cdots - \tilde{g}_1(x_q) - \theta_g$$

and observe that this is a symmetric core function with the property that

$$\mathbb{E}\left(w(X_1, \ldots, X_q)\right) = \mathbb{E}\left(w(X_1, \ldots, X_q) \mid X_1\right) = 0.$$

In particular, this implies $\xi_1(w) = 0$. Using that

$$\binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} \sum_{l=1}^{q} \tilde{g}_1(X_{i_l}) = \binom{m}{q}^{-1} \binom{m-1}{q-1} \sum_{l=1}^{m} \tilde{g}_1(X_l)$$

and (2.10) it follows that

$$U_m(g) - \widehat{U}_m(g) = \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} w(X_{i_1}, \ldots, X_{i_q}).$$

This means that $U_m(g) - \widehat{U}_m(g) = U_m(w)$ and we are in the setting where we can apply Theorem 2.28. This together with $\xi_1(w) = 0$ results in

$$\begin{aligned} \mathbb{E}\left[\left(U_m(g) - \widehat{U}_m(g)\right)^2\right] &= \operatorname{Var}\left(U_m(g) - \widehat{U}_m(g)\right) \\ &= \binom{m}{q}^{-1} \sum_{c=1}^{q} \binom{q}{c} \binom{m-q}{q-c} \xi_c(w) \\ &= \mathcal{O}\left(m^{-2}\right), \end{aligned}$$

which completes the proof of Theorem 2.29. □

**Remark 2.30** *The notion of projection can be generalized by defining the c-th order projection of a U-statistic $U_m(g)$ by*

$$\widehat{U}_{c,m}(g) := \sum_{\mathbf{C}_c(m)} \mathbb{E}\left(U_m(g) \mid X_{i_1}, \ldots, X_{i_c}\right) - \left(\binom{m}{c} - 1\right) \theta_g.$$

*Then using similar reasoning as above one gets*

$$\widehat{U}_{c,m}(g) = \frac{(q)_c}{(m)_c} \sum_{\mathbf{C}_c(m)} \tilde{g}_c\left(X_{i_1}, \ldots, X_{i_c}\right) + \theta_g$$

*and equivalent to Theorem 2.29 it holds that*

$$\mathbb{E}\left[\left(U_m(g) - \widehat{U}_m(g)\right)^2\right] = \mathcal{O}\left(m^{-(c+1)}\right)$$

*as $m \to \infty$.*

### 2.3.3 Consistency of U-statistics

The following theorem is a weaker version of Serfling (1980, Theorem A, Section 5.4).

**Theorem 2.31 (consistency of a U-statistic)**
*Assume Setting 2.26 and let $g \in \mathcal{L}^1((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a symmetric core function. Then*

$$U_m(g) \xrightarrow{\mathbb{P}} \theta_g$$

*as $m \to \infty$.*

**Proof** To prove this statement we make use of the projections introduced in Section 2.3.2. Note that by (2.10) $\widehat{U}_m(g)$ is the sum of iid random variables and therefore we can apply the weak law of large numbers to get,

$$\widehat{U}_m(g) \xrightarrow{\mathbb{P}} \theta_g$$

as $m \to \infty$. We then use Theorem 2.29 to get

$$\left(\widehat{U}_m(g) - U_m(g)\right) \xrightarrow{L^2} 0$$

as $m \to \infty$. Now since $L^2$-convergence also implies convergence in probability we get

$$U_m(g) = \widehat{U}_m(g) + \left(U_m(g) - \widehat{U}_m(g)\right) \xrightarrow{\mathbb{P}} \theta_g$$

as $m \to \infty$ which completes the proof of Theorem 2.31. □

### 2.3.4 Asymptotic distributions of U-statistics

A U-statistic is called degenerate if $\xi_1 = \operatorname{Var}(g_1(X_1)) = 0$ and non-degenerate if $\xi_1 > 0$. In this section we analyze the asymptotic distribution of

- $\sqrt{m}U_m(g)$ for the non-degenerate case ($\xi_1 > 0$) and
- $mU_m(g)$ for a special degenerate case ($\xi_1 = 0$, $\xi_2 > 0$).

In the degenerate case the asymptotic distribution depends on the eigenvalues of a particular integral operator. In order to avoid repeating its definition we introduce the following setting.

**Setting 2.32 (degenerate asymptotic)**
*Let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a core function, let $(Z_j)_{j\in\mathbb{N}}$ be a sequence of independent standard normal random variables on $\mathbb{R}$, let $T_{\tilde{g}_2} \in L\left(L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})\right)$ with the property that for every $f \in L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})$ and for every $x \in \mathcal{X}$ it holds that*

$$\left(T_{\tilde{g}_2}(f)\right)(x) = \int_{\mathcal{X}} \tilde{g}_2(x,y) f(y)\, \mathbb{P}^X(dy) \tag{2.11}$$

*and let $(\lambda_j)_{j\in\mathbb{N}}$ be the eigenvalues of $T_{\tilde{g}_2}$.*

### Non-degenerate case

The following theorem is taken from Serfling (1980, Theorem A, Section 5.5.1).

**Theorem 2.33 (asymptotic distribution of a U-statistic (non-degenerate))** *Assume Setting 2.26, let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a symmetric core function and assume $\xi_1 > 0$. Then it holds that*

$$\sqrt{m}\left(U_m(g) - \theta_g\right) \xrightarrow{d} \mathcal{N}\left(0, q^2\xi_1\right) \tag{2.12}$$

*as $m \to \infty$.*

**Proof** To prove this statement we make use of the projections introduced in Section 2.3.2. Note that by (2.10), $\widehat{U}_m(g)$ is the sum of iid random variables and therefore we applying the central limit theorem together with $\mathrm{Var}(q\tilde{g}_1(X_1)) = q^2\xi_1$ we get,

$$\sqrt{m}\left(\widehat{U}_m(g) - \theta_g\right) \xrightarrow{d} \mathcal{N}\left(0, q^2\xi_1\right)$$

as $m \to \infty$.

Hence, it only remains to show that $\sqrt{m}\widehat{U}_m(g)$ and $\sqrt{m}U_m(g)$ have the same limiting distribution. By Theorem 2.29 it holds that

$$\sqrt{m}\left(\widehat{U}_m(g) - U_m(g)\right) \xrightarrow{L^2} 0$$

as $m \to \infty$. Now since $L^2$-convergence also implies convergence in probability we may apply Slutsky's theorem to get that

$$\sqrt{m}\left(U_m(g) - \theta_g\right) = \sqrt{m}\left(\widehat{U}_m(g) - \theta_g\right) + \sqrt{m}\left(\widehat{U}_m(g) - U_m(g)\right) \xrightarrow{d} \mathcal{N}\left(0, q^2\xi_1\right)$$

as $m \to \infty$, which completes the proof of Theorem 2.33. □

### Degenerate case

The following theorem is taken from Serfling (1980, Theorem, Section 5.5.2). The proof is slightly modified using a different version Mercer's Theorem.

**Theorem 2.34 (asymptotic distribution of a U-statistic (degenerate))** *Assume Setting 2.26 and Setting 2.32, let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a symmetric core function and assume $0 = \xi_1 < \xi_2$. Then it holds that*

$$m\left(U_m(g) - \theta_g\right) \xrightarrow{d} \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j \left(Z_j^2 - 1\right)$$

*as $m \to \infty$.*

**Proof** Using Remark 2.30 it is enough to show the result for the second order projection of $\widehat{U}_m(g)$. So for $(Z_j)_{j\in\mathbb{N}}$ iid standard normal random variables defined on the same probability space and $Y = \sum_{j=1}^{\infty} \lambda_j(Z_j^2 - 1)$ we want to show

$$m\left(\widehat{U}_{2,m}(g) - \theta_g\right) \xrightarrow{d} \frac{q(q-1)}{2}Y$$

as $m \to \infty$.

Begin by setting

$$T_m = \frac{1}{m}\sum_{\mathbf{P}_2(m)} \tilde{g}_2\left(X_{i_1}, X_{i_2}\right)$$

and observe that

$$m\left(\widehat{U}_{2,m}(g) - \theta_g\right) = \frac{q(q-1)}{2}\frac{m}{m-1}T_m.$$

We have therefore reduced the problem to proving

$$T_m \xrightarrow{d} Y$$

as $m \to \infty$. We do this by proving that for all $x \in \mathbb{R}$ it holds that

$$\mathbb{E}\left(e^{ixT_m}\right) \longrightarrow \mathbb{E}\left(e^{ixY}\right) \tag{2.13}$$

as $m \to \infty$ and then applying the continuity theorem for characteristic functions.

Begin by applying Mercer's Theorem A.1 to $\tilde{g}_2$ to get an orthonormal sequence of eigenfunctions $(\varphi_j)_{j\in\mathbb{N}}$ and corresponding eigenvalues $(\lambda_j)_{j\in\mathbb{N}}$ of $T_{\tilde{g}_2}$ (see (2.11)) satisfying for $x, y \in \text{supp}(\mathbb{P}^X)$ that

$$\tilde{g}_2(x,y) = \sum_{j=1}^{\infty} \lambda_j\varphi_j(x)\varphi_j(y) \tag{2.14}$$

converges uniformly. Hence, using that we can exchange sum and expectation due to the uniform convergence we get for all $x \in \text{supp}(\mathbb{P}^X)$ that

$$\begin{aligned}\tilde{g}_1(x) &= \mathbb{E}\left(\tilde{g}_2(x, X_1)\right)\\ &= \mathbb{E}\left(\sum_{j=1}^{\infty}\lambda_j\varphi_j(x)\varphi_j(X_1)\right)\\ &= \sum_{j=1}^{\infty}\lambda_j\varphi_j(x)\mathbb{E}\left(\varphi_j(X_1)\right) \qquad (2.15)\end{aligned}$$

where the sum again converges uniformly. Since $\xi_1(g) = 0$ this implies $\tilde{g}_1(X_1) = 0$ $\mathbb{P}$-a.s. and thus for all $j \in \mathbb{N}$ that

$$\mathbb{E}\left(\varphi_j(X_1)\right) = 0. \tag{2.16}$$

Next using that $(\varphi_j)_{j\in\mathbb{N}}$ are orthonormal and again using that we can exchange sum and expectation we get

$$\begin{aligned}\mathbb{E}\left(\tilde{g}_2(X_1,X_2)^2\right) &= \mathbb{E}\left(\sum_{j=1}^{\infty}\sum_{i=1}^{\infty}\lambda_j\lambda_i\varphi_j(X_1)\varphi_j(X_2)\varphi_i(X_1)\varphi_i(X_2)\right)\\ &= \sum_{j=1}^{\infty}\sum_{i=1}^{\infty}\lambda_j\lambda_i\mathbb{E}\left(\varphi_j(X_1)\varphi_i(X_1)\right)\mathbb{E}\left(\varphi_j(X_2)\varphi_i(X_2)\right)\\ &= \sum_{j=1}^{\infty}\lambda_j^2. \qquad (2.17)\end{aligned}$$

This in particular implies $\sum_{j=1}^{\infty}\lambda_j^2 = \mathbb{E}\left(\tilde{g}_2(X_1,X_2)^2\right) < \infty$. Using the expansion (2.14) it holds that

$$T_m = \frac{1}{m}\sum_{\mathbf{P}_2(m)}\sum_{j=1}^{\infty}\lambda_j\varphi_j(X_{i_1})\varphi_j(X_{i_2}).$$

Next, define

$$T_{m,K} = \frac{1}{m}\sum_{\mathbf{P}_2(m)}\sum_{j=1}^{K}\lambda_j\varphi_j(X_{i_1})\varphi_j(X_{i_2}).$$

and

$$Y_K = \sum_{j=1}^{K}\lambda_j\left(Z_j^2-1\right).$$

The rest of the proof will be separated into four parts. The first part deals with the approximation of $T_m$ by $T_{m,K}$, the second part deals with the approximation of $Y_K$ by $T_{m,K}$, the third part deals with the approximation of $Y$ by $Y_K$ and the fourth part combines these results to conclude the proof.

**Part 1:**

Using Jensen's inequality and the inequality $|e^{iz}-1|\leq|z|$, we get

$$\begin{aligned}\left|\mathbb{E}\left(e^{ixT_m}\right)-\mathbb{E}\left(e^{ixT_{m,K}}\right)\right| &\leq \mathbb{E}\left|e^{ixT_m}-e^{ixT_{m,K}}\right|\\ &\leq |x|\mathbb{E}|T_m-T_{m,K}|\\ &\leq |x|\left[\mathbb{E}\left((T_m-T_{m,K})^2\right)\right]^{\frac{1}{2}}. \qquad (2.18)\end{aligned}$$

Due to the uniform convergence we can set

$$w^K(x,y) = \sum_{j=K+1}^{\infty}\lambda_j\varphi_j(x)\varphi_j(y)$$

and observe that $T_m - T_{m,K}$ can be written in terms of a U-statistic as

$$T_m - T_{m,K} = \frac{2}{m}\binom{m}{2}U_m(w^K). \qquad (2.19)$$

By uniform convergence and (2.16) it holds that $\mathbb{E}\left(w^K(X_1, X_2)\right) = 0$ which in turn implies that

$$\mathbb{E}\left(U_m(w^K)\right) = 0. \tag{2.20}$$

Using this and again making use of uniform convergence it also holds that

$$\xi_2(w^K) = \mathbb{E}\left(w^K(X_1, X_2)^2\right) = \sum_{j=K+1}^{\infty} \lambda_j^2. \tag{2.21}$$

A similar argument together with (2.16) leads to $\mathbb{E}\left(w^k(\cdot, X_1)\right) \equiv 0$ and hence

$$\xi_1(w^K) = \mathbb{E}_{X_1}\left[\left(\mathbb{E}_{X_2}\left(w^k(X_1, X_2)\right)\right)^2\right] = 0. \tag{2.22}$$

We can now apply Theorem 2.28 together with (2.20), (2.21) and (2.22) to get

$$\mathbb{E}\left(U_m(w^K)\right) = \binom{m}{2}^{-1} \sum_{j=K+1}^{\infty} \lambda_j^2$$

and combined with (2.19) this results in

$$\mathbb{E}\left((T_m - T_{m,K})^2\right) = \frac{2(m-1)}{m} \sum_{j=K+1}^{\infty} \lambda_j^2 \leq 2 \sum_{j=K+1}^{\infty} \lambda_j^2. \tag{2.23}$$

Now, for arbitrary $\varepsilon > 0$ and $x \in \mathbb{R}$ we can choose $K_0(\varepsilon, x) \in \mathbb{N}$ large enough such that for all $K > K_0(\varepsilon, x)$ it holds that

$$|x| \left(2 \sum_{j=K+1}^{\infty} \lambda_j^2\right)^{\frac{1}{2}} < \varepsilon \tag{2.24}$$

and therefore together with (2.18) and (2.23) it holds for all $m \in \mathbb{N}$ and all $K > K_0(\varepsilon, x)$ that

$$|\mathbb{E}\left(e^{ixT_m}\right) - \mathbb{E}\left(e^{ixT_{m,K}}\right)| < \varepsilon. \tag{2.25}$$

**Part 2:**

For this part set

$$Z_{jm} = m^{-\frac{1}{2}} \sum_{i=1}^{m} \varphi_j(X_i)$$

and

$$V_{jm} = m^{-1} \sum_{i=1}^{m} \varphi_j(X_i)^2.$$

Next, observe that we can write

$$T_{m,K} = \sum_{j=1}^{K} \lambda_j \left( Z_{jm}^2 - V_{jm} \right) \tag{2.26}$$

Using (2.16) it holds for all $j, m \in \mathbb{N}$ that

$$\mathbb{E}\left(Z_{jm}\right) = 0 \tag{2.27}$$

and hence for all $j, l, m \in \mathbb{N}$ it holds that

$$\operatorname{Cov}\left(Z_{jm}, Z_{lm}\right) = m^{-1} \sum_{\mathbf{M}_2(m)} \mathbb{E}\left(\varphi_j(X_{i_1})\varphi_l(X_{i_2})\right) = \begin{cases} 1, & \text{if } j = l \\ 0, & \text{if } j \neq l \end{cases} \tag{2.28}$$

Therefore, using the multidimensional version of the central limit theorem, (2.27) and (2.28) it holds that

$$\left(Z_{1m}, \ldots, Z_{Km}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Id}_{K \times K}\right) \tag{2.29}$$

as $m \to \infty$. Furthermore, since for all $j \in \mathbb{N}$ it holds that $\mathbb{E}\left(\varphi_j(X_1)\right) = 1$, applying the weak law of large numbers yields

$$\left(V_{1m}, \ldots, V_{Km}\right) \xrightarrow{\mathbb{P}} (1, \ldots, 1) \tag{2.30}$$

as $m \to \infty$. Combining (2.26), (2.29) and (2.30) together with Slutsky's theorem leads to

$$T_{m,K} \xrightarrow{d} Y_K$$

as $m \to \infty$. Hence, for arbitrary $\varepsilon > 0$ and $x \in \mathbb{R}$ we can choose $m_0(\varepsilon, x) \in \mathbb{N}$ large enough such that for all $m > m_0(\varepsilon, x)$ it holds that

$$|\mathbb{E}\left(e^{ixT_{m,K}}\right) - \mathbb{E}\left(e^{ixY_K}\right)| < \varepsilon. \tag{2.31}$$

**Part 3:**

By using Jensen inequality, the inequality $|e^{iz} - 1| \leq |z|$ and the independence of $(Z_j)_{j \in \mathbb{N}}$ it holds for any $x \in \mathbb{R}$ that

$$\begin{aligned} |\mathbb{E}\left(e^{ixY}\right) - \mathbb{E}\left(e^{ixY_K}\right)| &\leq \mathbb{E}|e^{ixY} - e^{ixY_K}| \\ &\leq |x| \left[\mathbb{E}\left((Y - Y_K)^2\right)\right]^{\frac{1}{2}} \\ &\leq |x| \left[\mathbb{E}\left((Z_1 - 1)^2\right)\right]^{\frac{1}{2}} \left( \sum_{j=K+1}^{\infty} \lambda_j^2 \right)^{\frac{1}{2}}. \end{aligned}$$

So for arbitrary $\varepsilon > 0$ and $x \in \mathbb{R}$ we can choose $K_1(\varepsilon, x) \in \mathbb{N}$ large enough such that for all $K > K_1(\varepsilon, x)$ it holds that

$$|\mathbb{E}\left(e^{ixY}\right) - \mathbb{E}\left(e^{ixY_K}\right)| < \varepsilon. \tag{2.32}$$

**Part 4:**

Finally combining (2.25), (2.31) and (2.32) it holds for all $\varepsilon > 0$ and all $x \in \mathbb{R}$ that for all $m > m_0(\varepsilon, x)$ and all $K > \max\{K_0(\varepsilon, x), K_1(\varepsilon, x)\}$ it holds that

$$\begin{aligned} \left|\mathbb{E}\left(e^{ixY}\right) - \mathbb{E}\left(e^{ixT_m}\right)\right| &\leq \left|\mathbb{E}\left(e^{ixY}\right) - \mathbb{E}\left(e^{ixY_K}\right)\right| + \left|\mathbb{E}\left(e^{ixY_k}\right) - \mathbb{E}\left(e^{ixT_{mK}}\right)\right| \\ &\quad + \left|\mathbb{E}\left(e^{ixT_{m,K}}\right) - \mathbb{E}\left(e^{ixT_m}\right)\right| \\ &< 3\varepsilon \end{aligned}$$

This completes the proof of Theorem 2.34. □

### 2.3.5 Connection between U-statistics and V-statistics

To derive the asymptotic distribution of V-statistics we show that V-statistics are in an appropriate sense good approximations of U-statistics. In order to show results of this type we require some kind of boundedness condition on the core function. The next definition introduces such a condition.

**Definition 2.35 (total boundedness condition)**
*Let $r \in \mathbb{N}$, assume Setting 2.26 and let $g \in \mathcal{L}^r((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a core function. Then we say that $g$ satisfies the total boundedness condition of order $r$ if for all $(i_1, \dots, i_q) \in \mathbf{M}_q(q)$ it holds that*

$$\mathbb{E}\left[|g(X_{i_1}, \dots, X_{i_q})|^r\right] < \infty.$$

In particular, this condition is fulfilled if the core function $g$ is a bounded function.

The following result is due to Serfling (1980, Lemma, Section 5.7.3).

**Lemma 2.36 (connection between U-and V-statistics)**
*Let $r \in \mathbb{N}$, assume Setting 2.26 and let $g \in \mathcal{L}^r((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a core function satisfying the total boundedness condition of order $r$. Then it holds that*

$$\mathbb{E}\left[|U_m(g) - V_m(g)|^r\right] = \mathcal{O}\left(m^{-r}\right)$$

*as $m \to \infty$.*

**Proof** Set

$$W_m(g) = \frac{1}{m^q - (m)_q} \sum_{\mathbf{M}_q(m) \setminus \mathbf{P}_q(m)} g(X_{i_1}, \dots, X_{i_q})$$

and observe that

$$\begin{aligned} m^q V_m(g) &= \sum_{\mathbf{M}_q(m)} g(X_{i_1}, \dots, X_{i_q}) \\ &= \sum_{\mathbf{P}_q(m)} g(X_{i_1}, \dots, X_{i_q}) + \sum_{\mathbf{M}_q(m) \setminus \mathbf{P}_q(m)} g(X_{i_1}, \dots, X_{i_q}) \\ &= (m)_q U_m^*(g) + \left(m^q - (m)_q\right) W_m(g) \\ &= (m)_q U_m(g) + \left(m^q - (m)_q\right) W_m(g). \end{aligned}$$

Therefore it holds that,

$$\begin{aligned} m^q \left(V_m(g) - U_m(g)\right) &= (m)_q U_m(g) + \left(m^q - (m)_q\right) W_m(g) - m^q U_m(g) \\ &= \left(m^q - (m)_q\right)\left(W_m(g) - U_m(g)\right). \end{aligned}$$

Next combining this result with Minkowski's inequality leads to

$$\begin{aligned} \mathbb{E}\left[|V_m(g) - U_m(g)|^r\right] &= \left(\frac{m^q - (m)_q}{m^q}\right)^r \mathbb{E}\left[|W_m(g) - U_m(g)|^r\right] \\ &\leq \left(\frac{m^q - (m)_q}{m^q}\right)^r \left[\left(\mathbb{E}\left[|W_m(g)|^r\right]\right)^{\frac{1}{r}} + \left(\mathbb{E}\left[|U_m(g)|^r\right]\right)^{\frac{1}{r}}\right]^r. \end{aligned}$$

Now noting that $\left(m^q - (m)_q\right) = \mathcal{O}\left(m^{q-1}\right)$ as $m \to \infty$ and using that $g$ satisfies the total boundedness condition we end up with

$$\mathbb{E}\left[|V_m(g) - U_m(g)|^r\right] = \mathcal{O}\left(m^{-r}\right),$$

which completes the proof of Lemma 2.36. □

In order to prove some of the asymptotic statements of V-statistics we require a stronger way of comparing V-statistics with U-statistics than that given in Lemma 2.36. For example, when computing the asymptotic variance of a V-statistic up to an order of $m^{-2}$ by comparison with the variance of a U-statistic, we need to estimate the second moment of the difference to an order of $m^{-(2+\varepsilon)}$. Hence, the result in Lemma 2.36 is not sufficient. The following technical lemma gives a decomposition of a V-statistic into the corresponding U-statistic and some remainder terms. We are not aware of a similar result in literature.

**Lemma 2.37 (decomposition of a V-statistic)**
*Assume Setting 2.26 and let $g \in \mathcal{L}^1((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a core function. For all $k \in \{1, \dots, q-1\}$, $l \in \{k+1, \dots, q\}$ let $\pi^{kl} : \{1, \dots, q\} \to \{1, \dots, q-1\}$ be the unique surjective functions with the property that $\pi^{kl}(k) = \pi^{kl}(l) = 1$ and for all $i, j \in \{1, \dots, q\} \setminus \{k, l\}$ with $i < j$ it holds that $\pi^{kl}(i) < \pi^{kl}(j)$. Define for all $x_1, \dots, x_{q-1} \in \mathcal{X}$ the function*

$$w(x_1, \dots, x_{q-1}) \coloneqq \sum_{k=1}^{q-1} \sum_{l=k+1}^{q} g(x_{\pi^{kl}(1)}, \dots, x_{\pi^{kl}(q)}).$$

*and set $B \coloneqq \{(i_1, \dots, i_q) \in \mathbf{M}_q(m) \mid$ at most $q-2$ distinct values$\}$. Then it holds that*

$$\begin{aligned} mV_m(g) = &\left(1 + \mathcal{O}\left(m^{-1}\right)\right) U_m(w) \\ &+ \left(1 + \mathcal{O}\left(m^{-1}\right)\right) \frac{1}{(m)_{q-1}} \sum_{B} g(X_{i_1}, \dots, X_{i_q}) \\ &- \left(\binom{q}{2} + \mathcal{O}\left(m^{-1}\right)\right) U_m(g) \\ &+ mU_m(g) \end{aligned}$$

*and $|B| = \mathcal{O}\left(m^{q-2}\right)$ as $m \to \infty$.*

**Proof** We begin by introducing

$$S_m = \frac{1}{(m)_{q-1}} \left( \sum_{\mathbf{M}_q(m)} g(X_{i_1}, \ldots, X_{i_q}) - \sum_{\mathbf{P}_q(m)} g(X_{i_1}, \ldots, X_{i_q}) \right)$$

and

$$A = \{(i_1, \ldots, i_q) \in \mathbf{M}_q(m) \mid \text{ at most } q-1 \text{ distinct values}\}.$$

Then, observe that $A = \mathbf{M}_q(m) \setminus \mathbf{P}_q(m)$ and

$$\begin{aligned} A \setminus B &= \{(i_1, \ldots, i_q) \in \mathbf{M}_q(m) \mid \text{ exactly } q-1 \text{ distinct values}\} \\ &= \Bigg\{ \left( i_{\pi^{kl}(1)}, \ldots, i_{\pi^{kl}(q)} \right) \Bigg| (i_1, \ldots, i_{q-1}) \in \mathbf{P}_{q-1}(m), \\ &\qquad\qquad k \in \{1, \ldots, q-1\},\ l \in \{k+1, \ldots, q\} \Bigg\}. \end{aligned}$$

Therefore, it holds that $|A| = m^q - (m)_q$ and $|A \setminus B| = \frac{q(q-1)}{2}(m)_{q-1}$. Using this we get

$$\begin{aligned} |B| &= |A| - |A \setminus B| \\ &= m^q - (m)_q - \frac{q(q-1)}{2}(m)_{q-1} \\ &= m^q - m(m-1)\cdots(m-(q-1)) - \frac{q(q-1)}{2} m(m-1)\cdots(m-(q-2)) \\ &= m^q - m^q + \frac{q(q-1)}{2} m^{q-1} + \mathcal{O}\left(m^{q-2}\right) - \frac{q(q-1)}{2} m^{q-1} + \mathcal{O}\left(m^{q-2}\right) \\ &= \mathcal{O}\left(m^{q-2}\right) \end{aligned}$$

as $m \to \infty$. We can now make the following calculation

$$\begin{aligned} S_m &= \frac{1}{(m)_{q-1}} \left( \sum_{\mathbf{M}_q(m)} g(X_{i_1}, \ldots, X_{i_q}) - \sum_{\mathbf{P}_q(m)} g(X_{i_1}, \ldots, X_{i_q}) \right) \\ &= \frac{1}{(m)_{q-1}} \sum_A g(X_{i_1}, \ldots, X_{i_q}) \\ &= \frac{1}{(m)_{q-1}} \sum_{\mathbf{P}_{q-1}(m)} w(X_{i_1}, \ldots, X_{i_{q-1}}) + \frac{1}{(m)_{q-1}} \sum_B g(X_{i_1}, \ldots, X_{i_q}) \\ &= U_m^*(w) + \frac{1}{(m)_{q-1}} \sum_B g(X_{i_1}, \ldots, X_{i_q}). \end{aligned} \tag{2.33}$$

Finally, we can decompose $mV_m(g)$ as follows

$$\begin{aligned} mV_m(g) &= \frac{1}{m^{q-1}} \sum_{\mathbf{M}_q(m)} g(X_{i_1}, \dots, X_{i_q}) - \frac{1}{m^{q-1}} \sum_{\mathbf{P}_q(m)} g(X_{i_1}, \dots, X_{i_q}) \\ &\quad + \frac{1}{m^{q-1}} \sum_{\mathbf{P}_q(m)} g(X_{i_1}, \dots, X_{i_q}) \\ &= \frac{(m)_{q-1}}{m^{q-1}} S_m + \frac{(m)_q}{m^{q-1}} U_m^*(g) \\ &= \left(1 + \mathcal{O}\left(m^{-1}\right)\right) S_m + \left(m - \frac{q(q-1)}{2} + \mathcal{O}\left(m^{-1}\right)\right) U_m(g) \\ &= \left(1 + \mathcal{O}\left(m^{-1}\right)\right) S_m - \left(\binom{q}{2} + \mathcal{O}\left(m^{-1}\right)\right) U_m(g) + mU_m(g). \end{aligned} \tag{2.34}$$

Combining (2.33) and (2.34) completes the proof of Lemma 2.37. □

### 2.3.6 Consistency of V-statistics

The following theorem is the counterpart of Theorem 2.31 for V-statistics. The proof is a direct application of Lemma 2.36 and Theorem 2.31.

**Theorem 2.38 (consistency of a V-statistic)**
*Assume Setting 2.26 and let $g \in \mathcal{L}^1((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a symmetric core function satisfying the total boundedness condition of order 1. Then*

$$V_m(g) \xrightarrow{\mathbb{P}} \theta_g$$

*as $m \to \infty$.*

**Proof** By Theorem 2.31 it holds that

$$U_m(g) \xrightarrow{\mathbb{P}} \theta_g$$

as $m \to \infty$. Furthermore, by Lemma 2.36 we have that

$$\mathbb{E}|U_m(g) - V_m(g)| = \mathcal{O}\left(m^{-1}\right)$$

as $m \to \infty$. Since convergence in $L^1$ implies convergence in probability we obtain

$$V_m(g) \xrightarrow{\mathbb{P}} \theta_g$$

as $m \to \infty$, which completes the proof of Theorem 2.38. □

### 2.3.7 Variance of V-statistics

In the degenerate setting $\xi_1 = 0$, Lemma 2.37 allows us to show that the variance of a V-statistic is equal to that of a U-statistic up to a certain order of $m$. Its proof relies on Lemma 2.37.

**Theorem 2.39 (asymptotic variance of a V-statistic)**
*Assume Setting 2.26 and let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a bounded core function satisfying $\xi_1 = 0$. Then it holds that*

$$\operatorname{Var}\left(V_m(g)\right) = \binom{m}{q}^{-1} \binom{q}{2} \binom{m-q}{q-2} \xi_2 + \mathcal{O}\left(m^{-\frac{5}{2}}\right).$$

*as $m \to \infty$.*

**Proof** It holds that

$$\operatorname{Var}\left(V_m(g)\right) = \operatorname{Var}\left(V_m(\tilde{g})\right),$$

which implies that without loss of generality we can assume that $\theta_g = 0$. By Lemma 2.37 we get that

$$\begin{aligned} mV_m(g) = &\left(1 + \mathcal{O}\left(m^{-1}\right)\right) S_m \\ &- \left(\binom{q}{2} + \mathcal{O}\left(m^{-1}\right) - m\right) U_m(g) \end{aligned} \tag{2.35}$$

as $m \to \infty$, where $S_m = U_m(w) + \frac{1}{(m)_{q-1}} \sum_B g(X_{i_1}, \dots, X_{i_q})$. Applying Theorem 2.28 results in

$$\operatorname{Var}\left(U_m(g)\right) = \binom{m}{q}^{-1} \binom{q}{2} \binom{m-q}{q-2} \xi_2 + \mathcal{O}\left(m^{-3}\right) \tag{2.36}$$

and

$$\operatorname{Var}\left(U_m(w)\right) = \mathcal{O}\left(m^{-1}\right). \tag{2.37}$$

Moreover, using that $g$ is bounded it holds that

$$\begin{aligned} &\operatorname{Var}\left(\frac{1}{(m)_{q-1}} \sum_B g(X_{i_1}, \dots, X_{i_q})\right) \\ &\leq \frac{1}{(m)_{q-1}^2} \mathbb{E}\left(\left|\sum_B g(X_{i_1}, \dots, X_{i_q})\right|^2\right) \\ &\leq \frac{1}{(m)_{q-1}^2} \sum_{(i_1,\dots,i_q)\in B} \sum_{(j_1,\dots,j_q)\in B} \mathbb{E}\left(\left|g(X_{i_1}, \dots, X_{j_q}) g(X_{i_1}, \dots, X_{j_q})\right|\right) \\ &\leq \frac{C|B|^2}{(m)_{q-1}^2} = \mathcal{O}\left(m^{-2}\right). \end{aligned} \tag{2.38}$$

So combining (2.37) and (2.38) shows that

$$\operatorname{Var}\left(S_m\right) = \mathcal{O}\left(m^{-1}\right) \tag{2.39}$$

and

$$\operatorname{Cov}\left(U_m(g), S_m\right) \leq \left(\operatorname{Var}\left(U_m(g)\right)\operatorname{Var}\left(S_m\right)\right)^{\frac{1}{2}} = \mathcal{O}\left(m^{-\frac{3}{2}}\right). \tag{2.40}$$

Finally, use (2.35), (2.36), (2.39) and (2.40) to get

$$\begin{aligned}\operatorname{Var}\left(mV_m(g)\right) &= \left(1+\mathcal{O}\left(m^{-1}\right)\right)^2 \operatorname{Var}\left(S_m\right)\\ &\quad + \left(\tbinom{q}{2} + \mathcal{O}\left(m^{-1}\right) - m\right)^2 \operatorname{Var}\left(U_m(g)\right)\\ &\quad - 2\left(1+\mathcal{O}\left(m^{-1}\right)\right)\left(m + \tbinom{q}{2} + \mathcal{O}\left(m^{-1}\right)\right)\operatorname{Cov}\left(U_m(g), S_m\right)\\ &= \mathcal{O}\left(1\right)\operatorname{Var}\left(S_m\right) + \left(m^2 + \mathcal{O}\left(m\right)\right)\operatorname{Var}\left(U_m(g)\right) + \mathcal{O}\left(m\right)\operatorname{Cov}\left(U_m(g), S_m\right)\\ &= m^2\binom{m}{q}^{-1}\binom{q}{2}\binom{m-q}{q-2}\xi_2 + \mathcal{O}\left(m^{-\frac{1}{2}}\right).\end{aligned}$$

Dividing by $m^2$ completes the proof of Theorem 2.39. □

### 2.3.8 Bias of V-statistics

As a further consequence of Lemma 2.37 the bias of a V-statistic can be explicitly expressed up to order $m^{-2}$.

**Theorem 2.40 (bias of a V-statistic)**
*Assume Setting 2.26 and let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a core function satisfying the total boundedness condition of order 2. Then it holds that*

$$\mathbb{E}\left(V_m(g) - \theta_g\right) = \frac{1}{m}\binom{q}{2}\mathbb{E}\left(\tilde{g}_2(X_1, X_1)\right) + \mathcal{O}\left(m^{-2}\right)$$

*as $m \to \infty$.*

**Proof** We use Lemma 2.37 to get that

$$\begin{aligned}mV_m(\tilde{g}) &= \left(1+\mathcal{O}\left(m^{-1}\right)\right)U_m(w)\\ &\quad + \left(1+\mathcal{O}\left(m^{-1}\right)\right)\tfrac{1}{(m)_{q-1}}\sum_{B}\tilde{g}(X_{i_1},\ldots,X_{i_q})\\ &\quad - \left(\binom{q}{2} - m + \mathcal{O}\left(m^{-1}\right)\right)U_m(\tilde{g}).\end{aligned} \tag{2.41}$$

Moreover, using the total boundedness condition of $g$ we can get a constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}\left|\frac{1}{(m)_{q-1}}\sum_{B}\tilde{g}(X_{i_1},\ldots,X_{i_q})\right| &\leq \frac{1}{(m)_{q-1}}\sum_{B}\mathbb{E}\left|\tilde{g}(X_{i_1},\ldots,X_{i_q})\right|\\ &\leq C\frac{|B|}{(m)_{q-1}}\\ &= \mathcal{O}\left(m^{-1}\right)\end{aligned} \tag{2.42}$$

as $m \to \infty$. Hence, using (2.41),(2.42) and the unbiasedness of U-statistics results in

$$\mathbb{E}\left(m\left(V_m(g) - \theta_g\right)\right) = \mathbb{E}\left(mV_m(\tilde{g})\right) = \theta_w + \mathcal{O}\left(m^{-1}\right). \tag{2.43}$$

We can compute $\theta_w$ by using the symmetry of $\tilde{g}$ to get

$$\theta_w = \mathbb{E}\left(w(X_1, \dots, X_{q-1})\right) = \binom{q}{2}\mathbb{E}\left(\tilde{g}_2(X_1, X_1)\right). \tag{2.44}$$

Finally, combining (2.43) and (2.44) and dividing by $m$ concludes the proof of Theorem 2.40. □

### 2.3.9 Asymptotic distribution of V-statistics

In this section we derive the asymptotic distributions for V-statistics based on the results from Section 2.3.4.

#### Non-degenerate case

The following theorem is the counterpart of Theorem 2.33 for V-statistics. The proof is a straightforward application of both Lemma 2.36 and Theorem 2.33.

**Theorem 2.41 (asymptotic distribution of a V-statistic (non-degenerate))** *Assume Setting 2.26, let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a core function satisfying the total boundedness condition of order* 2 *and assume $\xi_1 > 0$. Then it holds that*

$$\sqrt{m}\left(V_m(g) - \theta_g\right) \xrightarrow{d} \mathcal{N}\left(0, q^2\xi_1\right) \tag{2.45}$$

*as $m \to \infty$.*

**Proof** Since convergence in $L^2$ implies convergence in probability Lemma 2.36 in particular shows that

$$\sqrt{m}\left(V_m(g) - U_m(g)\right) \xrightarrow{\mathbb{P}} 0$$

as $m \to \infty$. Combining this with Theorem 2.33 and Slutsky's theorem we get

$$\sqrt{m}\left(V_m(g) - \theta_g\right) = \sqrt{m}\left(U_m(g) - \theta_g\right) + \sqrt{m}\left(V_m(g) - U_m(g)\right) \xrightarrow{d} \mathcal{N}\left(0, q^2\xi_1\right)$$

as $m \to \infty$ which completes the proof of Theorem 2.41. □

#### Degenerate case

Theorem 2.43 is the counterpart of Theorem 2.34 for V-statistics. Similar statements appear in literature (e.g. Gretton et al., 2007, Theorem 2). However, we are not aware of a complete proof of the statement. The proof requires the following intermediate result.

**Lemma 2.42 (eigenvalue representation of the bias)**
*Assume Setting 2.26 and Setting 2.32, let* $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ *be a core function satisfying the total boundedness condition of order* 2, *assume* $0 = \xi_1 < \xi_2$ *and assume* $\tilde{g}_2$ *is positive definite and continuous. Then it holds that*

$$\mathbb{E}\left(g_2(X_1, X_1)\right) = \sum_{j=1}^{\infty} \lambda_j + \theta_g$$

**Proof** Observe that $\tilde{g}_2$ is a continuous positive definite kernel. We can therefore apply Mercer's theorem A.1 to get that for all $x, y \in \operatorname{supp}(\mathbb{P}^X)$ it holds that

$$\tilde{g}_2(x, y) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x) \varphi_j(y)$$

converges uniformly. If we now take expectation and use that we can exchange the sum and expectation due the uniform convergence we get

$$\begin{aligned}
\mathbb{E}\left(\tilde{g}_2(X_1, X_1)\right) &= \mathbb{E}\left(\sum_{j=1}^{\infty} \lambda_j \varphi_j(X_1) \varphi_j(X_1)\right) \\
&= \sum_{j=1}^{\infty} \lambda_j \mathbb{E}\left(|\varphi_j(X_1)|^2\right) \\
&= \sum_{j=1}^{\infty} \lambda_j,
\end{aligned}$$

where in the last step we used that $(\varphi_j)_{j \in \mathbb{N}}$ forms an orthonormal basis of $L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})$. The result follows by noting that $g_2 \equiv \tilde{g}_2 + \theta_g$, which completes the proof of Lemma 2.42. □

We are now ready to state and prove the final result of this section.

**Theorem 2.43 (Asymptotic distribution of a V-statistic (degenerate))**
*Assume Setting 2.26 and Setting 2.32, let* $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ *be a core function satisfying the total boundedness condition of order* 2, *assume* $0 = \xi_1 < \xi_2$ *and assume* $\tilde{g}_2$ *is positive definite and continuous. Then it holds that*

$$m\left(V_m(g) - \theta_g\right) \xrightarrow{d} \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j Z_j^2$$

*as* $m \to \infty$.

**Proof** The idea of the proof is to use Lemma 2.37 to get the decomposition

$$\begin{aligned}
mV_m(\tilde{g}) = &\left(1 + \mathcal{O}\left(m^{-1}\right)\right) S_m \\
&- \left(\binom{q}{2} + \mathcal{O}\left(m^{-1}\right)\right) U_m(\tilde{g}) \\
&+ mU_m(\tilde{g})
\end{aligned} \tag{2.46}$$

as $m \to \infty$, where $S_m = U_m(w) + \frac{1}{(m)_{q-1}} \sum_B \tilde{g}(X_{i_1}, \ldots, X_{i_q})$ and $w$ is defined as in Lemma 2.37. We then calculate the asymptotic behavior of $S_m$ and use Theorem 2.34 to conclude.

Begin by analyzing the asymptotic behavior of $S_m$. To this end, note that by symmetry of the core function $g$ it holds that

$$\theta_w = \mathbb{E}(w(X_1, \ldots, X_{q-1})) = \binom{q}{2} \mathbb{E}(\tilde{g}_2(X_1, X_1)).$$

and together with Lemma 2.42 it holds that

$$\theta_w = \mathbb{E}(w(X_1, \ldots, X_{q-1})) = \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j.$$

Combining this with Theorem 2.31 it follows that

$$U_m(w) \xrightarrow{\mathbb{P}} \theta_w = \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j \tag{2.47}$$

as $m \to \infty$. Next, we use the total boundedness condition of $g$ to get a constant $C > 0$ such that

$$\begin{aligned} \mathbb{E}\Big|\frac{1}{(m)_{q-1}} \sum_B \tilde{g}(X_{i_1}, \ldots, X_{i_q})\Big| &\leq \frac{1}{(m)_{q-1}} \sum_B \mathbb{E}\big|\tilde{g}(X_{i_1}, \ldots, X_{i_q})\big| \\ &\leq C \frac{|B|}{(m)_{q-1}} \\ &= \mathcal{O}\left(m^{-1}\right) \end{aligned}$$

as $m \to \infty$. Using that $L^1$ convergence implies convergence in probability we get that

$$\frac{1}{(m)_{q-1}} \sum_{A_2} \tilde{g}(X_{i_1}, \ldots, X_{i_q}) \xrightarrow{\mathbb{P}} 0 \tag{2.48}$$

as $m \to \infty$. Finally, combining (2.47) and (2.48) this results in

$$S_m \xrightarrow{\mathbb{P}} \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j \tag{2.49}$$

as $m \to \infty$.

Now, by the properties of convergence in probability, (2.49) and Theorem 2.31 we have

$$\left(1 + \mathcal{O}\left(m^{-1}\right)\right) S_m \xrightarrow{\mathbb{P}} \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j \tag{2.50}$$

and

$$\left(\binom{q}{2} + \mathcal{O}\left(m^{-1}\right)\right) U_m(\tilde{g}) \xrightarrow{\mathbb{P}} 0 \tag{2.51}$$

as $m \to \infty$. Hence, (2.46), (2.50) and (2.51) together with Slutsky's theorem and Theorem 2.34 shows that

$$m\left(V_m(g) - \theta_g\right) = mV_m(\tilde{g}) \xrightarrow{d} \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j Z_j^2$$

as $m \to \infty$, which completes the proof of Theorem 2.43. □

### 2.3.10 Resampling results for U-statistics and V-statistics

In this section we want to consider what happens to the asymptotic behavior of $mU_m(g)$ and $mV_m(g)$ if instead of the original data sequence $(X_i)_{i\in\mathbb{N}}$ we consider a sequence of resampled data. The differences are quite subtle, therefore one needs to be very precise about what resampling means. Throughout this section we use the following setting.

**Setting 2.44 (resampling)**
*Let $\mathcal{X}$ be a separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X : \Omega \to \mathcal{X}$ be a random variable and let $(X_i)_{i\in\mathbb{N}}$ be a sequence of iid copies of $X$. For all $n \in \mathbb{N}$, let $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ be probability spaces, let $X_n^* : \Omega_n \to \mathcal{X}$ be random variables satisfying that $X_n^* \xrightarrow{d} X$ as $n \to \infty$ (i.e. $\lim_{n\to\infty} \mathbb{E}_n(f(X_n^*)) = \mathbb{E}(f(X))$ for all bounded and continuous functions $f : \mathcal{X} \to \mathbb{R}$) and let $(X_{n,i}^*)_{i\in\{1,\dots,n\}}$ be iid copies of $X_n^*$.*

The data $X_{m,1}^*, \dots, X_{m,m}^*$ should be interpreted as a new sample drawn from a distribution which converges to $\mathbb{P}^X$ as $m$ goes to infinity. Resampled data of this type often show up in different types of bootstrapping or permutation techniques.

We are interested in finding properties of the resampled U-statistc

$$\tilde{U}_m(g) := \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} g\left(X_{m,i_1}^*, \dots, X_{m,i_q}^*\right)$$

and the resampled V-statistic

$$\tilde{V}_m(g) := \frac{1}{m^q} \sum_{\mathbf{M}_q(m)} g\left(X_{m,i_1}^*, \dots, X_{m,i_q}^*\right).$$

The difference compared to the normal U-and V-statistic is that the distribution of the sample $X_{m,1}^*, \dots, X_{m,m}^*$ depends on $m$. Therefore, the results of the previous sections only carry over to the resampled U-and V-statistics if they are results for which $m$ is kept fixed. Results about the asymptotic behavior of the resampled U-and V-statistics need to be proved separately. A further more technical difficulty is that for different $m$ the random variables $\tilde{U}_m(g)$ and $\tilde{V}_m(g)$ are no longer defined on the same probability space. The following theorem gives us a way of dealing with this issue and is a slightly modified version of Skorohod's theorem (see Billingsley, 2008, Theorem 6.7).

**Theorem 2.45 (Skorohod's theorem)**
*Assume Setting 2.44. Then there exists a common probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and random variables $(\tilde{X}^*_{m,i})_{i\in\{1,\dots,m\}}$, $m \in \mathbb{N}$ and $(\tilde{X}_i)_{i\in\mathbb{N}}$ on this probability space satisfying*

*(i) for all $m \in \mathbb{N}$, for all $i \in \{1, \dots, m\}$: $\tilde{X}^*_{m,i} \sim \mathbb{P}^{X^*_m}$,*

*(ii) for all $i \in \mathbb{N}$: $\tilde{X}_i \sim \mathbb{P}^X$ and,*

*(iii) $\tilde{X}^*_{m,i} \xrightarrow{\tilde{\mathbb{P}}\text{-a.s.}} \tilde{X}_i$ as $m \to \infty$.*

In order to avoid ambiguity between the resampled and the original sample we introduce the following notation

(i) for all $m \in \mathbb{N}$ and all $c \in \{1, \dots, m\}$ define

$$g^m_c(x_1, \dots, x_c) := \mathbb{E}(g(x_1, \dots, x_c, X^*_{m,c+1}, \dots, X^*_{m,q})),$$

(ii) for all $m \in \mathbb{N}$ define

$$\theta^m_g := \mathbb{E}(g(X^*_{m,1}, \dots, X^*_{m,q})),$$

(iii) for all $m \in \mathbb{N}$ and all $c \in \{1, \dots, m\}$ define

$$\xi^m_c(g) := \mathbb{E}((g^m_c(X^*_{m,1}, \dots, X^*_{m,c}) - \theta^m_g)^2).$$

The following theorem shows that $\tilde{U}_m(g)$ is also consistent with $\theta_g$ in the appropriate sense.

**Lemma 2.46 (consistency of a resampling U-statistic)**
*Assume Setting 2.44 and let $g \in \mathcal{L}^1((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a continuous, bounded core function. Then it holds that*

$$\tilde{U}_m(g) \xrightarrow{d} \theta_g$$

*as $m \to \infty$.*

**Proof** Applying Theorem 2.45 results in a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and random variables $(\tilde{X}^*_{m,i})_{i\in\{1,\dots,m\}}$, $m \in \mathbb{N}$ and $(\tilde{X}_i)_{i\in\mathbb{N}}$ with properties specified in Theorem 2.45. Next, introduce the resampled U-statistic

$$\tilde{\mathcal{U}}_m(g) := \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}),$$

which has the same distribution under $\tilde{\mathbb{P}}$ as $\tilde{U}_m(g)$ under $\mathbb{P}_m$ and the U-statistic

$$\mathcal{U}_m(g) := \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q}),$$

which has the same distribution under $\tilde{\mathbb{P}}$ as $U_m(g)$ under $\mathbb{P}$. It holds that

$$\begin{aligned}\tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g) &= \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} \Big( g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) - g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q}) \Big) \\ &= \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} w((\tilde{X}_{i_1}, \tilde{X}^*_{m,i_1}), \dots, (\tilde{X}_{i_q}, \tilde{X}^*_{m,i_q})), \end{aligned} \tag{2.52}$$

where $w(\mathbf{x}_1, \dots, \mathbf{x}_q) := g(x_1^2, \dots, x_q^2) - g(x_1^1, \dots, x_q^1)$ is a symmetric core function. If we define for all $c \in \{1, \dots, q\}$ the functions

$$w_c^m(\mathbf{x}_1, \dots, \mathbf{x}_c) := \mathbb{E} \Big( g(x_1^2, \dots, x_c^2, \tilde{X}^*_{m,c+1}, \dots, \tilde{X}^*_{m,q}) - g(x_1^1, \dots, x_c^1, \tilde{X}_{c+1}, \dots, \tilde{X}_q) \Big)$$

it holds by the boundedness of $g$ that there exists a constant $C \in \mathbb{R}$ such that

$$\sup_{m \in \mathbb{N}} \xi_c^m(w) < C \tag{2.53}$$

By (2.52), it holds that for fixed $m$ we can apply Theorem 2.28 and together with (2.53) we get

$$\operatorname{Var}\Big( \tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g) \Big) = \binom{m}{q}^{-1} \sum_{c=1}^{m} \binom{q}{c} \binom{m-q}{q-c} \xi_c^m(w) = \mathcal{O}\left( m^{-1} \right). \tag{2.54}$$

For $(i_1, \dots, i_q) \in \mathbf{C}_q(m)$ it holds by continuity of $g$ that

$$g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) \xrightarrow{\tilde{\mathbb{P}}\text{-a.s.}} g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q})$$

as $m \to \infty$ and since $g$ is also bounded the dominated convergence theorem in particular implies

$$\lim_{m \to \infty} \mathbb{E} \Big( g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) - g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q}) \Big) = 0. \tag{2.55}$$

Combining (2.54) and (2.55) hence proves that

$$\begin{aligned} \lim_{m \to \infty} \mathbb{E} \left( \Big( \tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g) \Big)^2 \right) &= \lim_{m \to \infty} \mathbb{E} \Big( \tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g) \Big)^2 \\ &= \lim_{m \to \infty} \mathbb{E} \Big( g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) - g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q}) \Big)^2 \\ &= 0. \end{aligned}$$

Using that convergence in second moment implies convergence in probability we have therefore shown that

$$\tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g) \xrightarrow{\mathbb{P}} 0$$

as $m \to \infty$. Together with Theorem 2.31 it follows that

$$\tilde{\mathcal{U}}_m(g) - \theta_g = (\tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g)) - (\theta_g - \mathcal{U}_m(g)) \xrightarrow{\mathbb{P}} 0$$

as $m \to \infty$. This concludes the proof of Lemma 2.46. □

The following two theorems are extensions of results due to Leucht and Neumann (2009) that show that U-and V-statistics based on resampled data keep their respective asymptotic distributions. In Leucht and Neumann (2009) only U-and V-statistics of order 2 (i.e. $q = 2$) are considered. We adopted the proofs to work for arbitrary order.

**Theorem 2.47 (asymptotic distribution of degenerate a resampling U-statistic)** *Assume Setting 2.44 and Setting 2.32, let* $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ *be a continuous, bounded core function. Moreover, assume*

*(i) for all* $m \in \mathbb{N}$ *that* $g_1^m \equiv 0$*,*

*(ii)* $g_1 \equiv 0$ *(which implies* $\xi_1(g) = 0$*),*

*(iii)* $\xi_2(g) > 0$ *and*

*(iv)* $\theta_g = 0$.

*Then it holds that*

$$m\tilde{U}_m(g) \xrightarrow{d} \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$$

*as* $m \to \infty$.

**Proof** Applying Theorem 2.45 results in a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and random variables $(\tilde{X}^*_{m,i})_{i\in\{1,\dots,m\}}$, $m \in \mathbb{N}$ and $(\tilde{X}_i)_{i\in\mathbb{N}}$ with properties specified in Theorem 2.45. For $(i_1, \dots, i_q) \in \mathbf{C}_q(m)$ it holds by continuity of $g$ that

$$g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) \overset{\tilde{\mathbb{P}}\text{-a.s.}}{\longrightarrow} g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q})$$

as $m \to \infty$ and since $g$ is also bounded the dominated convergence theorem in particular implies

$$\lim_{m\to\infty} \mathbb{E}\left(\left(g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) - g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q})\right)^2\right) = 0. \tag{2.56}$$

Next, introduce the resampling U-statistic

$$\tilde{\mathcal{U}}_m(g) := \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}),$$

which has the same distribution under $\tilde{\mathbb{P}}$ as $\tilde{U}_m(g)$ under $\mathbb{P}_m$ and the U-statistic

$$\mathcal{U}_m(g) := \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q}),$$

which has the same distribution under $\tilde{\mathbb{P}}$ as $U_m(g)$ under $\mathbb{P}$. It holds that

$$\begin{aligned} \tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g) &= \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} \left(g(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) - g(\tilde{X}_{i_1}, \dots, \tilde{X}_{i_q})\right) \\ &= \binom{m}{q}^{-1} \sum_{\mathbf{C}_q(m)} w((\tilde{X}_{i_1}, \tilde{X}^*_{m,i_1}), \dots, (\tilde{X}_{i_q}, \tilde{X}^*_{m,i_q})), \end{aligned} \tag{2.57}$$

where $w(\mathbf{x}_1, \ldots, \mathbf{x}_q) := g(x_1^2, \ldots, x_q^2) - g(x_1^1, \ldots, x_q^1)$ is a symmetric core function. Define for all $c \in \{1, \ldots, q\}$ the functions

$$w_c^m(\mathbf{x}_1, \ldots, \mathbf{x}_c) := \mathbb{E}\left(g(x_1^2, \ldots, x_c^2, \tilde{X}_{m,c+1}^*, \ldots, \tilde{X}_{m,q}^*) - g(x_1^1, \ldots, x_c^1, \tilde{X}_{c+1}, \ldots, \tilde{X}_q)\right)$$

and the functions

$$\xi_c^m(w) := \mathbb{E}\left(w_c^m((\tilde{X}_1, \tilde{X}_{m,1}^*), \ldots, (\tilde{X}_c, \tilde{X}_{m,c}^*))^2\right).$$

Then, it holds by the boundedness of $g$ that there exists a constant $C \in \mathbb{R}$ such that

$$\sup_{m \in \mathbb{N}} \xi_c^m(w) < C \tag{2.58}$$

Moreover, it holds by assumption (i) and (ii) that

$$\begin{aligned} w_1^m(\mathbf{x}_1) &= \mathbb{E}\left(g(x_1^2, \tilde{X}_{m,2}^*, \ldots, \tilde{X}_{m,q}^*) - g(x_1^1, \tilde{X}_2, \ldots, \tilde{X}_q)\right) \\ &= g_1^m(x_1^2) - g_1(x_1^1) \\ &= 0, \end{aligned}$$

which immediately implies that

$$\xi_1^m(w) = \mathbb{E}\left(w_1^m((\tilde{X}_1, \tilde{X}_{m,1}^*))^2\right) = 0. \tag{2.59}$$

Furthermore, using Jensen's inequality it holds that

$$\begin{aligned} \xi_2^m(w) &= \mathbb{E}\left(w_2^m((\tilde{X}_1, \tilde{X}_{m,1}^*), (\tilde{X}_2, \tilde{X}_{m,2}^*))^2\right) \\ &\leq \mathbb{E}\left(w((\tilde{X}_1, \tilde{X}_{m,1}^*), \ldots, (\tilde{X}_q, \tilde{X}_{m,q}^*))^2\right) \\ &= \mathbb{E}\left(\left(g(\tilde{X}_{m,1}^*, \ldots, \tilde{X}_{m,q}^*) - g(\tilde{X}_1, \ldots, \tilde{X}_q)\right)^2\right). \end{aligned} \tag{2.60}$$

By (2.57), it holds that for fixed $m$ we can apply Theorem 2.28 and together with (2.58) and (2.59) we get

$$\operatorname{Var}\left(\tilde{\mathcal{U}}_m(g) - \mathcal{U}_m(g)\right) = \binom{m}{q}^{-1} \sum_{c=1}^{m} \binom{q}{c}\binom{m-q}{q-c} \xi_c^m(w) = \mathcal{O}\left(m^{-2}\right)\xi_2^m(w) + \mathcal{O}\left(m^{-3}\right).$$

Hence, together with (2.60) and (2.56) it holds that

$$\lim_{m \to \infty} \operatorname{Var}\left(m\left(\mathcal{U}_m(g) - \tilde{\mathcal{U}}_m(g)\right)\right) = 0$$

and consequently also that

$$m\left(\mathcal{U}_m(g) - \tilde{\mathcal{U}}_m(g)\right) \xrightarrow{\tilde{P}} 0$$

as $m \to \infty$. Finally, applying Slutsky's theorem together with Theorem 2.34 results in

$$m\tilde{U}_m(g) = mU_m(g) + m\left(\tilde{U}_m(g) - U_m(g)\right) \xrightarrow{d} \binom{q}{2}\sum_{j=1}^{\infty}\lambda_j(Z_j^2 - 1)$$

as $m \to \infty$, which completes the proof of Theorem 2.47. □

The same result also holds for V-statistics. The proof uses the same technique as the proof of Theorem 2.43 and reduces the V-statistic back to the U-statistic.

**Theorem 2.48 (asymptotic distribution of degenerate resampling V-statistic)** *Assume Setting 2.44 and Setting 2.32, let $g \in \mathcal{L}^2((\mathbb{P}^X)^{\otimes q}, |\cdot|_{\mathbb{R}})$ be a continuous, bounded core function. Moreover, assume*

*(i) for all $m \in \mathbb{N}$ that $g_1^m \equiv 0$,*

*(ii) $g_1 \equiv 0$ (which implies $\xi_1(g) = 0$),*

*(iii) $\xi_2(g) > 0$ and*

*(iv) $\theta_g = 0$.*

*Then it holds that*

$$m\tilde{V}_m(g) \xrightarrow{d} \binom{q}{2}\sum_{j=1}^{\infty}\lambda_j Z_j^2$$

*as $m \to \infty$.*

**Proof** Applying Theorem 2.45 results in a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ and random variables $(\tilde{X}_{m,i}^*)_{i\in\{1,\dots,m\}}$, $m \in \mathbb{N}$ and $(\tilde{X}_i)_{i\in\mathbb{N}}$ with properties specified in Theorem 2.45. Next, introduce the resampling U-statistic

$$\tilde{\mathcal{U}}_m(g) := \binom{m}{q}^{-1}\sum_{\mathbf{C}_q(m)} g(\tilde{X}_{m,i_1}^*, \dots, \tilde{X}_{m,i_q}^*),$$

which has the same distribution under $\tilde{\mathbb{P}}$ as $\tilde{U}_m(g)$ under $\mathbb{P}_m$ and the resampling V-statistic

$$\tilde{\mathcal{V}}_m(g) := \frac{1}{m^q}\sum_{\mathbf{M}_q(m)} g(\tilde{X}_{m,i_1}^*, \dots, \tilde{X}_{m,i_q}^*),$$

which has the same distribution under $\tilde{\mathbb{P}}$ as $\tilde{V}_m(g)$ under $\mathbb{P}_m$. For fixed $m \in \mathbb{N}$ we can view $\tilde{\mathcal{V}}_m(g)$ as a V-statistic and apply an adjusted version of Lemma 2.37 to get

$$m\tilde{\mathcal{V}}_m(\tilde{g}) = \left(1 + \mathcal{O}\left(m^{-1}\right)\right)S_m + \left(m - \binom{q}{2} + \mathcal{O}\left(m^{-1}\right)\right)\tilde{\mathcal{U}}_m(\tilde{g}) \tag{2.61}$$

as $m \to \infty$, where $S_m = \tilde{\mathcal{U}}_m(w) + \frac{1}{(m)_{q-1}}\sum_B \tilde{g}(\tilde{X}_{m,i_1}^*, \dots, \tilde{X}_{m,i_q}^*)$. By the symmetry of the core function $g$ and the definition of $w$ given in Lemma 2.37 it holds that

$$\theta_w = \mathbb{E}\left(w(X_1, \dots, X_{q-1})\right) = \binom{q}{2}\mathbb{E}\left(\tilde{g}_2(X_1, X_1)\right).$$

The consistency of resampled U-statistics given in Lemma 2.46 together with Lemma 2.42 imply that

$$\tilde{\mathcal{U}}_m(w) \xrightarrow{d} \theta_w = \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j \tag{2.62}$$

as $m \to \infty$. The boundedness of $g$ combined with the size of the set $B$ given in Lemma 2.37 shows that

$$\frac{1}{(m)_{q-1}} \sum_{B} \tilde{g}(\tilde{X}^*_{m,i_1}, \dots, \tilde{X}^*_{m,i_q}) \leq \frac{C|B|}{(m)_{q-1}} = \mathcal{O}\left(m^{-1}\right). \tag{2.63}$$

Moreover, also by Lemma 2.46 it holds that

$$\tilde{\mathcal{U}}_m(\tilde{g}) \xrightarrow{d} 0 \tag{2.64}$$

as $m \to \infty$ and by Theorem 2.47 it holds that

$$m\tilde{\mathcal{U}}_m(\tilde{g}) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j \left(Z_j^2 - 1\right). \tag{2.65}$$

Finally, we can combine (2.61), (2.62), (2.63), (2.64), (2.65) and use that convergence in distribution to a constant implies convergence in probability together with Slutsky's theorem to get that

$$m\tilde{\mathcal{V}}_m(\tilde{g}) \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j Z_j^2$$

as $m \to \infty$. This concludes the proof of Theorem 2.48. □

## 2.4 Mathematical statistics

### 2.4.1 Statistical Framework

This section intends to shortly introduce a basic statistical framework. For our purposes it will be important to have a setting that allows for asymptotic considerations. Similar settings can be found in almost every book on mathematical statistics (e.g. Lehmann and Casella, 1998).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{X}$ be a measurable space and let $X : \Omega \to \mathcal{X}$ be a random variable for which the exact law $\mathbb{P}^X$ is not fully known. In statistics one is often interested in determining different characteristics of the distribution $\mathbb{P}^X$, which are generally referred to as statistical functionals.

**Definition 2.49 (statistical functional)**
*Let $\mathcal{X}$ be a measurable space. Then a function*

$$\theta : \mathcal{P}(\mathcal{X}) \longrightarrow \mathbb{R}$$

*is a called statistical functional.*

A fundamental question in statistics is how to estimate the value of $\theta(\mathbb{P}^X)$ given a set of observations connected in some way to $X$. This estimation process is referred to as point estimation in literature. A common setting is to assume the existence of a sequence $(X_i)_{i\in\mathbb{N}}$ of iid copies of $X$, for which we observe one particular outcome $(x_i)_{i\in\mathbb{N}}$. The intuition being that the $X_i$ are a sequence of independent experiments with observed outcome $x_i$. One is then interested in a method that, based on the first $m$ observations $x_1, \dots, x_m$, outputs a 'reasonable' estimate of $\theta(\mathbb{P}^X)$. The following definition formalizes the estimation method.

**Definition 2.50 (statistic/estimator)**
*Let $\mathcal{X}$ be a measurable space. Then a family of measurable functions $T = (T_m)_{m\in\mathbb{N}}$ satisfying for all $m \in \mathbb{N}$ that*

$$T_m : \mathcal{X}^m \to \mathbb{R}$$

*is called a statistic (or estimator).*

Given a statistical functional $\theta$ we call a statistic $T = (T_m)_{m\in\mathbb{N}}$ an estimator of $\theta$ if $T_m(X_1, \dots, X_m)$ is in some sense a good approximation of $\theta$. A good approximation could for example mean that $T$ satisfies for all $X$ with $\mathbb{P}^X \in \mathcal{P}(\mathcal{X})$ and for all $m \in \mathbb{N}$ that

$$\mathbb{E}\left(T_m(X_1, \dots, X_m)\right) = \theta\left(\mathbb{P}^X\right),$$

where $X_1, X_2, \dots, \overset{\text{iid}}{\sim} \mathbb{P}^X$. Statistics satisfying this condition are called unbiased estimators of $\theta$. There are however also many other notions of 'good' approximations, many of which are concerned with asymptotic properties of $T_m$ as $m$ goes to infinity.

### 2.4.2 Hypothesis testing

In this section we intend to introduce a general framework for hypothesis testing. Similar definitions as given in this section can be found in Lehmann and Romano (2005).

Let $\Theta \subseteq \mathcal{P}(\mathcal{X})$ be the model class, i.e. the set of probability measures that fall within the statistical model. We will simply take $\Theta = \mathcal{P}(\mathcal{X})$, since we do not want to make any model restrictions.

Furthermore, let $H_0 \subseteq \Theta$ and $H_A \subseteq \Theta$ such that $H_0 \cap H_A = \varnothing$. We call $H_0$ the null hypothesis and $H_A$ the alternative hypothesis. Given a generating process $X$ with law $\mathbb{P}^X \in \Theta$ and a sequence of iid copies $(X_i)_{i\in\mathbb{N}}$ of $X$, the goal of a statistical hypothesis test is to decide, based on a finite sample $X_1, \dots, X_m$, whether to accept or reject the null hypothesis

$$\mathbb{P}^X \in H_0.$$

We formalize this in the following definition.

**Definition 2.51 (statistical hypothesis test)**
*Let $\mathcal{X}$ be a separable metric space. Let $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ be a family of measurable functions with the property that for all $m \in \mathbb{N}$ it holds that $\varphi_m : \mathcal{X}^m \to \{0, 1\}$. Then we call $\varphi$ a statistical hypothesis test.*

Given a specific outcome $x_1, \dots, x_m$ of the random variables $X_1, \dots, X_m$ we interpret

$$\varphi_m(x_1, \dots, x_m) = 1$$

as rejection of the null hypothesis (i.e. $\mathbb{P}^X \notin H_0$) and

$$\varphi_m(x_1, \dots, x_m) = 0$$

as acceptance of the null hypothesis (i.e. $\mathbb{P}^X \in H_0$). Of course such an interpretation only makes sense if $\varphi$ encodes information about whether $\mathbb{P}^X$ actually lies within the null hypothesis or not. To quantify this, we introduce two notions of error.

**Definition 2.52 (Type I and Type II error)**
*Let $\mathcal{X}$ be a separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ be a statistical hypothesis test. Then, for all $m \in \mathbb{N}$ define the Type I error (given $m$ observations) by*

$$\mathrm{E}_1(\varphi_m) \coloneqq \sup_{X\colon \mathbb{P}^X \in H_0} \mathbb{P}\left(\varphi_m(X_1, \dots, X_m) = 1\right),$$

*where $X_1, X_2, \dots \overset{iid}{\sim} \mathbb{P}^X$ and the Type II error (given $m$ observations) by*

$$\mathrm{E}_2(\varphi_m) \coloneqq \sup_{X\colon \mathbb{P}^X \in H_A} \mathbb{P}\left(\varphi_m(X_1, \dots, X_m) = 0\right),$$

*where $X_1, X_2, \dots \overset{iid}{\sim} \mathbb{P}^X$.*

The Type I error is the highest possible probability of rejecting the null hypothesis although it is true, while the Type II error is the highest probability of accepting the null hypothesis even though the alternative hypothesis holds. In traditional hypothesis testing one usually takes an asymmetric viewpoint and assumes that a Type I error is more severe than a Type II error. The first priority is thus to control the Type I error, which is done by enforcing that it lies below a certain threshold $\alpha \in (0, 1)$. This is formalized in the next definition.

**Definition 2.53 (level of a test)**
*Let $\mathcal{X}$ be a separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\alpha \in (0, 1)$ and let $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ be a statistical hypothesis test. If $\varphi$ satisfies for all $m \in \mathbb{N}$ that*

$$\mathrm{E}_1(\varphi_m) \leq \alpha,$$

*then we call $\varphi$ a hypothesis test at (valid) level $\alpha$. If $\varphi$ satisfies*

$$\limsup_{m\to\infty} \mathrm{E}_1(\varphi_m) \leq \alpha,$$

*then we call $\varphi$ a hypothesis test at uniform asymptotic level $\alpha$. If $\varphi$ satisfies for all $X$ with $\mathbb{P}^X \in H_0$ that*

$$\lim_{m\to\infty} \mathbb{P}\left(\varphi_m(X_1, \dots, X_m) = 1\right) \leq \alpha,$$

*where $X_1, X_2, \dots \overset{iid}{\sim} \mathbb{P}^X$ then we call $\varphi$ a hypothesis test at pointwise asymptotic level $\alpha$.*

A further desirable property of a test is that it is able to detect (at least in the large sample limit) if a given set of observations comes from a random variable, with law in the alternative hypothesis. For example, one can always consider the trivial test $phi_m \equiv 0$ which achieves any level but will never reject the null hypothesis. This is the subject of the next definition.

**Definition 2.54 (consistency of a test)**
*Let $\mathcal{X}$ be a separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ be a statistical hypothesis test. If $\varphi$ satisfies*

$$\lim_{m\to\infty} \mathrm{E}_2(\varphi_m) = 0,$$

*then we call $\varphi$ a uniformly consistent hypothesis test. If $\varphi$ satisfies for all $X$ with $\mathbb{P}^X \in H_A$ that*

$$\lim_{m\to\infty} \mathbb{P}\left(\varphi_m(X_1, \dots, X_m) = 0\right) = 0,$$

*where $X_1, X_2, \dots \overset{iid}{\sim} \mathbb{P}^X$ then we call $\varphi$ a pointwise consistent hypothesis test.*

### Construction of tests

In the previous section we introduced the statistical hypothesis test as an abstract estimator. It is, however, not immediately clear what such a test looks like. In fact, a statistical hypothesis test can have an arbitrarily complex form. For practical applications, there exist many heuristics that can be used to help construct explicit tests.

The starting point is generally a statistic $T = (T_m)_{m\in\mathbb{N}}$ on $\mathcal{X}$ called test statistic. It should behave differently under the null hypothesis $H_0$ than under the alternative hypothesis $H_A$. For example, $T$ could be negative in expectation under $H_0$ and positive in expectation under $H_A$. A statistical hypothesis test could then try to make use of these properties in an appropriate way. One option would be to introduce a threshold statistic $c = (c_m)_{m\in\mathbb{N}}$ (potentially depending on the observations) and define a test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ such that for all $m \in \mathbb{N}$ and all $x_1, \dots, x_m \in \mathcal{X}$ it holds that

$$\varphi_m(x_1, \dots, x_m) := \begin{cases} 0 & \text{if } T_m(x_1, \dots, x_m) \leq c_m(x_1, \dots, x_m) \\ 1 & \text{if } T_m(x_1, \dots, x_m) > c_m(x_1, \dots, x_m). \end{cases} \tag{2.66}$$

Given a fixed level $\alpha \in (0, 1)$, we then try to chose the threshold $c$ in such way that $\varphi$ is a consistent test at (asymptotic) level $\alpha$.

### Resampling tests

Generally, choosing the threshold $c$ requires some type of knowledge of the distribution of $T_m$. In practical applications, it is however often hard to get this type of information. Resampling tests are one option to avoid explicitly determining the distribution of $T$.

Let $\alpha \in (0, 1)$, let $\mathcal{X}$ be a separable metric space, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $X$ be a random variable with values in $\mathcal{X}$ and let $(X_i)_{i\in\mathbb{N}}$ be a sequence of iid copies.

The main idea behind resampling tests is to construct data sets based on the original observations $(X_1, \dots, X_m)$. These types of constructions are formalized by resampling methods.

**Definition 2.55 (resampling method)**
*Let $\mathcal{X}$ be a separable metric space and let $(M_m)_{m\in\mathbb{N}} \subseteq \mathbb{N}$ be a sequence. If*

$$g = \big((g_{m,k})_{k\in\{1,\dots,M_m\}}\big)_{m\in\mathbb{N}}$$

*is a family of functions satisfying for all $m \in \mathbb{N}$ and for all $k \in \{1, \dots, M_m\}$ that*

$$g_{m,k} : \mathcal{X}^m \to \mathcal{X}^m,$$

*then we call $g$ a resampling method.*

Based on a resampling method $g$ we can construct new observations for all $m \in \mathbb{N}$ and for all $k \in \{1, \dots, M_m\}$ by defining

$$Z_{m,k} \coloneqq g_{m,k}(X_1, \dots, X_m).$$

The new 'resampled' data $(Z_{m,k})_{k\in\{1,\dots,M_m\}} \subseteq \mathcal{X}^m$, $m \in \mathbb{N}$ is called resampling scheme and for each $m \in \mathbb{N}$ the sequence $Z_{m,1}, \dots, Z_{m,M_m}$ should be seen as $M_m$ resampled data sets constructed from the original observations $(X_1, \dots, X_m)$. A resampling method is therefore a formalization of the concept of resampling $M_m$ times from the original observations $(X_1, \dots, X_m)$.

Based on a resampling method we can introduce the resampling distribution function.

**Definition 2.56 (resampling distribution function)**
*Let $\mathcal{X}$ be a separable metric space, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $g$ be a resampling method. For all $m \in \mathbb{N}$ the functions $\widehat{R}_{T_m} : \mathcal{X}^m \times \mathbb{R} \to [0, 1]$ defined for all $(x_1, \dots, x_m) \in \mathcal{X}^m$ and for all $t \in \mathbb{R}$ by*

$$\widehat{R}_{T_m}(x_1, \dots, x_m)(t) \coloneqq \frac{1}{M_m} \sum_{k=1}^{M_m} \mathbb{1}_{\{T_m(g_{m,k}(x_1,\dots,x_m))\leq t\}}$$

*are called the resampling distribution functions (corresponding to test statistic $T$ and resampling method $g$).*

Fixing $m \in \mathbb{N}$ and $(x_1, \dots, x_m) \in \mathcal{X}^m$ it holds that

$$\widehat{R}_{T_m}(x_1, \dots, x_m) : \mathbb{R} \to [0, 1]$$

is non-decreasing, right-continuous and satisfies

$$\lim_{t\to-\infty} \widehat{R}_{T_m}(x_1, \dots, x_m)(t) = 0 \quad \text{and} \quad \lim_{t\to\infty} \widehat{R}_{T_m}(x_1, \dots, x_m)(t) = 1.$$

This implies that $\widehat{R}_{T_m}(x_1, \dots, x_m)$ is a distribution function and thus we can define the generalized inverse

$$\Big(\widehat{R}_{T_m}(x_1, \dots, x_m)\Big)^{-1} : (0, 1) \to \mathbb{R}$$

satisfying for all $\alpha \in (0,1)$ that

$$\left(\widehat{R}_{T_m}(x_1,\dots,x_m)\right)^{-1}(\alpha) := \inf\{t \in \mathbb{R} \mid \widehat{R}_{T_m}(x_1,\dots,x_m)(t) \geq \alpha\}.$$

Based on the resampling distribution functions we can define a resampling test as follows.

**Definition 2.57 (resampling test)**
*Let $\alpha \in (0,1)$, let $\mathcal{X}$ be a separable metric space, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $g$ be a resampling method and let $\widehat{R}_{T_m}$ be the corresponding resampling distribution functions. A hypothesis test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ defined for all $m \in \mathbb{N}$ and for all $(x_1,\dots,x_m) \in \mathcal{X}^m$ by*

$$\varphi_m(x_1,\dots,x_m) := \mathbb{1}_{\left\{T_m(x_1,\dots,x_m) > (\widehat{R}_{T_m}(x_1,\dots,x_m))^{-1}(1-\alpha)\right\}}$$

*is called $\alpha$-resampling test (corresponding to $g$).*

The advantage of resampling tests is that they can be constructed for any test statistic. We now define an important subclass of resampling methods.

**Definition 2.58 (resampling group)**
*Let $\mathcal{X}$ be a separable metric space, let $(M_m)_{m\in\mathbb{N}} \subseteq \mathbb{N}$ be a sequence and let $g$ be a resampling method. If $g$ satisfies that*

$$G := \{g_{m,1},\dots,g_{m,M_m}\}$$

*together with concatenation is a group of transformations on $\mathcal{X}^m$, then we call $g$ a resampling group.*

Resampling groups have the important property that for all test statistics $T = (T_m)_{m\in\mathbb{N}}$ the corresponding resampling distribution functions satisfy for all $m \in \mathbb{N}$, for all $k \in \{1,\dots,M_m\}$ and for all $(x_1,\dots,x_m) \in \mathcal{X}^m$ that

$$\widehat{R}_{T_m}(x_1,\dots,x_m) = \widehat{R}_{T_m}(g_{m,k}(x_1,\dots,x_m)). \tag{2.67}$$

This follows immediately from the group property of $g$. It allows us to prove, given an appropriate invariance of the resampling group under the null hypothesis, that the corresponding resampling test achieves level $\alpha$. The following theorem is a reformulation of Lehmann and Romano (2005, Theorem 15.2.1).

**Theorem 2.59 (level of resampling tests)**
*Let $\alpha \in (0,1)$, let $\mathcal{X}$ be a separable metric space, let $H_0, H_A \subseteq \mathcal{P}(\mathcal{X})$ be a null and alternative hypothesis respectively, let $g$ be a resampling group satisfying under $H_0$ that for all $m \in \mathbb{N}$ and for all $k \in \{1,\dots,M_m\}$ it holds that*

$$g_{m,k}(X_1,\dots,X_m) \text{ is equal in distribution to } (X_1,\dots,X_m).$$

*Then, the $\alpha$-resampling test $\varphi$ corresponding to $g$ is a test at level $\alpha$, when testing $H_0$ against $H_A$.*

**Proof** Fix $m \in \mathbb{N}$ and let $K$ be a uniformly distributed random variable on $\{1, \ldots, M_m\}$ independent of $(X_1, \ldots, X_m)$. Let $(x_1, \ldots, x_m) \in \operatorname{Im}((X_1, \ldots, X_m))$ and for all $k \in \{1, \ldots, M_m\}$ define $z_{m,k} \coloneqq g_{m,k}(x_1, \ldots, x_m)$ then it holds that

$$T_m(z_{m,K}) \tag{2.68}$$

has the distribution function $\widehat{R}_{T_m}(x_1, \ldots, x_m)$. Hence, using (2.67) and the the properties of the generalized inverse it holds that

$$\begin{aligned}
&\frac{1}{M_m}\sum_{k=1}^{M_m} \mathbb{1}_{\{T_m(z_{m,k})>(\widehat{R}_{T_m}(z_{m,k}))^{-1}(1-\alpha)\}}\\
&= \frac{1}{M_m}\sum_{k=1}^{M_m} \mathbb{1}_{\{T_m(z_{m,k})>(\widehat{R}_{T_m}(x_1,\ldots,x_m))^{-1}(1-\alpha)\}}\\
&= \mathbb{E}\left(\mathbb{1}_{\{T_m(z_{m,K})>(\widehat{R}_{T_m}(x_1,\ldots,x_m))^{-1}(1-\alpha)\}}\right)\\
&\leq \alpha,
\end{aligned}$$

which together with the monotonicity of the integral and the convention

$$Z_{m,k} = g_{m,k}(X_1, \ldots, X_m)$$

implies that

$$\mathbb{E}\left(\frac{1}{M_m}\sum_{k=1}^{M_m} \mathbb{1}_{\{T_m(Z_{m,k})>(\widehat{R}_{T_m}(Z_{m,k}))^{-1}(1-\alpha)\}}\right) \leq \alpha. \tag{2.69}$$

Moreover, under $H_0$, i.e. $X_1, X_2, \ldots \sim \mathbb{P}^X \in H_0$, it holds by assumption for all $k \in \{1, \ldots, M_m\}$ that $(X_1, \ldots, X_m)$ is equal in distribution to $Z_{m,k}$. This in particular implies that under $H_0$ it holds for all $k \in \{1, \ldots, M_m\}$ that

$$\mathbb{E}\left(\varphi_m(Z_{m,k})\right) = \mathbb{E}\left(\varphi_m(X_1, \ldots, X_m)\right). \tag{2.70}$$

Combining (2.69) and (2.70) results in

$$\begin{aligned}
\mathbb{P}\left(\varphi_m(X_1, \ldots, X_m) = 1\right) &= \mathbb{E}\left(\varphi_m(X_1, \ldots, X_m)\right)\\
&= \frac{1}{M_m}\sum_{k=1}^{M_m} \mathbb{E}\left(\varphi_m(Z_{m,k})\right)\\
&= \mathbb{E}\left(\frac{1}{M_m}\sum_{k=1}^{M_m} \mathbb{1}_{\{T_m(Z_{m,k})>(\widehat{R}_{T_m}(Z_{m,k}))^{-1}(1-\alpha)\}}\right)\\
&\leq \alpha,
\end{aligned}$$

which completes the proof of Theorem 2.59. □

The invariance assumption of the resampling group in the previous theorem is the same as the randomization hypothesis given by Lehmann and Romano (2005, Definition 15.2.1).

Unfortunately, there are no guarantees that an arbitrary resampling test controls the Type II error in any way. Results of this type need to be checked on a case by case basis by analyzing the resampling distribution function for the specific test statistic.

### Monte-Carlo approximated resampling tests

Finally, we want to discuss a computational difficulty that often arises in the context of resampling tests. The problem is that in practical applications the parameter $M_m$ from the definition of a resampling method grows very fast in $m$ and makes computations impossible for large $m$. One method of dealing with this is to approximate the resampling distribution $\widehat{R}_m$ using a Monte-Carlo approximated version.

**Definition 2.60 (Monte-Carlo approximated resampling distribution)** *Let $\mathcal{X}$ be a separable metric space, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $g$ be a resampling method and let $(K_i)_{i\in\mathbb{N}}$ be a sequence of independent uniformly distributed random variables on $\{1, \ldots, M_m\}$. For all $B \in \mathbb{N}$ let $\widehat{\mathcal{R}}^B_{T_m} : \mathcal{X}^m \times \mathbb{R} \to [0, 1]$ be the functions defined for all $(x_1, \ldots, x_m) \in \mathcal{X}^m$ and for all $t \in \mathbb{R}$ by*

$$\widehat{\mathcal{R}}^B_{T_m}(x_1, \ldots, x_m)(t) := \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}_{\{T_m(g_{m,K_i}(x_1,\ldots,x_m))\leq t\}}$$

*are called the Monte-Carlo approximated resampling distribution functions (corresponding to test statistic $T$ and resampling method $g$).*

The following proposition shows that $\widehat{\mathcal{R}}^B_{T_m}$ approximates $\widehat{R}_{T_m}$ in an appropriate way.

**Proposition 2.61 (Monte-Carlo approximation of resampling distribution)** *Let $\mathcal{X}$ be a separable metric space, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $g$ be a resampling method, let $\widehat{R}_{T_m}$ be the resampling distribution functions and for all $B \in \mathbb{N}$ let $\widehat{\mathcal{R}}^B_{T_m}$ be the Monte-Carlo approximated resampling distribution functions. Then for all $(x_1, \ldots, x_m) \in \mathcal{X}^m$ and for all $t \in \mathbb{R}$ it holds $\mathbb{P}$-a.s. that*

$$\lim_{B\to\infty} \widehat{\mathcal{R}}^B_{T_m}(x_1, \ldots, x_m)(t) = \widehat{R}_{T_m}(x_1, \ldots, x_m)(t).$$

**Proof** Let $(K_i)_{i\in\mathbb{N}}$ be the sequence of uniformly distributed random variables on $\{1, \ldots, M_m\}$ from the definition of $\widehat{\mathcal{R}}^B_{T_m}$, then introduce for all $k \in \{1, \ldots, M_m\}$ and for all $i \in \mathbb{N}$ the random variables

$$Y^k_i := \mathbb{1}_{\{K_i=k\}}.$$

$Y^k_i$ has a Bernoulli distribution with parameter $\frac{1}{M_m}$. Furthermore, we can write

$$\begin{aligned}\widehat{\mathcal{R}}^B_{T_m}(x_1, \ldots, x_m)(t) &= \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}_{\left\{T_m(g_{m,K_i}(x_1,\ldots,x_m))\leq t\right\}} \\ &= \sum_{k=1}^{M_m} \frac{\sum_{i=1}^{B} Y^k_i}{B} \mathbb{1}_{\left\{T_m(g_{m,k}(x_1,\ldots,x_m))\leq t\right\}}.\end{aligned}$$

By the strong law of large numbers this implies that $\mathbb{P}$-a.s. it holds that

$$\lim_{B\to\infty} \widehat{\mathcal{R}}^B_{T_m}(x_1, \ldots, x_m)(t) = \frac{1}{M_m}\sum_{k=1}^{M_m} \mathbb{1}_{\left\{T_m(g_{m,k}(x_1,\ldots,x_m))\leq t\right\}} = \widehat{R}_{T_m}(x_1, \ldots, x_m)(t),$$

which completes the proof of Proposition 2.61. □

We are now ready to define Monte-Carlo approximated resampling test. Instead of using the $(1-\alpha)$-quantile of the Monte-Carlo approximated resampling distribution we use a slightly larger critical value. Surprisingly, for resampling groups satisfying the invariance condition in Theorem 2.59, this allows us to achieve level $\alpha$ for any value of $B$. The trick is that slightly larger critical value accounts for the uncertainty due to the Monte-Carlo approximation.

We define the test using the p-value as this leads to easier calculations. The corresponding critical value can then be calculated via the standard correspondence between p-value and hypothesis test.

**Definition 2.62 (Monte-Carlo approximated resampling test)**
*Let $\alpha \in (0,1)$, let $\mathcal{X}$ be a separable metric space, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $g$ be a resampling method, let $B \in \mathbb{N}$, let $(K_i)_{i\in\mathbb{N}}$ be a sequence of independent uniformly distributed random variables on $\{1,\dots,M_m\}$ and let $(k_1,\dots,k_B)$ be a realization of $(K_1,\dots,K_B)$. For all $m \in \mathbb{N}$ define the function $\widehat{p}_m : \mathcal{X}^m \to [\frac{1}{B+1}, 1]$ satisfying*

$$\widehat{p}_m(x_1,\dots,x_m) \coloneqq \frac{1 + \left|\{i \in \{1,\dots,B\} : T_m(g_{m,k_i}(x_1,\dots,x_m)) \geq T_m(x_1,\dots,x_m)\}\right|}{1+B}.$$

*Then the hypothesis test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ defined for all $m \in \mathbb{N}$ and for all $(x_1,\dots,x_m) \in \mathcal{X}^m$ by*

$$\varphi_m(x_1,\dots,x_m) \coloneqq \mathbb{1}_{\{\widehat{p}_m(x_1,\dots,x_m)\leq\alpha\}},$$

*is called $\alpha$-Monte-Carlo approximated resampling test.*

The function $\widehat{p}_m$ is called p-value of the test $\varphi_m$. The following proposition shows that the Monte-Carlo approximated resampling test achieves level $\alpha$ given the appropriate invariance assumptions on $g$.

**Proposition 2.63 (Monte-Carlo approximated resampling test has valid level)**
*Let $\alpha \in (0,1)$, let $\mathcal{X}$ be a separable metric space, let $H_0, H_A \subseteq \mathcal{P}(\mathcal{X})$ be a null and alternative hypothesis respectively, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $B \in \mathbb{N}$ and let $g$ be a resampling group satisfying under $H_0$ that for all $m \in \mathbb{N}$ and for all $k \in \{1,\dots,M_m\}$ it holds that*

$$g_{m,k}(X_1,\dots,X_m) \text{ is equal in distribution to } (X_1,\dots,X_m), \tag{2.71}$$

*and for all $k, l \in \{1,\dots,M_m\}$ it holds that*

$$\mathbb{P}\left(T_m(g_{m,k}(X_1,\dots,X_m)) = T_m(g_{m,l}(X_1,\dots,X_m))\right) = 0. \tag{2.72}$$

*Then, the corresponding $\alpha$-Monte-Carlo approximated resampling test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ has level $\alpha$ when testing $H_0$ against $H_A$.*

**Proof** Begin by defining the function $f : \{1,\dots,M_m\}^B \times \mathcal{X}^m \to \{0,\dots,B\}$ satisfying for all $(k_1,\dots,k_B) \in \{1,\dots,M_m\}^B$ and for all $(x_1,\dots,x_m) \in \mathcal{X}^m$ that

$$f(k_1,\dots,k_B)(x_1,\dots,x_m) \coloneqq \left|\{i \in \{1,\dots,B\} : T_m(g_{m,k_i}(x_1,\dots,x_m)) \geq T_m(x_1,\dots,x_m)\}\right|,$$

and the function $f_{tot} : \mathcal{X}^m \to \{1, \dots, M_m\}$ satisfying for all $(x_1, \dots, x_m) \in \mathcal{X}^m$ that

$$f_{tot}(x_1, \dots, x_m) \coloneqq \big|\{i \in \{1, \dots, M_m\} : T_m(g_{m,i}(x_1, \dots, x_m)) \geq T_m(x_1, \dots, x_m)\}\big|.$$

Then, by the invariance assumption (2.71) it holds under $H_0$ for all $k, l \in \{1, \dots, M_m\}$ that

$$\mathbb{P}\left(f_{tot}(X_1, \dots, X_m) = l\right) = \mathbb{P}\left(f_{tot}(g_{m,k}(X_1, \dots, X_m)) = l\right). \tag{2.73}$$

Moreover, since $g$ is a group it holds $\mathbb{P}$-a.s. that

$$\begin{aligned} f_{tot}(g_{m,k}(X_1, \dots, X_m)) &= \sum_{i=1}^{M_m} \mathbb{1}_{\{T_m(g_{m,i}(g_{m,k}(X_1,\dots,X_m))) \geq T_m(g_{m,k}(X_1,\dots,X_m))\}} \\ &= \sum_{i=1}^{M_m} \mathbb{1}_{\{T_m(g_{m,i}(X_1,\dots,X_m)) \geq T_m(g_{m,k}(X_1,\dots,X_m))\}}, \end{aligned}$$

which implies together with (2.72) it holds $\mathbb{P}$-a.s. that

$$\sum_{k=1}^{M_m} \mathbb{1}_{\{f_{tot}(g_{m,k}(X_1,\dots,X_m))=l\}} = 1. \tag{2.74}$$

Combining (2.73) and (2.74) it holds under $H_0$ that

$$\begin{aligned} \mathbb{P}\left(f_{tot}(X_1, \dots, X_m) = l\right) &= \frac{1}{M_m} \sum_{k=1}^{M_m} \mathbb{P}\left(f_{tot}(g_{m,k}(X_1, \dots, X_m)) = l\right) \\ &= \frac{1}{M_m} \sum_{k=1}^{M_m} \mathbb{E}\left(\mathbb{1}_{\{f_{tot}(g_{m,k}(X_1,\dots,X_m))=l\}}\right) \\ &= \frac{1}{M_m} \mathbb{E}\left(\sum_{k=1}^{M_m} \mathbb{1}_{\{f_{tot}(g_{m,k}(X_1,\dots,X_m))=l\}}\right) \\ &= \frac{1}{M_m}, \end{aligned}$$

which proves that under $H_0$ it holds that $f_{tot}(X_1, \dots, X_m)$ is uniformly distributed on $\{1, \dots, M_m\}$. Furthermore, conditioned on $f_{tot}(X_1, \dots, X_m) = l$ it holds for all $i \in \{1, \dots, B\}$ that

$$\mathbb{1}_{\left\{T_m(g_{m,K_i}(X_1,\dots,X_m)) \geq T_m(X_1,\dots,X_m)\right\}}$$

is Bernoulli $\frac{l}{M_m}$ distributed which again conditioned on $f_{tot}(X_1, \dots, X_m) = l$ implies that

$$f(K_1, \dots, K_B)(X_1, \dots, X_m) = \sum_{i=1}^{B} \mathbb{1}_{\left\{T_m(g_{m,K_i}(X_1,\dots,X_m)) \geq T_m(X_1,\dots,X_m)\right\}}$$

has binomial distribution with parameters $B$ and $\frac{l}{M_m}$. It therefore holds under $H_0$ that

$$\begin{aligned}
&\mathbb{P}\left(\widehat{p}_m(X_1,\dots,X_m) \leq \alpha\right) \\
&\quad = \mathbb{P}\left(f(K_1,\dots,K_B)(X_1,\dots,X_m) \leq (B+1)\alpha - 1\right) \\
&\quad = \sum_{l=1}^{M_m} \mathbb{P}\left(f(K_1,\dots,K_B)(X_1,\dots,X_m) \leq (B+1)\alpha - 1 \,|\, f_{tot}(X_1,\dots,X_m) = l\right) \\
&\qquad\qquad\qquad \cdot \mathbb{P}\left(f_{tot}(X_1,\dots,X_m) = l\right) \\
&\quad = \frac{1}{M_m} \sum_{l=1}^{M_m} \sum_{i=0}^{\lfloor (B+1)\alpha - 1 \rfloor} \binom{B}{i} \left(\frac{l}{M_m}\right)^i \left(1 - \frac{l}{M_m}\right)^{B-i} \\
&\quad \leq \int_0^1 \sum_{i=0}^{\lfloor (B+1)\alpha - 1 \rfloor} \binom{B}{i} (x)^i (1-x)^{B-i} \, \lambda(\mathrm{d}x) \\
&\quad = \frac{\lfloor (B+1)\alpha - 1 \rfloor + 1}{B+1} \\
&\quad \leq \alpha,
\end{aligned}$$

where we approximated the sum by an integral and solved the integral using integration by parts. This completes the proof of Proposition 2.63. □

The p-value is underestimated by the choice we made. In fact, as described in Phipson and Smyth (2010), the level of the test would be preserved even if we chose the p-value slightly larger. This allows to construct a permutation test which is not only valid in level but actually achieves exact level.

The next proposition specifies the critical value that leads to the Monte-Carlo approximated resampling test.

**Proposition 2.64 (critical value of Monte-Carlo approximated resampling test)** *Let $\alpha \in (0,1)$, let $\mathcal{X}$ be a separable metric space, let $T = (T_m)_{m\in\mathbb{N}}$ be a test statistic on $\mathcal{X}$, let $g$ be a resampling method, let $B \in \mathbb{N}$, let $(K_i)_{i\in\mathbb{N}}$ be a sequence of uniformly distributed random variables on $\{1,\dots,M_m\}$ and let $(k_1,\dots,k_B)$ be a realization of $(K_1,\dots,K_B)$. For all $m \in \mathbb{N}$ define the function $c_m : \mathcal{X}^m \to \mathbb{R}$ satisfying that $c_m(x_1,\dots,x_m)$ is the*

$$\lceil (B+1)(1-\alpha) \rceil + \sum_{i=1}^{B} \mathbb{1}_{\{T_m(g_{m,k_i}(x_1,\dots,x_m)) = T_m(x_1,\dots,x_m)\}}\text{-}\textit{th largest value}$$

*in the vector $(T_m(g_{m,k_1}(x_1,\dots,x_m)),\dots,T_m(g_{m,k_B}(x_1,\dots,x_m)))$ if*

$$\lceil (B+1)(1-\alpha) \rceil + \sum_{i=1}^{B} \mathbb{1}_{\{T_m(g_{m,k_i}(x_1,\dots,x_m)) = T_m(x_1,\dots,x_m)\}} \leq B$$

*and $\infty$ otherwise. Then the hypothesis test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ defined for all $m \in \mathbb{N}$ and for all $(x_1,\dots,x_m) \in \mathcal{X}^m$ by*

$$\varphi(x_1,\dots,x_m) := \mathbb{1}_{\{T_m(x_1,\dots,x_m) \geq c_m(x_1,\dots,x_m)\}},$$

*is equal to the $\alpha$-Monte-Carlo approximated resampling test.*

**Proof** The following calculation is straight forward:

$$\begin{aligned}
&\mathbb{1}_{\{\widehat{p}_m(x_1,\ldots,x_m)\leq\alpha\}}\\
&\quad= \mathbb{1}_{\left\{\frac{1}{B}\sum_{i=1}^{B}\mathbb{1}_{\{T_m(g_{m,k_i}(x_1,\ldots,x_m))\geq T_m(x_1,\ldots,x_m)\}}\leq\frac{B+1}{B}\alpha-\frac{1}{B}\right\}}\\
&\quad= \mathbb{1}_{\left\{\frac{B+1}{B}(1-\alpha)\leq\frac{1}{B}\sum_{i=1}^{B}\mathbb{1}_{\{T_m(g_{m,k_i}(x_1,\ldots,x_m))<T_m(x_1,\ldots,x_m)\}}\right\}}\\
&\quad= \mathbb{1}_{\{T_m(x_1,\ldots,x_m)\geq c_m(x_1,\ldots,x_m)\}}\\
&\quad= \varphi_m(x_1,\ldots,x_m).
\end{aligned}$$

This completes the prove of Proposition 2.64. □

The Monte-Carlo approximated resampling test is closely related to the Monte-Carlo resampling distribution function. To see this observe that for large $B$ it holds for all $(x_1,\ldots,x_m)\in\mathcal{X}^m$ that

$$c_m(x_1,\ldots,x_m)\approx(\widehat{R}^{B}_{T_m}(x_1,\ldots,x_m))^{-1}(1-\alpha).$$

As mentioned above $c_m$ approximates the $(1-\alpha)$-quantile of the Monte-Carlo resampling distribution from above and gets closer as $B$ increases.

Chapter 3

# d-variable Hilbert-Schmidt independence criterion

## 3.1 Problem description

Our goal is to develop a non parametric hypothesis test to determine whether the components of a random vector $\mathbf{X} = (X^1, \dots, X^d)$ are mutually independent based on $m$ iid observations $\mathbf{X}_1, \dots, \mathbf{X}_m$ of the vector $\mathbf{X}$.

$X^1, \dots, X^d$ are mutually independent if and only if

$$\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d} = \mathbb{P}^{(X^1,\dots,X^d)}. \tag{3.1}$$

The central idea is to embed both $\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d}$ and $\mathbb{P}^{(X^1,\dots,X^d)}$ into an appropriate RKHS and then check whether the embedded elements are equal.

To keep an overview of all our assumptions, we summarize the setting used throughout the rest of this work.

**Setting 3.1 (dHSIC)**
*For all $j \in \{1, \dots, d\}$, let $\mathcal{X}^j$ be a separable metric space and denote by $\boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^d$ the product space. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and for every $j \in \{1, \dots, d\}$, let $X^j : \Omega \to \mathcal{X}^j$ be a random variable with law $\mathbb{P}^{X^j}$. Let $(\mathbf{X}_i)_{i \in \mathbb{N}}$ be a sequence of iid copies of $\mathbf{X} = (X^1, \dots, X^d)$. For $j \in \{1, \dots, d\}$, let $k^j : \mathcal{X}^j \times \mathcal{X}^j \to \mathbb{R}$ be a continuous, bounded, positive semi-definite and characteristic kernel on $\mathcal{X}^j$ and denote by $\mathcal{H}^j$ the corresponding RKHS. Let $\mathbf{k} = k^1 \otimes \cdots \otimes k^d$ be the tensor product of the kernels $k^j$ and let $\boldsymbol{\mathcal{H}} = \mathcal{H}^1 \otimes \cdots \otimes \mathcal{H}^d$ be the tensor product of the RKHSs $\mathcal{H}^j$. Let $\Pi : \mathcal{M}_f(\boldsymbol{\mathcal{X}}) \to \boldsymbol{\mathcal{H}}$ be the mean embedding function associated to $\mathbf{k}$.*

Observe that this setting has several nice consequences:

(i) $\boldsymbol{\mathcal{H}}$ is RKHS with reproducing kernel $\mathbf{k}$ (see Theorem A.5)

(ii) $\mathbf{k}$ is continuous and bounded (see Section 2.2.3)

(iii) $\boldsymbol{\mathcal{H}}$ is separable and only contains continuous functions (see Theorem 2.22)

(iv) $\Pi$ is injective because $\mathbf{k}$ is characteristic (see Definition 2.25)

## 3.2 dHSIC

The following section extends the Hilbert-Schmidt independence criterion (HSIC) from two variables as described by Gretton et al. (2007) to the case of $d$ variables.

The idea of the extension is to use the HSIC characterization via the mean embedding described by Smola et al. (2007).

**Definition 3.2 (dHSIC)**
*Assume Setting 3.1. Then, define the statistical functional*

$$\mathrm{dHSIC}\left(\mathbb{P}^{(X^1,\dots,X^d)}\right) := \left\|\Pi\left(\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^d}\right) - \Pi\left(\mathbb{P}^{(X^1,\dots,X^d)}\right)\right\|_{\mathcal{H}}^2$$

*and call it the d-variable Hilbert-Schmidt independence criterion (dHSIC).*

The essential property of dHSIC is stated in the next theorem.

**Theorem 3.3 (independence property of dHSIC)**
*Assume Setting 3.1. Then it holds that*

$$\mathrm{dHSIC} = 0 \quad\iff\quad \mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^d} = \mathbb{P}^{(X^1,\dots,X^d)}$$

**Proof** The proof of this statement follows from the definiteness of the norm and the fact that $\Pi$ is injective. □

In order to make dHSIC accessible for calculations, we express it in terms of the individual kernels $k^1,\dots,k^d$.

**Lemma 3.4 (expansion of dHSIC)**
*Assume Setting 3.1. Then it holds that*

$$\mathrm{dHSIC} = \mathbb{E}\left(\prod_{j=1}^d k^j\left(X_1^j, X_2^j\right)\right) + \mathbb{E}\left(\prod_{j=1}^d k^j\left(X_{2j-1}^j, X_{2j}^j\right)\right) - 2\mathbb{E}\left(\prod_{j=1}^d k^j\left(X_1^j, X_{j+1}^j\right)\right)$$

**Proof** Using the definition of the mean embedding we get

$$\begin{aligned}
\mathrm{dHSIC} &= \left\|\Pi\left(\mathbb{P}^{X^1}\otimes\cdots\otimes\mathbb{P}^{X^d}\right) - \Pi\left(\mathbb{P}^{(X^1,\dots,X^d)}\right)\right\|_{\mathcal{H}}^2 \\
&= \left\|\prod_{j=1}^d \mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right) - \mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1,\cdot\right)\right)\right\|_{\mathcal{H}}^2 \\
&= \left\|\prod_{j=1}^d \mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right)\right\|_{\mathcal{H}}^2 + \left\|\mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1,\cdot\right)\right)\right\|_{\mathcal{H}}^2 \\
&\quad - 2\left\langle\prod_{j=1}^d \mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right), \mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1,\cdot\right)\right)\right\rangle_{\mathcal{H}}
\end{aligned} \tag{3.2}$$

Next we simplify each term individually using the properties of the Bochner integral (see Appendix A.3) and the properties of tensor Hilbert spaces (see Section 2.2.3).

$$\begin{aligned}
\left\|\prod_{j=1}^{d} \mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right)\right\|_{\boldsymbol{\mathcal{H}}}^2 &= \prod_{j=1}^{d}\left\|\mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right)\right\|_{\mathcal{H}^j}^2 \\
&= \prod_{j=1}^{d}\left\langle \mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right), \mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right)\right\rangle_{\mathcal{H}^j} \\
&= \prod_{j=1}^{d} \mathbb{E}\left(\langle k^j\left(X_1^j,\cdot\right), k^j\left(X_2^j,\cdot\right)\rangle_{\mathcal{H}^j}\right) \\
&= \prod_{j=1}^{d} \mathbb{E}\left(k^j\left(X_1^j, X_2^j\right)\right) \\
&= \mathbb{E}\left(\prod_{j=1}^{d} k^j\left(X_{2j-1}^j, X_{2j}^j\right)\right)
\end{aligned} \tag{3.3}$$

$$\begin{aligned}
\left\|\mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1,\cdot\right)\right)\right\|_{\boldsymbol{\mathcal{H}}}^2 &= \left\langle \mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1,\cdot\right)\right), \mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1,\cdot\right)\right)\right\rangle_{\boldsymbol{\mathcal{H}}} \\
&= \mathbb{E}\Big(\langle \mathbf{k}\left(\mathbf{X}_1,\cdot\right), \mathbf{k}\left(\mathbf{X}_2,\cdot\right)\rangle_{\boldsymbol{\mathcal{H}}}\Big) \\
&= \mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1, \mathbf{X}_2\right)\right) \\
&= \mathbb{E}\left(\prod_{j=1}^{d} k^j\left(X_1^j, X_2^j\right)\right)
\end{aligned} \tag{3.4}$$

$$\begin{aligned}
\left\langle \prod_{j=1}^{d} \mathbb{E}\left(k^j\left(X_1^j,\cdot\right)\right), \mathbb{E}\left(\mathbf{k}\left(\mathbf{X}_1,\cdot\right)\right)\right\rangle_{\boldsymbol{\mathcal{H}}} &= \mathbb{E}\left(\left\langle \mathbb{E}\left(\prod_{j=1}^{d} k^j\left(X_{j+1}^j,\cdot\right)\right), \prod_{i=1}^{d} k^j\left(X_1^j,\cdot\right)\right\rangle_{\boldsymbol{\mathcal{H}}}\right) \\
&= \mathbb{E}\left(\left\langle \prod_{j=1}^{d} k^j\left(X_{j+1}^j,\cdot\right), \prod_{j=1}^{d} k^j\left(X_1^j,\cdot\right)\right\rangle_{\boldsymbol{\mathcal{H}}}\right) \\
&= \mathbb{E}\left(\prod_{j=1}^{d}\langle k^j\left(X_{j+1}^j,\cdot\right), k^j\left(X_1^j,\cdot\right)\rangle_{\mathcal{H}^j}\right) \\
&= \mathbb{E}\left(\prod_{j=1}^{d} k^j\left(X_1^j, X_{j+1}^j\right)\right)
\end{aligned} \tag{3.5}$$

Combining (3.2), (3.3), (3.4) and (3.5) completes the proof of Lemma 3.4. □

## 3.3 Estimating dHSIC

Next we define an estimator for dHSIC by estimating each of the expectation terms in Lemma 3.4 by a V-statistic.

**Definition 3.5 ($\widehat{\text{dHSIC}}$)**
*Assume Setting 3.1. Define the estimator* $\widehat{\text{dHSIC}} = (\widehat{\text{dHSIC}}_m)_{m\in\mathbb{N}}$ *in such a way that* $\widehat{\text{dHSIC}}_m : \boldsymbol{\mathcal{X}}^m \to \mathbb{R}$ *are the measurable functions with the property that for all* $m \in \{1,\dots,2d-1\}$ *it holds that* $\widehat{\text{dHSIC}}_m := 0$ *and for all* $m \in \{2d, 2d+1, \dots\}$ *and for all* $(\mathbf{x}_1,\dots,\mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ *it holds that*

$$\begin{aligned}\widehat{\text{dHSIC}}_m(\mathbf{x}_1,\dots,\mathbf{x}_m) := &\frac{1}{m^2}\sum_{\mathbf{M}_2(m)}\prod_{j=1}^{d} k^j\left(x^j_{i_1}, x^j_{i_2}\right)\\ &+\frac{1}{m^{2d}}\sum_{\mathbf{M}_{2d}(m)}\prod_{j=1}^{d} k^j\left(x^j_{i_{2j-1}}, x^j_{i_{2j}}\right)\\ &-\frac{2}{m^{d+1}}\sum_{\mathbf{M}_{d+1}(m)}\prod_{j=1}^{d} k^j\left(x^j_{i_1}, x^j_{i_{j+1}}\right).\end{aligned}$$

*We call* $\widehat{\text{dHSIC}}$ *the dHSIC V-estimator.*

Whenever it is clear from the context, we drop the functional arguments and just write $\widehat{\text{dHSIC}}_m$ instead of $\widehat{\text{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m)$.

Now, define $h : \boldsymbol{\mathcal{X}}^{2d} \to \mathbb{R}$ to be the function with the property that for all $\mathbf{z}_1,\dots,\mathbf{z}_{2d} \in \boldsymbol{\mathcal{X}}$ it holds

$$\begin{aligned}h(\mathbf{z}_1,\dots,\mathbf{z}_{2d}) = \frac{1}{(2d)!}\sum_{\pi\in S_{2d}}\Bigg[&\prod_{j=1}^{d} k^j\left(z^j_{\pi(1)}, z^j_{\pi(2)}\right) + \prod_{j=1}^{d} k^j\left(z^j_{\pi(2j-1)}, z^j_{\pi(2j)}\right)\\ &-2\prod_{j=1}^{d} k^j\left(z^j_{\pi(1)}, z^j_{\pi(j+1)}\right)\Bigg]\end{aligned} \tag{3.6}$$

where $S_{2d}$ is the set of permutations on $\{1,\dots,2d\}$. The function $h$ serves as a core function such that $\widehat{\text{dHSIC}}$ is a V-statistic. This is made precise in the following lemma.

**Lemma 3.6 (properties of h)**
*Assume Setting 3.1. It holds that*

*(i)* $h$ *is symmetric,*

*(ii)* $h$ *is continuous,*

*(iii)* $\exists C > 0$ *such that* $\forall \mathbf{z}_1,\dots,\mathbf{z}_{2d} \in \boldsymbol{\mathcal{X}}$: $|h(\mathbf{z}_1,\dots,\mathbf{z}_{2d})| < C$,

*(iv)* $V_m(h) = \widehat{\text{dHSIC}}_m$, *and*

*(v)* $\theta_h = \mathbb{E}\left(h(\mathbf{X}_1, \dots, \mathbf{X}_{2d})\right) = \text{dHSIC}$.

**Proof** (i) This is immediate by construction.

(ii) This follows from the continuity of the kernels $k^j$, which is assumed in Setting 3.1.

(iii) Under Setting 3.1 we assume that all $k^j$'s are bounded. Hence for all $j \in \{1, \dots, d\}$ let $C^j > 0$ such that for all $z_1, z_2 \in \mathcal{X}$ it holds that

$$|k^j(z_1, z_2)| < C^j.$$

Thus it is clear that for all $\mathbf{z}_1, \dots, \mathbf{z}_{2d} \in \boldsymbol{\mathcal{X}}$ it holds that

$$|h(\mathbf{z}_1, \dots, \mathbf{z}_{2d})| < 4 \prod_{j=1}^{d} C^j =: C.$$

(iv) Compute directly,

$$\begin{aligned}
V_m(h) &= \frac{1}{m^{2d}} \sum_{\mathbf{M}_{2d}(m)} h\left(\mathbf{X}_1, \dots, \mathbf{X}_{2p}\right) \\
&= \frac{1}{(2d)!} \sum_{\pi \in S_{2d}} \frac{1}{m^{2d}} \sum_{\mathbf{M}_{2d}(m)} \Bigg[ \prod_{j=1}^{d} k^j \left( X^j_{\pi(i_1)}, X^j_{\pi(i_2)} \right) + \prod_{j=1}^{d} k^j \left( X^j_{\pi({}_{2j-1})}, X^j_{\pi(i_{2j})} \right) \\
&\qquad\qquad - 2 \prod_{j=1}^{d} k^j \left( X^j_{\pi(i_1)}, X^j_{\pi(i_{j+1})} \right) \Bigg] \\
&= \frac{1}{(2d)!} \sum_{\pi \in S_{2d}} \Bigg[ \frac{1}{m^2} \sum_{\mathbf{M}_2(m)} \prod_{j=1}^{d} k^j \left( X^j_{\pi(i_1)}, X^j_{\pi(i_2)} \right) \\
&\qquad\qquad + \frac{1}{m^{2d}} \sum_{\mathbf{M}_{2d}(m)} \prod_{j=1}^{d} k^j \left( X^j_{\pi({}_{2j-1})}, X^j_{\pi(i_{2j})} \right) \\
&\qquad\qquad - \frac{2}{m^{d+1}} \sum_{\mathbf{M}_{d+1}(m)} \prod_{j=1}^{d} k^j \left( X^j_{\pi(i_1)}, X^j_{\pi(i_{j+1})} \right) \Bigg] \\
&= \widehat{\text{dHSIC}}_m \,.
\end{aligned}$$

(v) Again computing directly,

$$
\begin{aligned}
\mathbb{E}\left(h(\mathbf{X}_1,\dots,\mathbf{X}_{2d})\right) &= \frac{1}{(2d)!}\sum_{\pi\in S_{2d}}\Bigg[\mathbb{E}\left(\prod_{j=1}^{d} k^j\left(X^j_{\pi(1)},X^j_{\pi(2)}\right)\right)\\
&\qquad + \mathbb{E}\left(\prod_{j=1}^{d} k^j\left(z^j_{\pi(2j-1)},z^j_{\pi(2j)}\right)\right)\\
&\qquad - 2\mathbb{E}\left(\prod_{j=1}^{d} k^j\left(z^j_{\pi(1)},z^j_{\pi(j+1)}\right)\right)\Bigg]\\
&= \frac{1}{(2d)!}\sum_{\pi\in S_{2d}}\Bigg[\mathbb{E}\left(\prod_{j=1}^{d} k^j\left(X^j_1,X^j_2\right)\right) + \mathbb{E}\left(\prod_{j=1}^{d} k^j\left(z^j_{2j-1},z^j_{2j}\right)\right)\\
&\qquad - 2\mathbb{E}\left(\prod_{j=1}^{d} k^j\left(z^j_1,z^j_{j+1}\right)\right)\Bigg]\\
&= \mathrm{dHSIC}\,.
\end{aligned}
$$

This completes the proof of Lemma 3.6 □

This means that using $h$ as a core function we can express $\widehat{\mathrm{dHSIC}}$ as a V-statistic, which allows us to apply the asymptotic results that have been developed in Section 2.3 to analyze $\widehat{\mathrm{dHSIC}}$.

## 3.4 Implementation details

In this section we describe how one can efficiently implement the dHSIC V-estimator $\widehat{\mathrm{dHSIC}}$. One efficient implementation is given in Algorithm 1, where the function `Sum` takes the sum of all elements in a matrix, the function `ColumnSum` takes the sums of the columns of a matrix and the operator $*$ is the element-wise multiplication operator.

For the two variable case (i.e. the standard HSIC setting) Gretton et al. (2007) express $\widehat{\mathrm{dHSIC}}$ using the Gram matrices as

$$\widehat{\mathrm{dHSIC}}(\mathbf{x}_1,\dots,\mathbf{x}_m) = \frac{1}{m^2}\,\mathrm{trace}(\mathbf{K}^1\mathbf{H}\mathbf{K}^2\mathbf{H})$$

where $\mathbf{K}^j$ is the Gram matrix of $k^j$ given $\mathbf{x}_1,\dots,\mathbf{x}_m$ and $\mathbf{H}$ is the $(m\times m)$-matrix satisfying for all $i,j\in\{1,\dots,m\}$ that

$$\mathbf{H}_{i,j} = \begin{cases} -\frac{1}{m} & \text{if } i\neq j\\ \frac{m-1}{m} & \text{if } i=j. \end{cases}$$

**Algorithm 1** computing the dHSIC V-estimator

```
procedure DHSIC($\mathbf{x}_1, \dots, \mathbf{x}_m$)
    for $j = 1 : d$ do
        $\mathbf{K}^j \leftarrow$ Gram matrix of kernel $k^j$ given $\mathbf{x}_1, \dots, \mathbf{x}_m$
    term1 $\leftarrow$ $(m \times m)$-matrix with all entries equal to 1
    term2 $\leftarrow$ 1
    term3 $\leftarrow$ $(1 \times m)$-matrix with all entries equal to $\frac{2}{m}$
    for $j = 1 : d$ do
        term1 $\leftarrow$ term1 $*$ $\mathbf{K}^j$
        term2 $\leftarrow$ $\frac{1}{m^2} \cdot$ term2 $\cdot$ Sum($\mathbf{K}^j$)
        term3 $\leftarrow$ $\frac{1}{m} \cdot$ term3 $*$ ColumnSum($\mathbf{K}^j$)
    term1 $\leftarrow$ Sum(term1)
    term3 $\leftarrow$ Sum(term3)
    dHSIC $\leftarrow$ $\frac{1}{m^2} \cdot$ term1 $\cdot$ term2 $\cdot$ term3
    return dHSIC
```

While this representation is useful for computations by hand it should not be implemented directly as it leads to inefficient code. A better solution is to use following trace formula

$$\operatorname{trace}(\mathbf{A}\mathbf{B}^\top) = \sum_{i,j} \mathbf{A}_{i,j}\mathbf{B}_{i,j}. \tag{3.7}$$

The complexity of this operation is only $\mathcal{O}\left(m^2\right)$, while even the most efficient implementation of the matrix multiplication has a complexity strictly large than this. If the trace formula (3.7) is used the resulting code is similar to the one given in Algorithm 1.

Chapter 4

# Hypothesis tests based on dHSIC

As before, we write $\mathbf{X} = (X^1, \dots, X^d)$ and $\boldsymbol{\mathcal{X}}$ denotes the product state space, see Setting 3.1. In this section we derive four statistical hypothesis tests to test the null hypothesis

$$H_0 \coloneqq \Big\{ \mathbb{P}^{\mathbf{X}} \in \mathcal{P}(\boldsymbol{\mathcal{X}}) \,\Big|\, \exists \text{ measurable } \mathbf{X} : \Omega \to \boldsymbol{\mathcal{X}} \text{ with law } \mathbb{P}^{\mathbf{X}} \\ \text{such that } \mathbb{P}^{\mathbf{X}} = \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d} \Big\} \tag{4.1}$$

against the alternative

$$H_A \coloneqq \Big\{ \mathbb{P}^{\mathbf{X}} \in \mathcal{P}(\boldsymbol{\mathcal{X}}) \,\Big|\, \forall \text{ measurable } \mathbf{X} : \Omega \to \boldsymbol{\mathcal{X}} \text{ with law } \mathbb{P}^{\mathbf{X}} \\ \text{it holds that } \mathbb{P}^{\mathbf{X}} \neq \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d} \Big\}. \tag{4.2}$$

Since dHSIC is 0 under $H_0$ and positive otherwise we choose to use $m \cdot \widehat{\text{dHSIC}}_m$ as test statistic. Similarly as in (2.66) we define a test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ satisfying for all $m \in \{1, \dots, 2d-1\}$ that $\varphi_m \coloneqq 0$ and for all $m \in \{2d, 2d+1, \dots\}$ and for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that

$$\varphi_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \coloneqq \begin{cases} 0 & \text{if } m \cdot \widehat{\text{dHSIC}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \leq c_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \\ 1 & \text{if } m \cdot \widehat{\text{dHSIC}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) > c_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \end{cases} \tag{4.3}$$

where the threshold $c = (c_m)_{m\in\mathbb{N}}$ remains to be chosen.

Assume there exists a function $G : \mathcal{P}(\boldsymbol{\mathcal{X}}) \times \mathbb{R} \to [0, 1]$ satisfying for all $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$ and for all $t \in \mathbb{R}$ that

$$\lim_{m\to\infty} \mathbb{P}\Big( m \cdot \widehat{\text{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m) \leq t \Big) = G\left(\mathbb{P}^{\mathbf{X}}\right)(t). \tag{4.4}$$

In particular this means that for every fixed $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$ the random variable $m \cdot \widehat{\text{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m)$ has a limiting distribution. The idea is to find functions

$(G_m)_{m\in\mathbb{N}}$ with $G_m : \boldsymbol{\mathcal{X}}^m \times \mathbb{R} \to [0,1]$ satisfying for all $\mathbf{X}_1, \mathbf{X}_2, \ldots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$ it holds $\mathbb{P}$-a.s. that

$$\lim_{m\to\infty} G_m(\mathbf{X}_1, \ldots, \mathbf{X}_m)(t) = G\left(\mathbb{P}^{\mathbf{X}}\right)(t).$$

Then, we define the threshold to satisfy for all $(\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that

$$c_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) := (G_m(\mathbf{x}_1, \ldots, \mathbf{x}_m))^{-1}(1-\alpha).$$

Using a classical argument involving Slutsky's theorem it can be shown that the corresponding test has pointwise asymptotic level $\alpha$. If $G_m$ additionally satisfies for all $\mathbf{X}_1, \mathbf{X}_2, \ldots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_A$ that

$$\lim_{m\to\infty} G_m(\mathbf{X}_1, \ldots, \mathbf{X}_m)(t) = G\left(\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d}\right)(t),$$

then one can additionally show that this test is pointwise consistent.

In Section 4.1 we consider some of the asymptotic properties of the test statistic $m \cdot \widehat{\text{dHSIC}}_m$. In particular we show the existence of the function $G : \mathcal{P}(\boldsymbol{\mathcal{X}}) \times \mathbb{R} \to [0,1]$ satisfying (4.4). We then construct four hypothesis tests of the form (4.3). The first two are a permutation test and a bootstrap test which are discussed in Section 4.2. Both tests are based on resampling tests and hence do not relay on an explicit knowledge of the function $G$. In Section 4.3 we consider a third test which is based on an approximation of $G$ as a gamma distribution and finally in Section 4.4 we give a fourth test which uses the explicit form of $G$.

The following table summarizes the properties of the four tests.

| **Hypothesis test** | **consistency** | **level** | **speed** |
|---|---|---|---|
| Permutation[1] | unknown | valid (Prop. 4.5) | slow |
| Bootstrap[2] | pointwise (Prop. 4.10) | pointwise asymptotic (Prop. 4.9) | slow |
| Gamma approximation | no guarantee | no guarantee | fast |
| Eigenvalue approach[3] | pointwise (Conjecture 4.20) | pointwise asymptotic (Conjecture 4.20) | medium |

[1] For implementation purposes one can use the Monte-Carlo approximation. This leads to a reasonably fast implementation with similar level and consistency results. Further details are given at the end of Section 4.2.1 and Section 4.2.2.

[2] See footnote 1.

[3] The final results are missing a small step in the approximation.

## 4.1 Asymptotic behavior of the test statistic

All results in this section make use of the theory of U-and V-statistics introduced in Section 2.3.

### 4.1.1 Under the alternative hypothesis

We first determine the asymptotic distribution of $\sqrt{m}(\widehat{\mathrm{dHSIC}}_m - \mathrm{dHSIC})$ using Theorem 2.41 from the theory of V-statistics. This leads to the following theorem (for the two variable HSIC, see Gretton et al., 2007, Theorem 1).

**Theorem 4.1 (asymptotic distribution of $\sqrt{m} \cdot \widehat{\mathbf{dHSIC}}_m$ under $H_A$)**
*Assume Setting 3.1 and recall* (2.7). *If* $\xi_1(h) > 0$, *then under* $H_A$ *it holds that*

$$\sqrt{m}\left(\widehat{\mathrm{dHSIC}}_m - \mathrm{dHSIC}\right) \xrightarrow{d} \mathcal{N}\left(0, 4d^2\xi_1(h)\right)$$

*as* $m \to \infty$. *If* $\xi_1(h) = 0$, *then*

$$\sqrt{m}\left(\widehat{\mathrm{dHSIC}}_m - \mathrm{dHSIC}\right) \xrightarrow{d} 0$$

*as* $m \to \infty$.

**Proof** Use Lemma 3.6 to observe that $\widehat{\mathrm{dHSIC}}$ is simply the V-statistic $V_m(h)$ with $\theta_h =$ dHSIC. Moreover, again by Lemma 3.6 it holds that $h$ is bounded and continuous. If $\xi_1(h) > 0$ then we can apply Theorem 2.41 to see that,

$$\sqrt{m}\left(\widehat{\mathrm{dHSIC}}_m - \mathrm{dHSIC}\right) \xrightarrow{d} \mathcal{N}\left(0, (2d)^2\xi_1(h)\right)$$

as $m \to \infty$. Next assume $\xi_1(h) = 0$, then by Theorem 2.39 it holds that

$$\mathbb{E}\left(m\left(\widehat{\mathrm{dHSIC}}_m - \mathrm{dHSIC}\right)^2\right) = m\,\mathrm{Var}\left(\widehat{\mathrm{dHSIC}}\right) = \mathcal{O}\left(m^{-1}\right)$$

and since convergence in second moment implies convergence in distribution this completes the proof of Theorem 4.1. □

As a direct corollary we get that $m \cdot \widehat{\mathrm{dHSIC}}_m$ diverges under $H_A$ in the following sense.

**Corollary 4.2 (asymptotic distribution of $m \cdot \widehat{\mathbf{dHSIC}}_m$ under $H_A$)**
*Assume Setting 3.1. Then under* $H_A$ *it holds for all* $t \in \mathbb{R}$ *that*

$$\lim_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m) \leq t\right) = 0.$$

**Proof** Let $t \in \mathbb{R}$ and $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_A$ then

$$\begin{aligned}&\mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m) \leq t\right)\\ &\quad = \mathbb{P}\left(\sqrt{m}(\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m) - \mathrm{dHSIC}) \leq \frac{t}{\sqrt{m}} - \sqrt{m}\,\mathrm{dHSIC}\right).\end{aligned}$$

So Corollary B.3 together with Theorem 4.1 completes the proof of Corollary 4.2. □

This Corollary is the key result needed to show consistency for both the bootstrap test and the eigenvalue approach based test.

### 4.1.2 Under the null hypothesis

The next step is to determine the asymptotic distribution of $m \cdot \widehat{\mathrm{dHSIC}}_m$ under $H_0$. This is done in the following theorem (for the two variable HSIC, see Gretton et al., 2007, Theorem 2).

**Theorem 4.3 (asymptotic distribution of $m \cdot \widehat{\mathbf{dHSIC}}_m$ under $H_0$)**
*Assume Setting 3.1 and recall* (2.7)*. If $\xi_2(h) > 0$, let $(Z_i)_{i\in\mathbb{N}}$ be a sequence of independent standard normal random variables on $\mathbb{R}$, let $T_{h_2} \in L(L^2(\mathbb{P}^{(X^1,\dots,X^d)}, |\cdot|_{\mathbb{R}}))$ with the property that for every $f \in L^2(\mathbb{P}^{(X^1,\dots,X^d)}, |\cdot|_{\mathbb{R}})$ and for every $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ it holds that*

$$(T_{h_2}(f))(\mathbf{x}) = \int_{\boldsymbol{\mathcal{X}}} h_2(\mathbf{x},\mathbf{y}) f(\mathbf{y})\, \mathbb{P}^{(X^1,\dots,X^d)}(d\mathbf{y})$$

*and let $(\lambda_i)_{i\in\mathbb{N}}$ be the eigenvalues of $T_{h_2}$, then under $H_0$ it holds that*

$$m \cdot \widehat{\mathrm{dHSIC}}_m \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2$$

*as $m \to \infty$. If $\xi_2(h) = 0$ then under $H_0$ it holds that*

$$m \cdot \widehat{\mathrm{dHSIC}}_m \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i$$

*as $m \to \infty$.*

**Proof** Use Lemma 3.6 to observe that $\widehat{\mathrm{dHSIC}}_m$ is simply the V-statistic $V_m(h)$ with $\theta_h = \mathrm{dHSIC}$. By Lemma C.3 it holds that $\xi_1(h) = 0$ under $H_0$ and moreover, again by Lemma 3.6 it holds that $h$ is bounded and continuous. If $\xi_2(h) > 0$, we can apply Theorem 2.43 to see that,

$$m \cdot \widehat{\mathrm{dHSIC}}_m \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2$$

as $m \to \infty$. If $\xi_2(h) = 0$, we can apply Theorem 2.39 to see that $\lim_{n\to\infty} \mathrm{Var}(n \cdot \widehat{\mathrm{dHSIC}}_n) = 0$. Combining this with Theorem 2.40 and Lemma 2.42 hence leads to

$$n \cdot \widehat{\mathrm{dHSIC}}_n \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i.$$

as $m \to \infty$, which completes the proof of Theorem 4.3. □

Unfortunately, the asymptotic distribution under $H_0$ depends on whether $\xi_2(h) > 0$ or $\xi_2(h) = 0$. This means that we always have to consider both cases, which becomes quite repetitive. In the following analysis we will therefore always assume that $\xi_2(h) > 0$ when applying Theorem 4.3. The case $\xi_2(h) = 0$ has to be analyzed in a similar fashion.

## 4.2 Permutation/Bootstrap

In this section we construct two types of resampling tests for dHSIC. The first is a permutation test and the second is a bootstrap test. The main ingredient for a resampling test is the resampling method.

Assume Setting 3.1 and fix $m \in \mathbb{N}$. For every function $\boldsymbol{\psi} = (\psi^1, \dots, \psi^d)$ such that for all $i \in \{1, \dots, d\}$ it holds that $\psi^i : \{1, \dots, m\} \to \{1, \dots, m\}$ define the function $g_{m,\boldsymbol{\psi}} : \boldsymbol{\mathcal{X}}^m \to \boldsymbol{\mathcal{X}}^m$ satisfying for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that

$$g_{m,\boldsymbol{\psi}}(\mathbf{x}_1, \dots, \mathbf{x}_m) := \left(\mathbf{x}^{\boldsymbol{\psi}}_{m,1}, \dots, \mathbf{x}^{\boldsymbol{\psi}}_{m,m}\right) \tag{4.5}$$

where $\mathbf{x}^{\boldsymbol{\psi}}_{m,i} := \left(x^1_{\psi^1(i)}, \dots, x^d_{\psi^d(i)}\right)$. The diagram in (4.6) illustrates how $g_{m,\boldsymbol{\psi}}$ acts on the sample $(\mathbf{x}_1, \dots, \mathbf{x}_m)$.

$$\begin{array}{c|ccc} \mathbf{x}_1 & x^1_1 & \cdots & x^d_1 \\ \vdots & \vdots & & \vdots \\ \mathbf{x}_m & x^1_m & \cdots & x^d_m \end{array} \quad \overset{g_{m,\boldsymbol{\psi}}}{\longrightarrow} \quad \begin{array}{c|ccc} \mathbf{x}^{\boldsymbol{\psi}}_{m,1} & x^1_{\psi^1(1)} & \cdots & x^d_{\psi^d(1)} \\ \vdots & \vdots & & \vdots \\ \mathbf{x}^{\boldsymbol{\psi}}_{m,m} & x^1_{\psi^1(m)} & \cdots & x^d_{\psi^d(m)} \end{array} \tag{4.6}$$

Define

$$B_m := \left\{\psi : \{1, \dots, m\} \to \{1, \dots, m\} \mid \psi \text{ is a function}\right\} \tag{4.7}$$

then for a subset $A_m \subseteq B^d_m$ using the terminology of Section 2.4.2 we can define a resampling method

$$g := \left((g_{m,\boldsymbol{\psi}})_{\boldsymbol{\psi} \in A_m}\right)_{m \in \mathbb{N}}. \tag{4.8}$$

In the next two section we consider two explicit tests based on this resampling method.

### 4.2.1 Permutation test

The permutation test is the resampling test corresponding to the resampling method in (4.8) with $A_m = (S_m)^d$, where $S_m$ is the set of permutations on $\{1, \dots, m\}$.

**Definition 4.4 (permutation test for dHSIC)**
*Assume Setting 3.1, assume $\alpha \in (0, 1)$ and for all $m \in \mathbb{N}$ and for all $\boldsymbol{\psi} \in (S_m)^d$ let $g_{m,\boldsymbol{\psi}}$ be defined as in (4.5). Moreover, for all $m \in \{2d, 2d+1, \dots\}$ let $\widehat{R}_m : \boldsymbol{\mathcal{X}}^m \times \mathbb{R} \to [0, 1]$ be the resampling distribution functions defined for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ and for all $t \in \mathbb{R}$ by*

$$\widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)(t) := \frac{1}{(m!)^d} \sum_{\boldsymbol{\psi} \in (S_m)^d} \mathbb{1}_{\{m \cdot \widehat{\mathrm{dHSIC}}_m(g_{m,\boldsymbol{\psi}}(\mathbf{x}_1, \dots, \mathbf{x}_m)) \leq t\}}.$$

*Then the $\alpha$-resampling hypothesis test $\varphi = (\varphi_m)_{m \in \mathbb{N}}$ defined for all $m \in \{1, \dots, 2d-1\}$ by $\varphi_m := 0$ and for all $m \in \{2d, 2d+1, \dots\}$ and for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by*

$$\varphi_m(\mathbf{x}_1, \dots, \mathbf{x}_m) := \mathbb{1}_{\left\{m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) > (\widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m))^{-1}(1-\alpha)\right\}},$$

*is called $\alpha$-permutation test for dHSIC.*

In literature, resampling tests, that similar to the permutation test for dHSIC, are based on resampling schemes constructed by permutations are called permutation tests. As an immediate consequence of Theorem 2.59 it follows that $\varphi$ has valid level.

**Proposition 4.5 (permutation test for dHSIC has valid level)**
*Assume Setting 3.1 and let $H_0$ and $H_A$ be defined as in (4.1) and (4.2). Then for all $\alpha \in (0,1)$ the $\alpha$-permutation test for dHSIC has level $\alpha$ when testing $H_0$ against $H_A$.*

**Proof** Fix $m \in \mathbb{N}$, under $H_0$, i.e. $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$, it holds that the individual coordinates of $\mathbf{X}_i$ are independent. Hence, for all $\boldsymbol{\psi} \in (S_m)^d$ it holds that $(\mathbf{X}_1, \dots, \mathbf{X}_m)$ is equal in distribution to $(\mathbf{X}_1^{\boldsymbol{\psi}}, \dots, \mathbf{X}_m^{\boldsymbol{\psi}})$, so in particular, we have that

$$g_{m,\boldsymbol{\psi}}(\mathbf{X}_1, \dots, \mathbf{X}_m) \text{ is equal in distribution to } (\mathbf{X}_1, \dots, \mathbf{X}_m). \tag{4.9}$$

Moreover since $(S_m)^d$ has a group structure we can apply Theorem 2.59 to get that $\varphi$ has level $\alpha$, which completes the proof of Proposition 4.5. □

It turns out that proving that the permutation test for dHSIC has pointwise consistency is rather difficult because for $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_A$ it is not straight forward to link the resampling distribution function $\widehat{R}_m(\mathbf{X}_1, \dots, \mathbf{X}_m)$ with the limit distribution function $G(\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d})$.

The size of the set $(S_m)^d$ is given by $(m!)^d$ which grows very fast. For implementation purposes we therefore generally use a Monte-Carlo approximated version as defined in Definition 2.62. Given that the underlying probability distribution $\mathbb{P}^{\mathbf{X}}$ is continuous we satisfy all requirements of Proposition 2.63, which implies that the Monte-Carlo approximated permutation test for dHSIC also has valid level. For more details see Section 4.5.

### 4.2.2 Bootstrap test

The bootstrap test is the resampling test corresponding to the resampling method in (4.8) with $A_m = B_m^d$.

**Definition 4.6 (bootstrap test for dHSIC)**
*Assume Setting 3.1, assume $\alpha \in (0,1)$ and for all $m \in \mathbb{N}$ and for all $\boldsymbol{\psi} \in B_m^d$ let the function $g_{m,\boldsymbol{\psi}}$ be defined as in (4.5). Moreover, for all $m \in \{2d, 2d+1, \dots\}$ let $\widehat{R}_m : \boldsymbol{\mathcal{X}}^m \times \mathbb{R} \to [0,1]$ be the resampling distribution functions defined for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ and for all $t \in \mathbb{R}$ by*

$$\widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)(t) := \frac{1}{m^{md}} \sum_{\boldsymbol{\psi} \in B_m^d} \mathbb{1}_{\{m \cdot \widehat{\text{dHSIC}}_m(g_{m,\boldsymbol{\psi}}(\mathbf{x}_1, \dots, \mathbf{x}_m)) \leq t\}}.$$

*Then the resampling hypothesis test $\varphi = (\varphi_m)_{m \in \mathbb{N}}$ defined for all $m \in \{1, \dots, 2d-1\}$ by $\varphi_m := 0$ and for all $m \in \{2d, 2d+1, \dots\}$ and for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by*

$$\varphi_m(\mathbf{x}_1, \dots, \mathbf{x}_m) := \mathbb{1}_{\left\{m \cdot \widehat{\text{dHSIC}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) > (\widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m))^{-1}(1-\alpha)\right\}}$$

*is called $\alpha$-bootstrap test for dHSIC.*

In order to prove level and consistency results we need to introduce some additional notation. For $j \in \{1, \dots, d\}$ denote by $\widehat{\mathbb{P}}_m^{X^j}$ the empirical distribution corresponding to $X_1^j, \dots, X_m^j$, i.e.

$$\widehat{\mathbb{P}}_m^{X^j} = \frac{1}{m} \sum_{i=1}^{m} \delta_{X_i^j}.$$

**Definition 4.7 (empirical product distribution function)**
*Assume Setting 3.1, then the function $\widehat{F}_m : \boldsymbol{\mathcal{X}}^m \times \mathbb{R}^d \to [0,1]$ satisfying for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ and for all $\mathbf{t} \in \mathbb{R}^d$ that*

$$\widehat{F}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)(\mathbf{t}) := \prod_{j=1}^{d} \left( \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{x_i^j \leq t^j\}} \right)$$

*is called the empirical product distribution function.*

It can be easily shown that the distribution function corresponding to the distribution $\widehat{\mathbb{P}}_m^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_m^{X^d}$ is given by the empirical product distribution function $\widehat{F}_m$.

The following proposition shows that random draws from the resampling distribution corresponds to independent draws from the empirical product distribution $\widehat{\mathbb{P}}_m^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_m^{X^d}$.

**Proposition 4.8 (bootstrapping property)**
*Assume Setting 3.1, let $m \in \mathbb{N}$, and for all $\boldsymbol{\psi} \in B_m^d$ let $g_{m,\boldsymbol{\psi}}$ be defined as in (4.5), let $\boldsymbol{\Psi}$ be a random variable with uniform distribution on $B_m^d$ and let $\widehat{F}_m$ be the empirical product distribution function. Then it holds for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that*

$$g_{m,\boldsymbol{\Psi}}(\mathbf{x}_1, \dots, \mathbf{x}_m) = \left( \mathbf{x}_{m,1}^{\boldsymbol{\Psi}}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}} \right)$$

*are $m$ iid random variables with distribution function $\widehat{F}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)$.*

**Proof** Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ be the probability space such that $\boldsymbol{\Psi} = (\Psi^1, \dots, \Psi^d) : \tilde{\Omega} \to B_n^d$. Then, by the properties of the uniform distribution it holds that $\Psi^1, \dots, \Psi^d$ are iid with uniform distribution on $B_m$. This implies that for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$, for all $i \in \{1, \dots, m\}$ and for all $\mathbf{t} \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
\tilde{\mathbb{P}}\left(\mathbf{x}_{m,i}^{\boldsymbol{\Psi}} \leq \mathbf{t}\right) &= \prod_{j=1}^{d} \tilde{\mathbb{P}}\left(x_{m,\Psi^j(i)}^j \leq t^j\right) \\
&= \prod_{j=1}^{d} \left( \frac{1}{|B_m|} \sum_{\psi \in B} \mathbb{1}_{\{x_{m,\psi(i)}^j \leq t^j\}} \right) \\
&= \prod_{j=1}^{d} \left( \frac{1}{m} \sum_{l=1}^{m} \mathbb{1}_{\{x_{m,i}^j \leq t^j\}} \right).
\end{aligned}$$

Hence, it holds for all $i \in \{1, \dots, m\}$ that $\mathbf{x}^{\mathbf{\Psi}}_{m,i}$ has distribution function $\widehat{F}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)$. Moreover, by an explicit calculation it holds for all $i, j \in \{1, \dots, m\}$ with $i \neq j$ and for all $\mathbf{k}, \mathbf{l} \in \{1, \dots, m\}^d$ that

$$\tilde{\mathbb{P}}\big(\mathbf{\Psi}(i) = \mathbf{k}, \mathbf{\Psi}(j) = \mathbf{l}\big) = \tilde{\mathbb{P}}\big(\mathbf{\Psi}(i) = \mathbf{k}\big)\tilde{\mathbb{P}}\big(\mathbf{\Psi}(j) = \mathbf{l}\big),$$

which implies that $\mathbf{\Psi}(i)$ is independent of $\mathbf{\Psi}(j)$. Consequently, it also holds that $\mathbf{x}^{\mathbf{\Psi}}_{m,i}$ is independent of $\mathbf{x}^{\mathbf{\Psi}}_{j}$. This finally proves that for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ it holds that

$$g_{m,\mathbf{\Psi}}(\mathbf{x}_1, \dots, \mathbf{x}_m) = \big(\mathbf{x}^{\mathbf{\Psi}}_{m,1}, \dots, \mathbf{x}^{\mathbf{\Psi}}_{m,m}\big)$$

are iid random variables with distribution function $\widehat{F}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)$, which completes the proof of Proposition 4.8. □

We can now prove that the $\alpha$-bootstrap test for dHSIC has pointwise asymptotic level $\alpha$.

**Proposition 4.9 (bootstrap test for dHSIC has pointwise asymptotic level)** *Assume Setting 3.1 and let $H_0$ and $H_A$ be defined as in* (4.1) *and* (4.2)*. Then for all $\alpha \in (0, 1)$ the $\alpha$-bootstrap test for mHSIC has pointwise asymptotic level $\alpha$ when testing $H_0$ against $H_A$.*

**Proof** Let $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$ be fixed and begin by introducing the following notation:

· Let $(Z_j)_{j\in\mathbb{N}}$ be a sequence of independent standard normal random variables on $\mathbb{R}$, let $T_{h_2} \in L(L^2(\mathbb{P}^{\mathbf{X}}, |\cdot|_{\mathbb{R}}))$ with the property that for every $f \in L^2(\mathbb{P}^{\mathbf{X}}, |\cdot|_{\mathbb{R}})$ and for every $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ it holds that

$$\left(T_{h_2}(f)\right)(\mathbf{x}) = \int_{\boldsymbol{\mathcal{X}}} h_2(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \, \mathbb{P}^{\mathbf{X}}(\mathrm{d}\mathbf{y})$$

and let $(\lambda_j)_{j\in\mathbb{N}}$ be the eigenvalues of $T_{h_2}$.

· Let $\widehat{F}_m$ be the empirical product distribution function and define for all $\mathbf{t} \in \mathbb{R}^d$ the population product distribution function by

$$F(\mathbf{t}) := \left(\mathbb{P}^{\mathbf{X}}\right)\left((-\infty, t^1] \times \cdots \times (-\infty, t^d]\right) = \prod_{j=1}^{d} \mathbb{P}\left(X^j \leq t^j\right)$$

By Theorem 4.3 it follows that

$$m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m) \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \tag{4.10}$$

as $m \to \infty$. Moreover, applying the Glivenko-Cantelli theorem (e.g. van der Vaart, 1998, Theorem 19.1), which extends the strong law of large numbers for empirical distributions to uniform convergence, shows that there exists a subset $A_0 \subseteq \Omega$ such that $\mathbb{P}(A_0) = 1$ and such that for all $\omega \in A_0$ it holds for all $\mathbf{t} \in \mathbb{R}^d$ that

$$\lim_{m\to\infty} \widehat{F}_m(\mathbf{X}_1(\omega), \dots, \mathbf{X}_m(\omega))(\mathbf{t}) = F(\mathbf{t}). \tag{4.11}$$

Next, for all $m \in \mathbb{N}$ let $\boldsymbol{\Psi}_m$ be a uniformly distributed random variable on $B_m^d$ independent of $\mathbf{X}$ and $(\mathbf{X}_i)_{i\in\mathbb{N}}$. Then, by Proposition 4.8 it holds for all $m \in \mathbb{N}$ and for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that

$$g_{m,\boldsymbol{\Psi}_m}(\mathbf{x}_1, \dots, \mathbf{x}_m) = \left(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m}\right)$$

are iid random variables with distribution function $\widehat{F}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)$. Hence, for fixed $\omega \in A_0$ and for all $i \in \mathbb{N}$ define $\mathbf{x}_i \coloneqq \mathbf{X}_i(\omega)$, then by (4.11) it holds that

$$\mathbf{x}_{m,i}^{\boldsymbol{\Psi}_m} \xrightarrow{d} \mathbf{X}$$

as $m \to \infty$. Hence, we are in the same setting as described in Setting 2.44.

Since both $\mathbb{P}^{\mathbf{X}} \in H_0$ and $\widehat{\mathbb{P}}_m^{X^1} \otimes \cdots \otimes \widehat{\mathbb{P}}_m^{X^d} \in H_0$ it holds by Lemma C.3 for all $\mathbf{z} \in \boldsymbol{\mathcal{X}}$ that

$$h_1(\mathbf{z}) = \mathbb{E}\left(h(\mathbf{z}, \mathbf{X}_2, \dots, \mathbf{X}_{2d})\right) = 0$$

and for all $m \in \{2d, 2d+1, \dots\}$ and for all $\mathbf{z} \in \boldsymbol{\mathcal{X}}$ that

$$h_1^m(\mathbf{z}) = \mathbb{E}\left(h(\mathbf{z}, \mathbf{x}_{m,2}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,2d}^{\boldsymbol{\Psi}_m})\right) = 0,$$

where $h$ is defined as in (3.6). Moreover, it holds by Theorem 3.3 that

$$\theta_h = \mathbb{E}\left(h(\mathbf{X}_1, \dots, \mathbf{X}_{2d})\right) = \text{dHSIC}\left(\mathbb{P}^{\mathbf{X}}\right) = 0.$$

We therefore satisfy all requirements of Theorem 2.48 (the condition $\xi_2(h) > 0$ is assumed, see remark after Theorem 4.3) and get that

$$m \cdot \widehat{\text{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m}) = m\tilde{V}_m(h) \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \tag{4.12}$$

as $m \to \infty$.

Let $G : \mathbb{R} \to (0,1)$ be the distribution function of $\binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2$, then by (4.12) it holds for all $t \in \mathbb{R}$ that

$$\begin{aligned}
\lim_{m\to\infty} \widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)(t) &= \lim_{m\to\infty} \frac{1}{m^{md}} \sum_{\boldsymbol{\psi} \in B_m^d} \mathbb{1}_{\{m \cdot \widehat{\text{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\psi}}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\psi}}) \le t\}} \\
&= \lim_{m\to\infty} \mathbb{E}\left(\mathbb{1}_{\{m \cdot \widehat{\text{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m}) \le t\}}\right) \\
&= \lim_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\text{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m}) \le t\right) \\
&= G(t).
\end{aligned}$$

Since $G$ is continuous Theorem B.1 implies that for all $t \in (0,1)$ that

$$\lim_{m\to\infty} \left(\widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)\right)^{-1}(t) = G^{-1}(t).$$

Recall that $\mathbb{P}(A_0) = 1$ which implies that it holds $\mathbb{P}$-a.s. that

$$\lim_{m\to\infty} \left(\widehat{R}_m(\mathbf{X}_1, \dots, \mathbf{X}_m)\right)^{-1} (1-\alpha) = G^{-1}(1-\alpha). \tag{4.13}$$

Finally, we can perform the following calculation

$$\begin{aligned}
&\limsup_{m\to\infty} \mathbb{P}\left(\varphi_m(\mathbf{X}_1, \dots, \mathbf{X}_m) = 1\right)\\
&\quad = \limsup_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m) > \left(\widehat{R}_m(\mathbf{X}_1, \dots, \mathbf{X}_m)\right)^{-1} (1-\alpha)\right)\\
&\quad = 1 - \liminf_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1, \dots, \mathbf{X}_m) \leq \left(\widehat{R}_m(\mathbf{X}_1, \dots, \mathbf{X}_m)\right)^{-1} (1-\alpha)\right)\\
&\quad = 1 - G(G^{-1}(1-\alpha)) = \alpha,
\end{aligned}$$

where the last step uses Corollary B.3 together with (4.10) and (4.13). This completes the proof of Proposition 4.9. □

The following proposition shows that the bootstrap test is also pointwise consistent against any alternative. The proof is very similar to the proof of Proposition 4.9 the main difference being that in Proposition 4.9 we consider the case $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$ while in Proposition 4.10 we consider $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_A$.

**Proposition 4.10 (consistency of the bootstrap test for dHSIC)**
*Assume Setting 3.1 and let $H_0$ and $H_A$ be defined as in* (4.1) *and* (4.2)*. Then for all $\alpha \in (0,1)$ the $\alpha$-bootstrap test is pointwise consistent when testing $H_0$ against $H_A$.*

**Proof** Let $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_A$ be fixed and begin by introducing the following notation:

· Let $(Z_j)_{j\in\mathbb{N}}$ be a sequence of independent standard normal random variables on $\mathbb{R}$, let $T_{h_2} \in L(L^2(\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d}, |\cdot|_{\mathbb{R}}))$ with the property that for every $f \in L^2(\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d}, |\cdot|_{\mathbb{R}})$ and for every $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ it holds that

$$\left(T_{h_2}(f)\right)(\mathbf{x}) = \int_{\boldsymbol{\mathcal{X}}} h_2(\mathbf{x}, \mathbf{y}) f(\mathbf{y})\, \mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d}(\mathrm{d}\mathbf{y}) \tag{4.14}$$

and let $(\lambda_j)_{j\in\mathbb{N}}$ be the eigenvalues of $T_{h_2}$.

· Let $\widehat{F}_m$ be the empirical product distribution function and define for all $\mathbf{t} \in \mathbb{R}^d$ the population product distribution function by

$$F(\mathbf{t}) := \left(\mathbb{P}^{X^1} \otimes \cdots \otimes \mathbb{P}^{X^d}\right) \left((-\infty, t^1] \times \cdots \times (-\infty, t^d]\right) = \prod_{j=1}^{d} \mathbb{P}\left(X^j \leq t^j\right)$$

Applying the Glivenko-Cantelli theorem (e.g. van der Vaart, 1998, Theorem 19.1), which extends the strong law of large numbers for empirical distributions to uniform convergence, shows that there exists a subset $A_0 \subseteq \Omega$ such that $\mathbb{P}(A_0) = 1$ and such that for all $\omega \in A_0$ it holds for all $\mathbf{t} \in \mathbb{R}^d$ that

$$\lim_{m\to\infty} \widehat{F}_m(\mathbf{X}_1(\omega), \dots, \mathbf{X}_m(\omega))(\mathbf{t}) = F(\mathbf{t}). \tag{4.15}$$

Next, for all $m \in \mathbb{N}$ let $\boldsymbol{\Psi}_m$ be a uniformly distributed random variable on $B_m^d$ independent of $\mathbf{X}$ and $(\mathbf{X}_i)_{i\in\mathbb{N}}$. Then, by Proposition 4.8 it holds for all $m \in \mathbb{N}$ and for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that

$$g_{m,\boldsymbol{\Psi}_m}(\mathbf{x}_1, \dots, \mathbf{x}_m) = \left(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m}\right)$$

are iid random variables with distribution function $\widehat{F}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)$.

Fix $\omega \in A_0$, let $(\mathbf{X}_i^*)_{i\in\mathbb{N}}$ be iid sequence of random variables with distribution $\mathbb{P}^{X^1} \otimes \dots \otimes \mathbb{P}^{X^d}$ and for all $i \in \mathbb{N}$ define $\mathbf{x}_i \coloneqq \mathbf{X}_i(\omega)$. Then, by (4.15) it holds that

$$\mathbf{x}_{m,i}^{\boldsymbol{\Psi}_m} \xrightarrow{d} \mathbf{X}_i^*$$

as $m \to \infty$. Hence, we are in the same setting as described in Setting 2.44.

Since both $\mathbb{P}^{X^1} \otimes \dots \otimes \mathbb{P}^{X^d} \in H_0$ and $\widehat{\mathbb{P}}_m^{X^1} \otimes \dots \otimes \widehat{\mathbb{P}}_m^{X^d} \in H_0$ it holds by Lemma C.3 for all $\mathbf{z} \in \boldsymbol{\mathcal{X}}$ that

$$h_1(\mathbf{z}) = \mathbb{E}\left(h(\mathbf{z}, \mathbf{X}_2^*, \dots, \mathbf{X}_{2d}^*)\right) = 0$$

and for all $m \in \{2d, 2d+1, \dots\}$ and for all $\mathbf{z} \in \boldsymbol{\mathcal{X}}$ that

$$h_1^m(\mathbf{z}) = \mathbb{E}\left(h(\mathbf{z}, \mathbf{x}_{m,2}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,2d}^{\boldsymbol{\Psi}_m})\right) = 0,$$

where $h$ is defined as in (3.6). Moreover, it holds by Theorem 3.3 that

$$\theta_h = \mathbb{E}\left(h(\mathbf{X}_1^*, \dots, \mathbf{X}_{2d}^*)\right) = \mathrm{dHSIC}\left(\mathbb{P}^{X^1} \otimes \dots \otimes \mathbb{P}^{X^d}\right) = 0.$$

We therefore satisfy all requirements of Theorem 2.48 (the condition $\xi_2(h) > 0$ is assumed, see remark after Theorem 4.3) and get that

$$m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m}) = m\tilde{V}_m(h) \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2 \tag{4.16}$$

as $m \to \infty$, where the $\lambda_i$'s are defined as the eigenvalues of the operator in (4.14).

Let $G : \mathbb{R} \to (0,1)$ be the distribution function of $\binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2$, then by (4.16) it holds for all $t \in \mathbb{R}$ that

$$\begin{aligned}
\lim_{m\to\infty} \widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)(t) &= \lim_{m\to\infty} \frac{1}{m^{md}} \sum_{\boldsymbol{\psi}\in B_m^d} \mathbb{1}_{\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\psi}},\dots,\mathbf{x}_{m,m}^{\boldsymbol{\psi}})\le t\}} \\
&= \lim_{m\to\infty} \mathbb{E}\left(\mathbb{1}_{\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m},\dots,\mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m})\le t\}}\right) \\
&= \lim_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{x}_{m,1}^{\boldsymbol{\Psi}_m}, \dots, \mathbf{x}_{m,m}^{\boldsymbol{\Psi}_m}) \le t\right) \\
&= G(t).
\end{aligned}$$

Since $G$ is continuous Theorem B.1 implies for all $t \in (0,1)$ that

$$\lim_{m\to\infty} \left(\widehat{R}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)\right)^{-1}(t) = G^{-1}(t).$$

Therefore we have shown that for all $\omega \in A_0$ it holds that

$$\lim_{m\to\infty} \left(\widehat{R}_m(\mathbf{X}_1(\omega),\dots,\mathbf{X}_m(\omega))\right)^{-1}(1-\alpha) = G^{-1}(1-\alpha). \tag{4.17}$$

Introduce the set

$$A_1 := \left\{\omega \in \Omega \,\middle|\, \forall t \in \mathbb{R} : \lim_{m\to\infty} \mathbb{1}_{\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1(\omega),\dots,\mathbf{X}_m(\omega))\leq t\}} = 0\right\}.$$

By Corollary 4.2 it holds that $\mathbb{P}(A_1) = 1$, which implies that $\mathbb{P}(A_0 \cap A_1) = 1$. Let $\omega \in A_0 \cap A_1$, then by (4.17) there exists a constant $t^* \in \mathbb{R}$ such that for all $m \in \mathbb{N}$ it holds that

$$\left(\widehat{R}_m(\mathbf{X}_1(\omega),\dots,\mathbf{X}_m(\omega))\right)^{-1}(1-\alpha) \leq t^*$$

and hence

$$\begin{aligned}&\lim_{m\to\infty} \mathbb{1}_{\left\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1(\omega),\dots,\mathbf{X}_m(\omega))\leq\left(\widehat{R}_m(\mathbf{X}_1(\omega),\dots,\mathbf{X}_m(\omega))\right)^{-1}(1-\alpha)\right\}}\\ &\qquad\leq \lim_{m\to\infty} \mathbb{1}_{\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1(\omega),\dots,\mathbf{X}_m(\omega))\leq t^*\}} = 0.\end{aligned}$$

This proves that $\mathbb{P}$-a.s. it holds that

$$\lim_{m\to\infty} \mathbb{1}_{\left\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m)\leq\left(\widehat{R}_m(\mathbf{X}_1,\dots,\mathbf{X}_m)\right)^{-1}(1-\alpha)\right\}} = 0$$

and applying the dominated convergence theorem we also get

$$\begin{aligned}&\lim_{m\to\infty} \mathbb{P}\left(\varphi_m(\mathbf{X}_1,\dots,\mathbf{X}_m) = 0\right)\\ &\qquad= \lim_{m\to\infty} \mathbb{E}\left(\mathbb{1}_{\left\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m)\leq\left(\widehat{R}_m(\mathbf{X}_1,\dots,\mathbf{X}_m)\right)^{-1}(1-\alpha)\right\}}\right)\\ &\qquad= 0,\end{aligned}$$

which completes the proof of Proposition 4.10. □

Similar to the case of the permutation test the size of the set $(B_m)^d$ is given by $m^{md}$ which grows very fast. For implementation purposes we therefore generally use a Monte-Carlo approximated version as defined in Definition 2.62. Using Proposition 2.61 and remarks at the end of Section 2.4.2, it follows that the asymptotic properties of the boostrap based test are preserved in the Monte-Carlo approximations. For more details see Section 4.5.

## 4.3 Gamma approximation

We showed in Theorem 4.3 that the asymptotic distribution of $m \cdot \widehat{\mathrm{dHSIC}}_m$ is given by

$$\binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2.$$

The essential idea behind the gamma approximation is that a distribution of the form

$$\sum_{i=1}^{\infty} \lambda_i Z_i^2$$

can be approximated fairly well by a gamma distribution with matched first and second moments (see Satterthwaite, 1946, for basic empirical evidence). This has, however, only been shown empirically and there are no guarantees that it leads to good results in the large sample limit. Nevertheless, since it is very fast and in most settings leads to good results it turns out to be very useful.

The gamma distribution with parameters $\alpha$ and $\beta$ is denoted by $\mathrm{Gamma}(\alpha, \beta)$ and corresponds to the distribution with density

$$f(x) = \frac{x^{\alpha-1} e^{\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)}$$

where $\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$ is the gamma function. The first two moments of the $\mathrm{Gamma}(\alpha, \beta)$-distributed random variable $Y$ are given by $\mathbb{E}(Y) = \alpha\beta$ and $\mathrm{Var}(Y) = \alpha\beta^2$. In order to match the first two moments we define for $\mathbf{X}_1, \mathbf{X}_2, \ldots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$ the two parameters

$$\alpha_m(\mathbb{P}^{\mathbf{X}}) := \frac{\left(\mathbb{E}\left(\widehat{\mathrm{dHSIC}}_m\right)\right)^2}{\mathrm{Var}\left(\widehat{\mathrm{dHSIC}}_m\right)} \quad \text{and} \quad \beta_m(\mathbb{P}^{\mathbf{X}}) := \frac{m \,\mathrm{Var}\left(\widehat{\mathrm{dHSIC}}_m\right)}{\mathbb{E}\left(\widehat{\mathrm{dHSIC}}_m\right)}. \tag{4.18}$$

Then we make the approximation

$$m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1, \ldots, \mathbf{X}_m) \sim \mathrm{Gamma}\left(\alpha_m(\mathbb{P}^{\mathbf{X}}), \beta_m(\mathbb{P}^{\mathbf{X}})\right). \tag{4.19}$$

In order to use this approximation in a hypothesis test we first need to find a method to calculate $\alpha_m(\mathbb{P}^{\mathbf{X}})$ and $\beta_m(\mathbb{P}^{\mathbf{X}})$ based only on the first $m$ observations $\mathbf{X}_1, \ldots, \mathbf{X}_m$. The following two theorems give expansions of the moments in terms of the kernel.

**Lemma 4.11 (mean of $\widehat{\mathbf{dHSIC}}$)**
*Assume Setting 3.1. Then under $H_0$ it holds that,*

$$\mathbb{E}\left(\widehat{\mathrm{dHSIC}}_m\right) = \frac{1}{m} - \frac{1}{m} \sum_{r=1}^{d} \left( \prod_{j \neq r} \mathbb{E}\left(k^j(X_1^j, X_2^j)\right) \right) + \frac{d-1}{m} \prod_{j=1}^{d} \mathbb{E}\left(k^j(X_1^j, X_2^j)\right) + \mathcal{O}\left(m^{-2}\right)$$

*as $m \to \infty$.*

**Proof** Due to Lemma 3.6 we know that $\widehat{\text{dHSIC}}_m$ is a V-statistic with core function $h$. Under $H_0$ it holds that $\theta_h = 0$ and thus applying Lemma 2.40 results in

$$\mathbb{E}\left(\widehat{\text{dHSIC}}_m\right) = \frac{1}{m}\binom{2d}{2}\mathbb{E}\left(h_2(\mathbf{X}_1, \mathbf{X}_1)\right) + \mathcal{O}\left(m^{-2}\right).$$

We can use Lemma C.2 to explicitly calculate $\binom{2p}{2}\mathbb{E}(h_2(\mathbf{X}_1, \mathbf{X}_1))$, which together with the independence assumption under $H_0$ simplifies to the desired expression. This concludes the proof of Lemma 4.11. □

**Lemma 4.12 (variance of $\widehat{\textbf{dHSIC}}$)**
*Assume Setting 3.1. Then under $H_0$ it holds that,*

$$\begin{aligned}\operatorname{Var}\left(\widehat{\text{dHSIC}}_m\right) = 2\frac{(m-2d)!}{m!}\frac{(m-2d)!}{(m-4d+2)!}\Bigg[&\prod_{j=1}^{d} e_1(j) + (d-1)^2\prod_{j=1}^{d} e_0(j)^2 \\ &+ 2(d-1)\prod_{j=1}^{d} e_2(j) + \sum_{j=1}^{d} e_1(j)\prod_{r\neq j} e_0(r)^2 \\ &- 2\sum_{j=1}^{d} e_1(j)\prod_{r\neq j} e_2(r) - 2(d-1)\sum_{j=1}^{d} e_2(j)\prod_{r\neq j} e_0(r)^2 \\ &+ \sum_{j\neq l} e_2(j)e_2(l)\prod_{r\neq j,l} e_0(r)^2\Bigg] + \mathcal{O}\left(m^{-\frac{5}{2}}\right)\end{aligned}$$

*as $m \to \infty$ and where for all $j \in \{1, \dots, d\}$ we set*

$$e_0(j) = \mathbb{E}\left(k^j(X_1^j, X_2^j)\right), \quad e_1(j) = \mathbb{E}\left(k^j(X_1^j, X_2^j)^2\right), \quad e_2(j) = \mathbb{E}_{X_1^j}\left(\mathbb{E}_{X_2^j}\left(k^j(X_1^j, X_2^j)\right)^2\right).$$

**Proof** Due to Lemma 3.6 we know that $\widehat{\text{dHSIC}}$ is a V-statistic with core function $h$. Applying Lemma 2.39 thus results in

$$\operatorname{Var}\left(\widehat{\text{dHSIC}}_m\right) = \binom{m}{2d}^{-1}\binom{2d}{2}\binom{m-2d}{2d-2}\xi_2 + \mathcal{O}\left(m^{-\frac{5}{2}}\right).$$

Using Lemma C.2 we get that

$$\begin{aligned}\xi_2 &= \mathbb{E}\left(h_2(\mathbf{X}_1, \mathbf{X}_2)^2\right) \\ &= \binom{2d}{2}^{-2}\mathbb{E}\left(\left(\sum_{i=1}^{10} a_i\right)^2\right) \\ &= \binom{2d}{2}^{-2}\sum_{i,j=1}^{10}\mathbb{E}\left(a_i a_j\right).\end{aligned}$$

Each term $\mathbb{E}(a_i a_j)$ can be explicitly calculated and simplified using the independence properties under $H_0$ (very tedious). This concludes the proof of Lemma 4.12. □

Based on these two lemmas we only need a method to calculate the terms

(i) $e_0(j) := \mathbb{E}\left(k^j(X_1^j, X_2^j)\right)$,

(ii) $e_1(j) := \mathbb{E}\left(k^j(X_1^j, X_2^j)^2\right)$,

(iii) $e_2(j) := \mathbb{E}_{X_1^j}\left(\mathbb{E}_{X_2^j}\left(k^j(X_1^j, X_2^j)\right)^2\right)$.

One obvious choice would be to use a U-statistic for each expectation term as this would not add any bias. It however turns out that a V-statistic also does not add any bias in this particular case. This is due to Theorem 2.40, which shows that the bias of a V-statistic is of order $\mathcal{O}\left(m^{-1}\right)$ and hence is consumed by the error terms in Lemma 4.11 and Lemma 4.12. The V-statistics for these terms are given for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by

(i) $\widehat{e}_0(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) := \frac{1}{m^2}\sum_{i_1,i_2=1}^m k^j(x_{i_1}^j, x_{i_2}^j)$,

(ii) $\widehat{e}_1(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) := \frac{1}{m^2}\sum_{i_1,i_2=1}^m k^j(x_{i_1}^j, x_{i_2}^j)^2$,

(iii) $\widehat{e}_2(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) := \frac{1}{m^3}\sum_{i_2=1}^m \left(\sum_{i_1=1}^m k^j(x_{i_1}^j, x_{i_2}^j)\right)^2$.

Based on these terms we can define the estimator $\widehat{\mathrm{Exp}}_m : \boldsymbol{\mathcal{X}}^m \to \mathbb{R}$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by

$$\widehat{\mathrm{Exp}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) := \frac{1}{m} - \frac{1}{m}\sum_{r=1}^d \prod_{j\neq r} \widehat{e}_0(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) + \frac{d-1}{m}\prod_{j=1}^d \widehat{e}_0(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) \tag{4.20}$$

and the estimator $\widehat{\mathrm{Var}}_m : \boldsymbol{\mathcal{X}}^m \to \mathbb{R}$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by

$$\begin{aligned}\widehat{\mathrm{Var}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \coloneqq 2\frac{(m-2d)!}{m!}\frac{(m-2d)!}{(m-4d+2)!}\Bigg[&\prod_{j=1}^{d} \widehat{e}_1(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) \\ &+ (d-1)^2 \prod_{j=1}^{d} \widehat{e}_0(j)(\mathbf{x}_1, \dots, \mathbf{x}_m)^2 \\ &+ 2(d-1) \prod_{j=1}^{d} \widehat{e}_2(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) \\ &+ \sum_{j=1}^{d} \widehat{e}_1(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) \prod_{r\neq j} \widehat{e}_0(r)(\mathbf{x}_1, \dots, \mathbf{x}_m)^2 \\ &- 2\sum_{j=1}^{d} \widehat{e}_1(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) \prod_{r\neq j} \widehat{e}_2(r)(\mathbf{x}_1, \dots, \mathbf{x}_m) \\ &- 2(d-1)\sum_{j=1}^{d} \widehat{e}_2(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) \prod_{r\neq j} \widehat{e}_0(r)(\mathbf{x}_1, \dots, \mathbf{x}_m)^2 \\ &+ \sum_{j\neq l} \widehat{e}_2(j)(\mathbf{x}_1, \dots, \mathbf{x}_m)\widehat{e}_2(l)(\mathbf{x}_1, \dots, \mathbf{x}_m) \prod_{r\neq j,l} \widehat{e}_0(r)(\mathbf{x}_1, \dots, \mathbf{x}_m)^2\Bigg].\end{aligned} \tag{4.21}$$

Finally we can define the estimator $\widehat{\alpha}_m : \boldsymbol{\mathcal{X}}^m \to \mathbb{R}$ of $\alpha_m$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by

$$\widehat{\alpha}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \coloneqq \frac{\widehat{\mathrm{Exp}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)^2}{\widehat{\mathrm{Var}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)} \tag{4.22}$$

and the estimator $\widehat{\beta}_m : \boldsymbol{\mathcal{X}}^m \to \mathbb{R}$ of $\beta_m$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by

$$\widehat{\beta}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \coloneqq \frac{m\widehat{\mathrm{Var}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)}{\widehat{\mathrm{Exp}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)}. \tag{4.23}$$

Using these estimators we can define the following hypothesis test.

**Definition 4.13 (gamma approximation based test for dHSIC)**
*Assume Setting 3.1, assume $\alpha \in (0,1)$ and for all $m \in \mathbb{N}$ let $F_m : \boldsymbol{\mathcal{X}}^m \times \mathbb{R} \to [0,1]$ be the functions satisfying for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that $F_m(\mathbf{x}_1, \dots, \mathbf{x}_m)$ is the distribution function associated to the* Gamma$(\widehat{\alpha}_m(\mathbf{x}_1, \dots, \mathbf{x}_m), \widehat{\beta}_m(\mathbf{x}_1, \dots, \mathbf{x}_m))$*-distribution, where $\widehat{\alpha}_m$ and $\widehat{\beta}_m$ are defined as in (4.22), (4.23) respectively. Then the hypothesis test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ satisfying for all $m \in \{1, \dots, 2d-1\}$ that $\varphi_m \coloneqq 0$ and for all $m \in \{2d, 2d+1, \dots\}$ and for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that*

$$\varphi_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \coloneqq \mathbb{1}_{\left\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{x}_1,\dots,\mathbf{x}_m) > F_m(\mathbf{x}_1,\dots,\mathbf{x}_m)^{-1}(1-\alpha)\right\}}$$

*is called gamma approximation based $\alpha$-test for dHSIC.*

To illustrate that the gamma approximation based $\alpha$-test for dHSIC $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ is a good choice assume the approximation (4.19) is exact with $\widehat{\alpha}_m$ replacing $\alpha_m$ and with $\widehat{\beta}_m$ replacing $\beta_m$ (which is not true), then it would hold that

$$\begin{aligned}
\mathrm{E}_1(\varphi_m) &= \sup_{\mathbf{X}\sim\mathbb{P}^{\mathbf{X}}\in H_0} \mathbb{P}\left(\varphi_m(\mathbf{X}_1,\dots,\mathbf{X}_m)=1\right)\\
&= \sup_{\mathbf{X}\sim\mathbb{P}^{\mathbf{X}}\in H_0} \mathbb{P}\left(m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) > F_m^{-1}(\mathbf{X}_1,\dots,\mathbf{X}_m)(1-\alpha)\right)\\
&= 1-\inf_{\mathbf{X}\sim\mathbb{P}^{\mathbf{X}}\in H_0} \mathbb{P}\left(m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) \leq F_m^{-1}(\mathbf{X}_1,\dots,\mathbf{X}_m)(1-\alpha)\right)\\
&= 1-(1-\alpha)=\alpha
\end{aligned}$$

Of course, since (4.19) is only a heuristic we have no guarantee that the gamma approximation test actually has exact level.

## 4.4 Eigenvalue approach

In this section we show a method that aims at approximating the null distribution

$$\binom{2d}{2}\sum_{i=1}^{\infty}\lambda_i Z_i^2$$

by estimating the eigenvalues $\lambda_i$ based on the data. For this to be a sensible approach we have to find a way to estimate the eigenvalues of integral operators in such a way that the distribution calculated from the estimated eigenvalues converges to the desired distribution in the large sample limit.

### 4.4.1 General setting

We begin by formalizing the required integral operators.

**Definition 4.14 ($T_k$-operator)**
*Let $\mathcal{X}$ be a metric space and $k$ a continuous, bounded, positive definite kernel on $\mathcal{X}$. Then for every $\mu\in\mathcal{P}(\mathcal{X})$ let $T_k(\mu): L^2(\mu,|\cdot|_{\mathbb{R}})\to L^2(\mu,|\cdot|_{\mathbb{R}})$ be the functions with the property that for every $f\in L^2(\mu,|\cdot|_{\mathbb{R}})$ it holds that*

$$(T_k(\mu))(f)=\int_{\mathcal{X}} k(x,\cdot)f(x)\,\mu(dx).$$

The Bochner integral is well-defined, since the function $x\mapsto k(x,\cdot)f(x)$ is in $\mathcal{L}^1(\mu,\|\cdot\|_{\mathcal{H}})$.

The following theorem connects the eigenvalues of the integral operator $T_k(\mu)$ with the eigenvalues of the covariance operator. It is due to Blanchard et al. (2007, Theorem 4.1).

**Theorem 4.15 (eigenvalue correspondence)**
*Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a metric space and $k$ a continuous, bounded, positive definite kernel on $\mathcal{X}$. Then, for every random variable $X : \Omega \to \mathcal{X}$ with law $\mathbb{P}^X$ it holds that*

$$\sigma_p\left(T_k(\mathbb{P}^X)\right) = \sigma_p\left(\mathrm{CovOp}(k(X, \cdot))\right),$$

*where the covariance operator is defined in Definition 2.5, and the eigenvalues have the same multiplicity.*

**Proof** Observe that by the reproducing property of $k$, the Cauchy-Schwarz inequality and the fact that $k$ is bounded, it holds for all $f \in \mathcal{H}$ that

$$\begin{aligned}\mathbb{E}\left(|f(X)|^2\right) &= \mathbb{E}\left(|\langle f, k(X, \cdot)\rangle_{\mathcal{H}}|^2\right)\\ &\leq \mathbb{E}\left(\|f\|_{\mathcal{H}}^2 \|k(X, \cdot)\|_{\mathcal{H}}^2\right)\\ &= \|f\|_{\mathcal{H}}^2 \mathbb{E}\left(k(X, X)\right)\\ &\leq C\|f\|_{\mathcal{H}}^2\end{aligned}$$

for some constant $C > 0$. Therefore we have shown that the function $A : \mathcal{H} \to L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})$ with the property that for all $f \in \mathcal{H}$ it holds that $Af = f$ is a well-defined, bounded, linear operator. Therefore its adjoint $A^* : L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}}) \to \mathcal{H}$ exists and satisfies for all $f \in L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})$ and all $g \in \mathcal{H}$ that

$$\langle A^* f, g\rangle_{\mathcal{H}} = \langle f, Ag\rangle_{L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})}. \tag{4.24}$$

Next, observe that by the reproducing property for all $f \in L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})$ and all $g \in \mathcal{H}$ it holds that

$$\begin{aligned}\langle f, Ag\rangle_{L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})} &= \mathbb{E}\left(f(X)g(X)\right)\\ &= \mathbb{E}\left(f(X)\langle g, k(X, \cdot)\rangle_{\mathcal{H}}\right)\\ &= \mathbb{E}\left(\langle g, f(X)k(X, \cdot)\rangle_{\mathcal{H}}\right)\\ &= \langle g, \mathbb{E}\left(f(X)k(X, \cdot)\right)\rangle_{\mathcal{H}},\end{aligned}$$

where in the last step we use a property of the Bochner integral and that $k$ is bounded and $f \in L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})$ ensuring the existence of the integral. Together with (4.24) this implies that

$$A^* f = \mathbb{E}\left(f(X)k(X, \cdot)\right). \tag{4.25}$$

This immediately implies that $AA^* = T_k(\mathbb{P}^X)$ and since for all $f, g \in \mathcal{H}$ it holds that

$$\langle f, A^*Ag\rangle_{\mathcal{H}} = \langle Af, Ag\rangle_{L^2(\mathbb{P}^X, |\cdot|_{\mathbb{R}})} = \mathbb{E}\left(f(X)g(X)\right) = \mathbb{E}\left(\langle f, k(X, \cdot)\rangle_{\mathcal{H}}\langle g, k(X, \cdot)\rangle_{\mathcal{H}}\right)$$

the uniqueness property of the covariance operator implies $A^*A = \mathrm{CovOp}(k(X, \cdot))$. Denote by

$$E_\lambda(B) = \{x \mid Bx = \lambda x\}$$

the eigenspace of an operator $B$ associated to the eigenvalue $\lambda$. Next assume $0 < \lambda \in \sigma_p(T_k(\mathbb{P}^X))$ and let $f$ be an associated eigenfunction. Then it holds that

$$\mathrm{CovOp}(k(X,\cdot))A^*f = A^*AA^*f = A^*T_k(\mathbb{P}^X) = \lambda A^*f.$$

This implies that

$$A^*E_\lambda(T_k(\mathbb{P}^X)) \subseteq E_\lambda(\mathrm{CovOp}(k(X,\cdot))). \tag{4.26}$$

In particular this means that $\lambda$ is an eigenvalue of $\mathrm{CovOp}(k(X,\cdot))$. Thus we can apply the same argument to $\mathrm{CovOp}(k(X,\cdot))$ which leads to

$$AE_\lambda\left(\mathrm{CovOp}(k(X,\cdot))\right) \subseteq E_\lambda\left(T_k(\mathbb{P}^X)\right). \tag{4.27}$$

Now, since $\lambda > 0$ applying $A$ to both sides of the inclusion in (4.26) results in

$$AA^*E_\lambda\left(T_k(\mathbb{P}^X)\right) = T_k(\mathbb{P}^X)E_\lambda\left(T_k(\mathbb{P}^X)\right) = E_\lambda\left(T_k(\mathbb{P}^X)\right) \subseteq AE_\lambda\left(\mathrm{CovOp}(k(X,\cdot))\right). \tag{4.28}$$

Finally, combining (4.27) and (4.28) gives us

$$AE_\lambda\left(\mathrm{CovOp}(k(X,\cdot))\right) = E_\lambda\left(T_k(\mathbb{P}^X)\right).$$

This implies that multiplicities coincide, which completes the proof of Theorem 4.15. □

Next, we define for all $z \in \mathcal{H}$ the function $\mathcal{C}_z \in L_1(\mathcal{H})$ by

$$\mathcal{C}_z = z \otimes z.$$

Furthermore, for an $\mathcal{H}$-valued random variable $Z$ and a sequence of independent copies $(Z_i)_{i\in\mathbb{N}}$ of $Z$ define

$$\begin{aligned} \mathcal{C} &:= \mathbb{E}\left(Z \otimes Z\right) = \mathbb{E}\left(\mathcal{C}_Z\right) \\ \widehat{\mathcal{C}}_m &:= \frac{1}{m}\sum_{i=1}^m Z_i \otimes Z_i = \frac{1}{m}\sum_{i=1}^m \mathcal{C}_{Z_i}. \end{aligned} \tag{4.29}$$

The next theorem shows that $\widehat{\mathcal{C}}_m$ is a consistent estimator of $\mathcal{C}$.

**Theorem 4.16 (consistency of the empirical covariance operator)**
*Let $\mathcal{H}$ be a separable Hilbert space, $Z$ an $\mathcal{H}$-valued random variable with finite second moment and $(Z_i)_{i\in\mathbb{N}}$ a sequence of independent copies of $Z$. Then it holds that*

$$\|\widehat{\mathcal{C}}_m - \mathcal{C}\|_1 \overset{\mathbb{P}-a.s.}{\longrightarrow} 0$$

*as $m \to \infty$, where $\widehat{\mathcal{C}}_m$ and $\mathcal{C}$ are defined as in* (4.29).

**Proof** For all $i \in \mathbb{N}$ set $Y_i = \mathcal{C}_{Z_i} - \mathbb{E}\left(\mathcal{C}_Z\right)$. Then we get

$$\begin{aligned} \widehat{\mathcal{C}}_m - \mathcal{C} &= \frac{1}{m}\sum_{i=1}^m \mathcal{C}_{Z_i} - \mathbb{E}\left(\mathcal{C}_Z\right) \\ &= \frac{1}{m}\sum_{i=1}^m Y_i. \end{aligned} \tag{4.30}$$

Next observe that the sequence $(Y_i)_{i\in\mathbb{N}}$ satisfies the following conditions

(i) $(Y_i)_{i\in\mathbb{N}}$ is a sequence of random variables with values in $L_1(\mathcal{H})$,

(ii) for all $i \in \mathbb{N}$ it holds that $\mathbb{E}(Y_i) = 0$, and

(iii) $(Y_i)_{i\in\mathbb{N}}$ are independent and identically distributed.

Since $\mathcal{H}$ is a separable Hilbert space it holds that $L_1(\mathcal{H})$ is a separable Banach space (see Theorem 2.2). Therefore we can apply the extension of the strong law of large numbers given in Theorem B.6 to get that

$$\left\|\frac{1}{m}\sum_{i=1}^{m} Y_i\right\|_1 \overset{\mathbb{P}-a.s.}{\longrightarrow} 0, \tag{4.31}$$

as $m \to \infty$. Hence combining (4.30) with (4.31) completes the proof of Theorem 4.16. □

### 4.4.2 Application to dHSIC

Throughout this section we assume Setting 3.1. Recall that by Theorem 4.3 the asymptotic distribution of $m \cdot \widehat{\text{dHSIC}}_m$ under $H_0$ is given by

$$m \cdot \widehat{\text{dHSIC}}_m \xrightarrow{d} \binom{2d}{2} \sum_{i=1}^{\infty} \lambda_i Z_i^2$$

where $(Z_i)_{i\in\mathbb{N}}$ is an iid sequence of standard normal random variables and $(\lambda_i)_{i\in\mathbb{N}}$ are the eigenvalues of the operator $T_{h_2}(\mathbb{P}^{\mathbf{X}})$.

Now consider the empirical distribution based on the observations $\mathbf{X}_1, \dots, \mathbf{X}_m$ which is given by

$$\widehat{\mathbb{P}}_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{\mathbf{X}_i}, \tag{4.32}$$

where $\delta_{\mathbf{x}}$ is the Dirac measure at the point $\mathbf{x}$. For $f \in \mathcal{H}$ it holds that

$$\begin{aligned}(T_{h_2}(\widehat{\mathbb{P}}_m))(f) &= \int_{\mathcal{X}} h_2(\mathbf{x}, \cdot) f(\mathbf{x})\, \widehat{\mathbb{P}}_m(\mathrm{d}\mathbf{x}) \\ &= \frac{1}{m}\sum_{i=1}^{m} h_2(\mathbf{X}_i, \cdot) f(\mathbf{X}_i).\end{aligned}$$

The idea is to calculate the eigenvalues $\{\nu_{m,1}, \dots, \nu_{m,m}\}$ of the Gram matrix $\mathbf{H_2}$ of the kernel $h_2$ given the observations $\mathbf{X}_1, \dots, \mathbf{X}_m$, then it follows that

$$\sigma_p\left(T_{h_2}(\widehat{\mathbb{P}}_m)\right) = \left\{0, \frac{\nu_{m,1}}{m}, \dots, \frac{\nu_{m,m}}{m}\right\}.$$

For the following analysis, we assume that $h_2$ is positive semi-definite. In the case $d = 2$ this can be shown relatively easy, while in the general case a proof of this is not so clear. Assume $\{\nu_{m,1}, \dots, \nu_{m,m}\}$ are sorted in descending order, then set

$$\widehat{\lambda}_{m,i} = \begin{cases} \frac{\nu_{m,i}}{m} & \text{for all } i \in \{1, \dots, m\} \\ 0 & \text{for all } i \in \{m+1, \dots\}. \end{cases} \tag{4.33}$$

The following theorem shows that using these empirical eigenvalues leads to a consistent estimate of the null distribution. The proof follows Gretton et al. (2009, Theorem 1).

**Theorem 4.17 (consistency of empirical eigenvalues)**
*Assume Setting 3.1. Let $(Z_i)_{i\in\mathbb{N}}$ be a sequence of iid standard normal random variables, let $(\lambda_i)_{i\in\mathbb{N}}$ and $(\widehat{\lambda}_{m,i})_{i\in\mathbb{N}}$ be the non-decreasing sequences of eigenvalues of the operators $T_{h_2}(\mathbb{P}^{\mathbf{X}})$ and $T_{h_2}(\widehat{\mathbb{P}}_m)$, respectively. Then it holds that*

$$\sum_{i=1}^{\infty}\widehat{\lambda}_{m,i}Z_i^2 \xrightarrow{d} \sum_{i=1}^{\infty}\lambda_i Z_i^2$$

*as $m\to\infty$.*

**Proof** Using Kolmogorov's extension theorem (e.g. Klenke, 2014, Theorem 14.36) there exists a probability space $(\Omega',\mathcal{F}',\mathbb{P}')$ and independent random variables $(Z_i)_{i\in\mathbb{N}}$, $(\mathbf{X}'_i)_{i\in\mathbb{N}}$ and $\mathbf{X}'$ satisfying that

(i) for all $i\in\mathbb{N}$ it holds that $Z_i:\Omega'\to\mathbb{R}$ is a standard normal random variable,

(ii) for all $i\in\mathbb{N}$ it holds that $\mathbf{X}'_i:\Omega'\to\boldsymbol{\mathcal{X}}$ has the same distribution as $\mathbf{X}$ and

(iii) $\mathbf{X}':\Omega'\to\boldsymbol{\mathcal{X}}$ has the distribution as $\mathbf{X}$.

This immediately implies that $\mathbb{P}^{\mathbf{X}}=\mathbb{P}'^{\mathbf{X}'}$ and therefore the eigenvalues of $T_{h_2}(\mathbb{P}'^{\mathbf{X}'})$ are also $(\lambda_i)_{i\in\mathbb{N}}$. Similarly, if we denote by $(\widehat{\lambda}'_{m,i})_{i\in\mathbb{N}}$ the eigenvalues of $T_{h_2}(\widehat{\mathbb{P}'}_m)$, where $\widehat{\mathbb{P}'}_m$ is the empirical distribution corresponding to $\mathbf{X}'_1,\dots,\mathbf{X}'_m$, then $(\widehat{\lambda}'_{m,i})_{i\in\mathbb{N}}$ have the same distribution as $(\widehat{\lambda}_{m,i})_{i\in\mathbb{N}}$. It is therefore sufficient to show the result for $(\widehat{\lambda}'_{m,i})_{i\in\mathbb{N}}$.

Next, it holds that

$$\operatorname{CovOp}(h_2(\mathbf{X}',\cdot))=\mathbb{E}\left(h_2(\mathbf{X}',\cdot)\otimes h_2(\mathbf{X}',\cdot)\right)=:\mathcal{C}$$

and for a random variable $\widehat{\mathbf{X}'}$ with distribution $\widehat{\mathbb{P}'}_m$ it holds that

$$\operatorname{CovOp}(h_2(\widehat{\mathbf{X}'},\cdot))=\mathbb{E}\left(h_2(\widehat{\mathbf{X}'},\cdot)\otimes h_2(\widehat{\mathbf{X}'},\cdot)\right)=\frac{1}{m}\sum_{i=1}^{m}h_2(\mathbf{X}'_i,\cdot)\otimes h_2(\mathbf{X}'_i,\cdot)=:\widehat{\mathcal{C}}_m,$$

where $\widehat{\mathcal{C}}_m$ is defined as in (4.29). Then, applying Theorem 4.15 proves that $(\lambda_i)_{i\in\mathbb{N}}$ and $(\widehat{\lambda}'_{m,i})_{i\in\mathbb{N}}$ are also the eigenvalues of $\mathcal{C}$ and $\widehat{\mathcal{C}}_m$ respectively. By Lemma 2.7 it also holds that both $\mathcal{C}$ and $\widehat{\mathcal{C}}_m$ are non-negative nuclear operators and therefore it in particular holds that

$$\|\mathcal{C}\|_1=\operatorname{trace}(\mathcal{C})=\sum_{i=1}^{\infty}\lambda_i<\infty \tag{4.34}$$

and $\mathbb{P}'$-a.s. that

$$\|\widehat{\mathcal{C}}_m\|_1=\operatorname{trace}(\widehat{\mathcal{C}}_m)=\sum_{i=1}^{\infty}\widehat{\lambda}'_{m,i}<\infty. \tag{4.35}$$

So by (4.34) and since $Z_i$ have finite moments it holds that

$$\sum_{i=1}^{\infty} \mathbb{E}\left(\lambda_i Z_i^2\right) = \sum_{i=1}^{\infty} \lambda_i < \infty \quad \text{and} \quad \sum_{i=1}^{\infty} \operatorname{Var}\left(\lambda_i Z_i^2\right) = \sum_{i=1}^{\infty} \lambda_i^2 < \infty.$$

Hence, we can apply Kolmogorov's three series theorem (e.g. Klenke, 2014, Theorem 15.50) to get that

$$\sum_{i=1}^{\infty} \lambda_i Z_i^2 < \infty \quad \mathbb{P}'\text{-a.s.}$$

which together with (4.35) implies

$$\sum_{i=1}^{\infty} \widehat{\lambda}'_{m,i} Z_i^2 - \sum_{i=1}^{\infty} \lambda_i Z_i^2 = \sum_{i=1}^{\infty} \left(\widehat{\lambda}'_{m,i} - \lambda_i\right) Z_i^2 \quad \mathbb{P}'\text{-a.s..}$$

Therefore it only remains to prove that

$$\sum_{i=1}^{\infty} \left(\widehat{\lambda}'_{m,i} - \lambda_i\right) Z_i^2 \xrightarrow{\mathbb{P}'\text{-a.s.}} 0 \tag{4.36}$$

as $m \to \infty$. To this end observe that $\mathbb{P}'$-a.s. it holds that

$$\begin{aligned}
&\left|\sum_{i=1}^{\infty} \left(\widehat{\lambda}'_{m,i} - \lambda_i\right) Z_i^2\right| \\
&\quad \leq \left|\sum_{i=1}^{\infty} \left((\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}\right) (\widehat{\lambda}'_{m,i})^{\frac{1}{2}} Z_i^2\right| + \left|\sum_{i=1}^{\infty} \left((\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}\right) \lambda_i^{\frac{1}{2}} Z_i^2\right| \\
&\quad \leq \left[\sum_{i=1}^{\infty} \widehat{\lambda}'_{m,i} Z_i^4\right]^{\frac{1}{2}} \left[\sum_{i=1}^{\infty} \left|(\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}\right|^2\right]^{\frac{1}{2}} + \left[\sum_{i=1}^{\infty} \lambda_i Z_i^4\right]^{\frac{1}{2}} \left[\sum_{i=1}^{\infty} \left|(\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}\right|^2\right]^{\frac{1}{2}}
\end{aligned} \tag{4.37}$$

where in the last step we used Hölder's inequality. Next, we show that this expression converges to 0 $\mathbb{P}'$-a.s.. By Markov's inequality we have that

$$\mathbb{P}'\left(\left|\sum_{i=1}^{\infty} \lambda_i Z_i^4\right| > k\right) \leq \frac{1}{k} \mathbb{E}\left(\left|\sum_{i=1}^{\infty} \lambda_i Z_i^4\right|\right) \leq \frac{C_1}{k} \tag{4.38}$$

and by using independence of $\widehat{\lambda}'_{m,i}$ and $Z_i$ and again Markov's inequality we have that

$$\mathbb{P}'\left(\left|\sum_{i=1}^{\infty} \widehat{\lambda}'_{m,i} Z_i^4\right| > k\right) \leq \frac{1}{k} \mathbb{E}\left(\left|\sum_{i=1}^{\infty} \widehat{\lambda}'_{m,i} Z_i^4\right|\right) \leq \frac{1}{k} \sum_{i=1}^{\infty} \mathbb{E}\left(|\widehat{\lambda}'_{m,i}|\right) \mathbb{E}\left(|Z_i|^4\right) \leq \frac{C_2}{k} \tag{4.39}$$

where $C_1, C_2 > 0$ are two constants. Furthermore, it holds $\mathbb{P}'$-a.s. that

$$\sum_{i=1}^{\infty} \left((\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}\right)^2 \leq \sum_{i=1}^{\infty} |(\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}||(\widehat{\lambda}'_{m,i})^{\frac{1}{2}} + \lambda_i^{\frac{1}{2}}| = \sum_{i=1}^{\infty} |\widehat{\lambda}'_{m,i} - \lambda_i|. \tag{4.40}$$

Now, since both $\mathcal{C}$ and $\widehat{\mathcal{C}}_m$ are nuclear operators it follows that $(\widehat{\lambda}'_{m,i} - \lambda_i)_{i\in\mathbb{N}}$ are the eigenvalues of the operator $\widehat{\mathcal{C}}_m - \mathcal{C}$, which together with (4.40) and Theorem 4.16 implies that

$$\sum_{i=1}^{\infty}\left((\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}\right)^2 \leq \sum_{i=1}^{\infty}|\widehat{\lambda}'_{m,i} - \lambda_i| = \|\widehat{\mathcal{C}}_m - \mathcal{C}\|_1 \overset{\mathbb{P}'\text{-a.s.}}{\longrightarrow} 0 \tag{4.41}$$

as $m \to \infty$. Finally, combining (4.37) with (4.38), (4.39) and (4.41) shows that for all $k \in \mathbb{N}$ we have

$$\left|\sum_{i=1}^{\infty}\left(\widehat{\lambda}'_{m,i} - \lambda_i\right) Z_i^2\right| \leq 2\sqrt{k}\left[\sum_{i=1}^{\infty}\left|(\widehat{\lambda}'_{m,i})^{\frac{1}{2}} - \lambda_i^{\frac{1}{2}}\right|^2\right]^{\frac{1}{2}} \overset{\mathbb{P}'\text{-a.s.}}{\longrightarrow} 0$$

as $m \to \infty$ with probability at least $(1 - \frac{1}{k}\max\{C_1, C_2\})$. Therefore letting $k$ tend to infinity completes the proof of Theorem 4.17. □

### 4.4.3 Constructing a hypothesis test

In order to construct a hypothesis we need to calculate the values $\widehat{\lambda}_{m,i}$ from the observations. Using (4.33) we get that

$$\widehat{\lambda}_{m,i} = \frac{\nu_{m,i}}{m}\mathbb{1}_{\{i\in\{1,\dots,m\}\}},$$

where $\nu_{m,1}, \dots, \nu_{m,m}$ are the $m$ eigenvalues of the Gram matrix $\mathbf{H_2}$ of the kernel $h_2$ given observations $\mathbf{x}_1, \dots, \mathbf{x}_m$. Unfortunately, the explicit expansion of $h_2$ under $H_0$ given in Lemma C.2 contains the terms

(i) $e_0(j) := \mathbb{E}\left(k^j(X_1^j, X_2^j)\right)$,

(ii) $e_1(j)(\cdot) := \mathbb{E}\left(k^j(\cdot, X_1^j)\right)$,

where $j \in \{1, \dots, d\}$, which cannot be calculated from the observational data alone.

We hence need an extra approximation step. The U-statistic of (i) and (ii) are given for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by

(i) $\widehat{e}_0(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) := \frac{1}{m(m-1)}\sum_{i_1\neq i_2}^m k^j(x_{i_1}^j, x_{i_2}^j)$,

(ii) $\widehat{e}_1(j)(\mathbf{x}_1, \dots, \mathbf{x}_m)(\cdot) := \frac{1}{m}\sum_{i=1}^m k^j(\cdot, x_i^j)$.

Based on these estimators we can define the estimator $\widehat{\mathbf{H}}_2 : \boldsymbol{\mathcal{X}}^m \to \mathbb{R}^{m\times m}$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by replacing

$$e_0(j) \text{ by } \widehat{e}_0(j)(\mathbf{x}_1, \dots, \mathbf{x}_m) \text{ and } e_1(j)(\cdot) \text{ by } \widehat{e}_1(j)(\mathbf{x}_1, \dots, \mathbf{x}_m)(\cdot)$$

in the expansion of $\mathbf{H}_2$ given in Lemma C.2. Moreover, for all $i \in \{1, \dots, m\}$ we define an estimator $\widehat{\nu}_{m,i} : \boldsymbol{\mathcal{X}}^m \to \mathbb{R}$ of $\nu_{m,i}$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ by letting $\widehat{\nu}_{m,i}(\mathbf{x}_1, \dots, \mathbf{x}_m)$ be the i-th largest eigenvalue of $\widehat{\mathbf{H}}_2(\mathbf{x}_1,\dots,\mathbf{x}_m)$. Finally, the distribution of the test statistic under $H_0$ can be approximated by the distribution

$$\binom{2d}{2}\sum_{i=1}^{\infty}\frac{\widehat{\nu}_{m,i}(\mathbf{x}_1,\dots,\mathbf{x}_m)}{m}Z_i^2. \tag{4.42}$$

Using this distribution we can construct the following hypothesis test.

**Definition 4.18 (eigenvalue based test for dHSIC)**
*Assume Setting 3.1, let $\alpha \in (0,1)$ and for all $m \in \mathbb{N}$ let $\widehat{F}_m : \boldsymbol{\mathcal{X}}^m \times \mathbb{R} \to [0,1]$ be the function satisfying for all $(\mathbf{x}_1,\dots,\mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that $\widehat{F}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)$ is the distribution function corresponding to the distribution of (4.42). Then the hypothesis test $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ satisfying for all $m \in \{1,\dots,2d-1\}$ that $\varphi_m := 0$ and for all $m \in \{2d, 2d+1,\dots\}$ and for all $(\mathbf{x}_1,\dots,\mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ that*

$$\varphi_m(\mathbf{x}_1,\dots,\mathbf{x}_m) := \mathbb{1}_{\{m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)>\widehat{F}_m^{-1}(\mathbf{x}_1,\dots,\mathbf{x}_m)(1-\alpha)\}}$$

*is called the eigenvalue based $\alpha$-test for dHSIC.*

Together with the tools described in the previous sections we can show that the first step in the approximation is consistent in the following sense.

**Lemma 4.19 (Type I and Type II error for first approximation step)**
*Assume Setting 3.1, let $H_0$ and $H_A$ be defined as in (4.1) and (4.2), respectively, let $\alpha \in (0,1)$ and let $F_m$ be the distribution function of $\binom{2d}{2}\sum_{j=1}^{\infty}\widehat{\lambda}_{m,j}Z_j^2$, with $\widehat{\lambda}_{m,j}$ defined as in (4.33). If $\mathbf{X}_1,\mathbf{X}_2,\dots \overset{iid}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$ it holds that*

$$\limsup_{m\to\infty}\mathbb{P}\left(m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) > F_m^{-1}(1-\alpha)\right) = \alpha$$

*and if $\mathbf{X}_1,\mathbf{X}_2,\dots \overset{iid}{\sim} \mathbb{P}^{\mathbf{X}} \in H_A$ it holds that*

$$\limsup_{m\to\infty}\mathbb{P}\left(m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) \leq F_m^{-1}(1-\alpha)\right) = 0.$$

**Proof** Theorem 4.17 implies for all $t \in \mathbb{R}$ that

$$\lim_{m\to\infty}F_m(t) = \lim_{m\to\infty}\mathbb{P}'\left(\binom{2d}{2}\sum_{i=1}^{\infty}\widehat{\lambda}'_{m,i}Z_i^2 \leq t\right) = \mathbb{P}'\left(\binom{2d}{2}\sum_{i=1}^{\infty}\lambda_i Z_i^2 \leq t\right) =: F(t).$$

Since $F$ is continuous everywhere it is well-known (see Theorem B.1) that for all $x \in (0,1)$ it holds that

$$\lim_{m\to\infty}F_m^{-1}(x) = F^{-1}(t). \tag{4.43}$$

Now, let $\mathbf{X}_1,\mathbf{X}_2,\dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_0$. Using Theorem 4.3 (and assuming that $\xi_2(h) > 0$ see remark below the theorem) it holds

$$m\cdot\widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) \xrightarrow{d} \binom{2d}{2}\sum_{j=1}^{\infty}\lambda_j Z_j^2 \tag{4.44}$$

as $m \to \infty$. Finally, combining Corollary B.3 with (4.43) and (4.44) it holds that

$$\begin{aligned}
&\limsup_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) > F_m^{-1}(1-\alpha)\right)\\
&\quad = 1 - \liminf_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) \leq F_m^{-1}(1-\alpha)\right)\\
&\quad = 1 - F(F^{-1}(1-\alpha))\\
&\quad = \alpha.
\end{aligned}$$

This shows that $\varphi$ has pointwise asymptotic level $\alpha$.

Now, let $\mathbf{X}_1, \mathbf{X}_2, \dots \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}} \in H_A$. Using that a convergent sequence is bounded (4.43) implies that there exists a $t^* \in \mathbb{R}$ such that

$$\sup_{m\in\mathbb{N}} F_m^{-1}(1-\alpha) \leq t^*. \tag{4.45}$$

Hence, (4.45) together with Corollary 4.2 shows that

$$\begin{aligned}
&\limsup_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) \leq F_m^{-1}(1-\alpha)\right)\\
&\quad \leq \limsup_{m\to\infty} \mathbb{P}\left(m \cdot \widehat{\mathrm{dHSIC}}_m(\mathbf{X}_1,\dots,\mathbf{X}_m) \leq t^*\right)\\
&\quad = 0.
\end{aligned}$$

This proves that $\varphi$ is pointwise consistent and therefore completes the proof of Theorem 4.20. □

In order to show that the second approximation step does not affect the Type I and Type II errors, one needs to show that the distribution function $\widehat{F}_m(\mathbf{X}_1,\dots,\mathbf{X}_m)$ from Definition 4.18 is a good approximation of $F_m$ from Lemma 4.19. Then we could prove the following desirable result.

**Conjecture 4.20 (level and consistency of the eigenvalue based test for dHSIC)** *Assume Setting 3.1, let $H_0$ and $H_A$ be defined as in* (4.1) *and* (4.2)*, respectively, let $\alpha \in (0,1)$ and let $\varphi = (\varphi_m)_{m\in\mathbb{N}}$ be the eigenvalue based $\alpha$-test for dHSIC. Then, $\varphi$ is a pointwise consistent hypothesis test with pointwise asymptotic level $\alpha$ for testing $H_0$ against $H_A$.*

## 4.5 Implementation details

In the previous sections, we introduced

(i) the permutation test for dHSIC (Definition 4.4),

(ii) the bootstrap test for dHSIC (Definition 4.6),

(iii) the gamma approximation based test for dHSIC (Definition 4.13) and

(iv) the eigenvalue based test for dHSIC (Definition 4.18).

The definitions are quite abstract, as they are tailored for a theoretical analysis. This section aims at giving additional details on how to implement these abstractly defined hypothesis tests on a more practical level. The details on how to compute $\widehat{\text{dHSIC}}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_m) \in \boldsymbol{\mathcal{X}}^m$ efficiently are discussed in Section 3.4 and are omitted here.

### 4.5.1 Permutation/Bootstrap

Fix $\alpha \in (0, 1)$ and assume we observe $(\mathbf{X}_1, \dots, \mathbf{X}_m) = (\mathbf{x}_1, \dots, \mathbf{x}_m)$. In order to apply the permutation test for dHSIC or the bootstrap test for dHSIC, one has to calculate the quantile

$$\left(\widehat{R}_m(\mathbf{x}_1, \dots, \mathbf{x}_m)\right)^{-1}(1 - \alpha). \tag{4.46}$$

From the definition of $\widehat{R}_m$ it is clear that this involves $(m!)^d$ evaluations of $\widehat{\text{dHSIC}}$ for the permutation test and $m^{mp}$ evaluations of $\widehat{\text{dHSIC}}$ for the bootstrap test. In both settings this becomes computationally impossible rather fast as $m$ grows. Instead of computing $\widehat{R}_m$ explicitly one can use the Monte-Carlo approximation defined in Definition 2.60. Essentially this involves calculating the p-value given by

$$\widehat{p}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) := \frac{1 + \left|\{i \in \{1, \dots, B\} : T_m(g_{m,\boldsymbol{\psi}_i}(\mathbf{x}_1, \dots, \mathbf{x}_m)) \geq T_m(\mathbf{x}_1, \dots, \mathbf{x}_m)\}\right|}{1 + B},$$

where $(\boldsymbol{\psi}_i)_{i \in \mathbb{N}}$ is a sequence drawn from the uniform distribution on $A_m$ (i.e. on $(S_m)^d$ for the permutation test and on $B_m^d$ for the bootstrap test). The test then rejects the null hypothesis whenever $\widehat{p}_m(\mathbf{x}_1, \dots, \mathbf{x}_m) \leq \alpha$. The corresponding critical value is calculated according to Proposition 2.64. For further details see Section 2.4.2

For practical applications one generally fixes the value of $B$. Davison and Hinkley (1997) suggest to use a value of $B$ between 99 and 999.

In the following two section we give some additional details specific to the permutation test and the bootstrap test.

#### Permutation test

As shown in the proof of Proposition 4.5 the resampling method $g$ for the permutation test is a resampling group which satisfies the invariance condition (4.9). This allows us to apply Proposition 2.63 to see that the Monte-Carlo approximated permutation test has valid level, given that we have continuous random variables as input.

Algorithm 2 shows how to implement the p-value and the critical value for the Monte-Carlo approximated permutation test.

### Bootstrap test

We showed that the bootstrap test for dHSIC has pointwise asymptotic level and is pointwise consistent. Both of these properties are preserved if we use the Monte-Carlo approximation and let $B$ tend to infinity.

Algorithm 2 shows how to implement the p-value and the critical value for the Monte-Carlo approximated bootstrap test.

**Algorithm 2** computing p-value and critical value for the permutation/bootstrap test

```
procedure MONTECARLO-PVALUE($\mathbf{x}_1, \dots \mathbf{x}_m, B$)
    initialize empty $B$-dimensional vector $\mathbf{T}$
    for $k = 1 : B$ do
        initialize $d$-dimensional vectors $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_m$
        for $j = 1 : d$ do
            $\psi \leftarrow$ random element from $S_m$ (permutation) or $\{1, \dots, m\}^m$ (bootstrap)
            for $i = 1 : m$ do
                $\tilde{\mathbf{x}}_i[j] \leftarrow \mathbf{x}_{\psi(i)}[j]$
        $\mathbf{T}[k] \leftarrow$ dHSIC($\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_m$)
    tmp $\leftarrow \#\{k \in \{1, \dots, B\} \mid \mathbf{T}[k] \geq$ dHSIC($\mathbf{x}_1, \dots \mathbf{x}_m$)$)\}$
    pval $\leftarrow$ (tmp $+ 1)/(B + 1)$
    return pval

procedure MONTECARLO-CRITVAL($\mathbf{x}_1, \dots \mathbf{x}_m, B, \alpha$)
    initialize empty $B$-dimensional vector $\mathbf{T}$
    for $k = 1 : B$ do
        initialize $d$-dimensional vectors $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_m$
        for $j = 1 : d$ do
            $\psi \leftarrow$ random element from $S_m$ (permutation) or $\{1, \dots, m\}^m$ (bootstrap)
            for $i = 1 : m$ do
                $\tilde{\mathbf{x}}_i[j] \leftarrow \mathbf{x}_{\psi(i)}[j]$
        $\mathbf{T}[k] \leftarrow m \cdot$ dHSIC($\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_m$)
    tmp $\leftarrow \#\{k \in \{1, \dots, B\} \mid \mathbf{T}[k] =$ dHSIC($\mathbf{x}_1, \dots \mathbf{x}_m$)$)\}$
    ind $\leftarrow \lceil (B + 1) \cdot (1 - \alpha) \rceil +$ tmp
    if ind $\leq B$ then
        $\mathbf{S} \leftarrow$ sort($\mathbf{T}$) (in ascending order)
        critval $\leftarrow \mathbf{S}$[ind]
    else
        critval $\leftarrow \infty$
    return critval
```

### 4.5.2 Gamma approximation

Fix $\alpha \in (0,1)$ and assume we observe $(\mathbf{X}_1,\dots,\mathbf{X}_m) = (\mathbf{x}_1,\dots,\mathbf{x}_m)$. Implementing the gamma approximation based $\alpha$-test for dHSIC is essentially a 4 step process, where we use the notation from Section 4.3:

1) for all $j \in \{1,\dots,d\}$ implement the estimators $\widehat{e}_0(j),\dots,\widehat{e}_3(j)$,

2) based on (4.20) and (4.21) compute the estimates $\widehat{\text{Exp}}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)$ and $\widehat{\text{Var}}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)$,

3) using (4.22) and (4.23) compute the estimates $\widehat{\alpha}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)$ and $\widehat{\beta}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)$ and

4) compute the $1-\alpha$ quantile of the $\text{Gamma}(\widehat{\alpha}_m(\mathbf{x}_1,\dots,\mathbf{x}_m),\widehat{\beta}_m(\mathbf{x}_1,\dots,\mathbf{x}_m))$-distribution.

The hypothesis test rejects $H_0$ if $m\cdot\widehat{\text{dHSIC}}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)$ is greater than the threshold given by the $1-\alpha$ quantile of the $\text{Gamma}(\widehat{\alpha}_m(\mathbf{x}_1,\dots,\mathbf{x}_m),\widehat{\beta}_m(\mathbf{x}_1,\dots,\mathbf{x}_m))$-distribution calculated in step 4).

### 4.5.3 Eigenvalue approach

Fix $\alpha \in (0,1)$ and assume we observe $(\mathbf{X}_1,\dots,\mathbf{X}_m) = (\mathbf{x}_1,\dots,\mathbf{x}_m)$. Implementing the eigenvalue approach based $\alpha$-test for dHSIC can be broken down into the following steps, where we use the notation of Section 4.4.3:

1) for all $j \in \{1,\dots,d\}$ implement the estimators $\widehat{e}_0(j)$ and $\widehat{e}_1(j)$,

2) based on Lemma C.2 compute the estimates $\widehat{a}_1(\mathbf{x}_1,\dots,\mathbf{x}_m),\dots,\widehat{a}_{10}(\mathbf{x}_1,\dots,\mathbf{x}_m)$ by replacing the expectations by the corresponding estimator $\widehat{e}_0(j)$ and $\widehat{e}_1(j)$ implemented in step 1),

3) also based on Lemma C.2 compute the estimate

$$\widehat{\mathbf{H}}_2(\mathbf{x}_1,\dots,\mathbf{x}_m) = \binom{2d}{2}^{-1}\sum_{i=1}^{10}\widehat{a}_i(\mathbf{x}_1,\dots,\mathbf{x}_m),$$

4) compute the estimates $\widehat{\nu}_{m,1}(\mathbf{x}_1,\dots,\mathbf{x}_m),\dots,\widehat{\nu}_{m,m}(\mathbf{x}_1,\dots,\mathbf{x}_m)$ by computing the eigenvalues of the matrix $\widehat{\mathbf{H}}_2(\mathbf{x}_1,\dots,\mathbf{x}_m)$,

5) use a Monte-Carlo approach to approximate $(1-\alpha)$-quantile of the distribution

$$\sum_{i=1}^{m}\frac{\widehat{\nu}_{m,1}(\mathbf{x}_1,\dots,\mathbf{x}_m)}{m}Z_i,$$

where $(Z_i)_{i\in\mathbb{N}}$ are standard normal random variables.

The hypothesis test rejects $H_0$ if $m\cdot\widehat{\text{dHSIC}}_m(\mathbf{x}_1,\dots,\mathbf{x}_m)$ is greater than then the $(1-\alpha)$-quantile computed in the last step.

Chapter 5

# Simulations

In this section we perform some basic simulations to verify the properties of the four hypothesis tests derived in the previous section. We separate the simulations into three parts; level analysis, power analysis and runtime analysis.

## 5.1 Competing method

For comparison purposes we use a multiple testing version of the two variable HSIC test. The idea is that if we can group several variables together, we can test the newly constructed multivariate variable against a different variable using the two variable HSIC. Grouping is for example possible if the variables are vectors, in which case we can simply bind them together to a matrix. In order to test for joint independence we use the following testing sequence,

1. use HSIC to test whether $X^d$ is independent of $[X^1, \dots, X^{d-1}]$,

2. use HSIC to test whether $X^{d-1}$ is independent of $[X^1, \dots, X^{d-2}]$,

...

d-1. use HSIC to test whether $X^2$ is independent of $X^1$.

Finally, we account for the increased family-wise error rate using the Bonferroni correction, i.e. we perform all tests at level $\frac{\alpha}{d-1}$ and reject the null hypothesis if any of the individual tests reject the null hypothesis. In the following sections we refer to this test simply as HSIC.

This method is of course not restricted to HSIC but can be performed for any two variable independence test.

## 5.2 Level analysis

In Chapter 4 we proved the following results related to the significance level of the hypothesis tests:

(i) the permutation test has valid level (even the Monte-Carlo approximated test),

(ii) the bootstrap test has pointwise asymptotic level,

(iii) the gamma approximation based test has no guarantee for level, and

(iv) the eigenvalue approximation based test has pointwise asymptotic level.

We verify these results numerically by considering two examples of fixed elements $\mathbb{P}^{\mathbf{X}} \in H_0$. In both examples we simulate $n = 1000$ realizations of $\mathbf{X}_1, \dots, \mathbf{X}_m \overset{\text{iid}}{\sim} \mathbb{P}^{\mathbf{X}}$ for different sample sizes $m$ and check how often each of the four hypothesis tests rejects the null hypothesis.

**Simulation 1 (testing level)**

Consider $X^1, X^2, X^3 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then for $\mathbf{X} = (X^1, X^2, X^3)$ it holds that

$$\mathbb{P}^{\mathbf{X}} = \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2} \otimes \mathbb{P}^{X^3} \in H_0,$$

where $H_0$ is the null hypothesis defined in (4.1). Set $\alpha = 0.05$, $B = 25$, and $m \in \{100, 200, \dots, 1000\}$. The rejection rates for the corresponding four hypothesis tests (permutation, bootstrap, gamma approximation and eigenvalue) based on $n = 1000$ repeated draws of $\mathbf{X}$ are plotted in Figure 5.1.

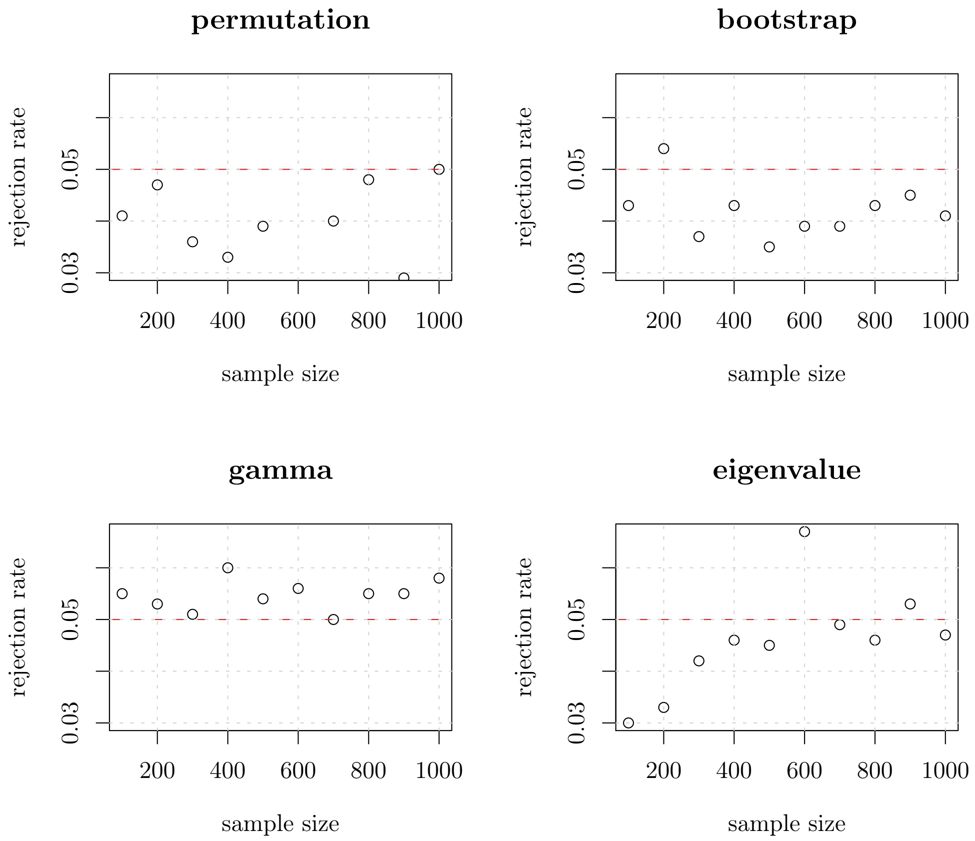


Figure 5.1: Simulation 1 (testing level): Rejection rates (based on n=1000 repetitions) for each of the four different hypothesis tests based on dHSIC. The test has valid level if the rejection rate lies below the dotted red line at 0.05.

**Simulation 2 (testing level)**

Consider $X^1 \sim \mathcal{N}(0,1)$ and $X^2 \sim \text{Bin}(20, 0.2)$ with $X^1$ and $X^2$ independent. Then for $\mathbf{X} = (X^1, X^2)$ it holds that

$$\mathbb{P}^{\mathbf{X}} = \mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2} \in H_0,$$

where $H_0$ is the null hypothesis defined in (4.1). Set $\alpha = 0.05$, $B = 25$, and $m \in \{100, 200, \ldots, 1000\}$. The rejection rates for the corresponding four hypothesis tests (permutation, bootstrap, gamma approximation and eigenvalue) based on $n = 1000$ repeated draws of $\mathbf{X}$ are plotted in Figure 5.2.

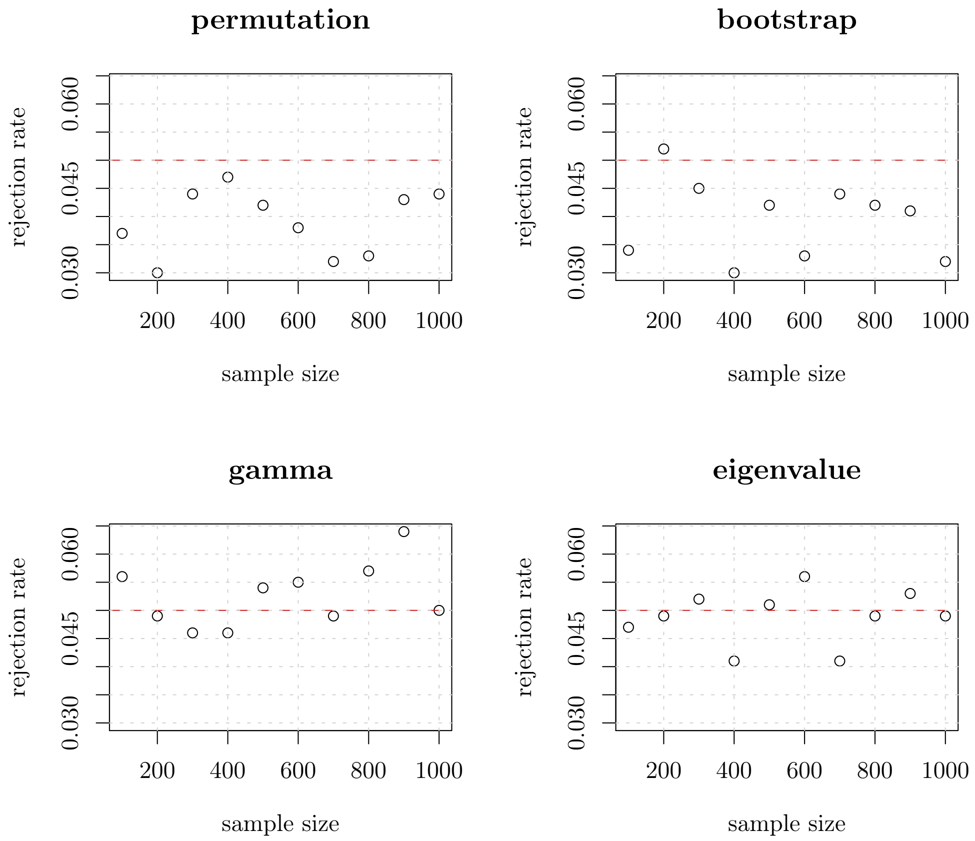


Figure 5.2: Simulation 2 (testing level): Rejection rates (based on n=1000 repetitions) for each of the four different hypothesis tests based on dHSIC. The test has valid level if the rejection rate lies below the dotted red line at 0.05.

In both simulations we get similar results. We collect the most important observations.

(i) The permutation test is the only test that achieves level $\alpha$. This corresponds to what has been proved in the previous section. As mentioned above, this result is rather surprising as it is does not depend on the choice of $B$, which in these examples is very small ($B = 25$).

(ii) The gamma approximation based test, at least in these two examples, has level close to $\alpha$. It however also shows that one has to be very careful, when analyzing results based on this test, since it often exceeds the required level.

(iii) It has been proved in the previous section that both the bootstrap test and the eigenvalue test have pointwise asymptotic level. In both examples this convergence cannot be observed for sample sizes between 100 and 1000. This shows that the theoretically nice result of pointwise asymptotic level is rather weak in practical applications.

(iv) The bootstrap test appears to achieve level $\alpha$ in most cases. This is due to the conservative choice of the p-value in the Monte-Carlo approximation of the bootstrap test.

## 5.3 Power analysis

Assessing the power of a test can be done for many different alternatives. Here, we show a single simulation setting that compares the four hypothesis tests based on dHSIC.

**Simulation 3 (comparing power)**

Consider $N^1, N^2, N^3 \sim \mathcal{N}(0,1)$ and let $\mathbb{P}^{\mathbf{X}}$ be generated by the SEM,

$$\begin{aligned} X^1 &= N^1 \\ X^2 &= \lambda \cos(X^1) + N^2 \\ X^3 &= \lambda \cos(X^1) + \lambda \cos(X^2) + N^3. \end{aligned}$$

Then for $\lambda = 0$ it holds that $\mathbb{P}^{\mathbf{X}} \in H_0$ and for $\lambda \in (0,1]$ it holds that $\mathbb{P}^{\mathbf{X}} \in H_0$, where $H_0$ and $H_A$ are defined in (4.1) and (4.2). For $\alpha = 0.05$ and $B = 100$, the rejection rates for the corresponding four hypothesis tests (permutation, bootstrap, gamma approximation and eigenvalue) and the multiple testing approach using HSIC (with the permutation test) based on $n = 1000$ repeated draws of $\mathbf{X}$ are plotted in Figure 5.3 for different values of $\lambda$.

Although it appears like the dHSIC test has more power than the multiple testing approach with HSIC, this depends strongly on the dependence under considerations. For some dependencies dHSIC is more powerful and for other dependencies the multiple testing approach with HSIC is more powerful.

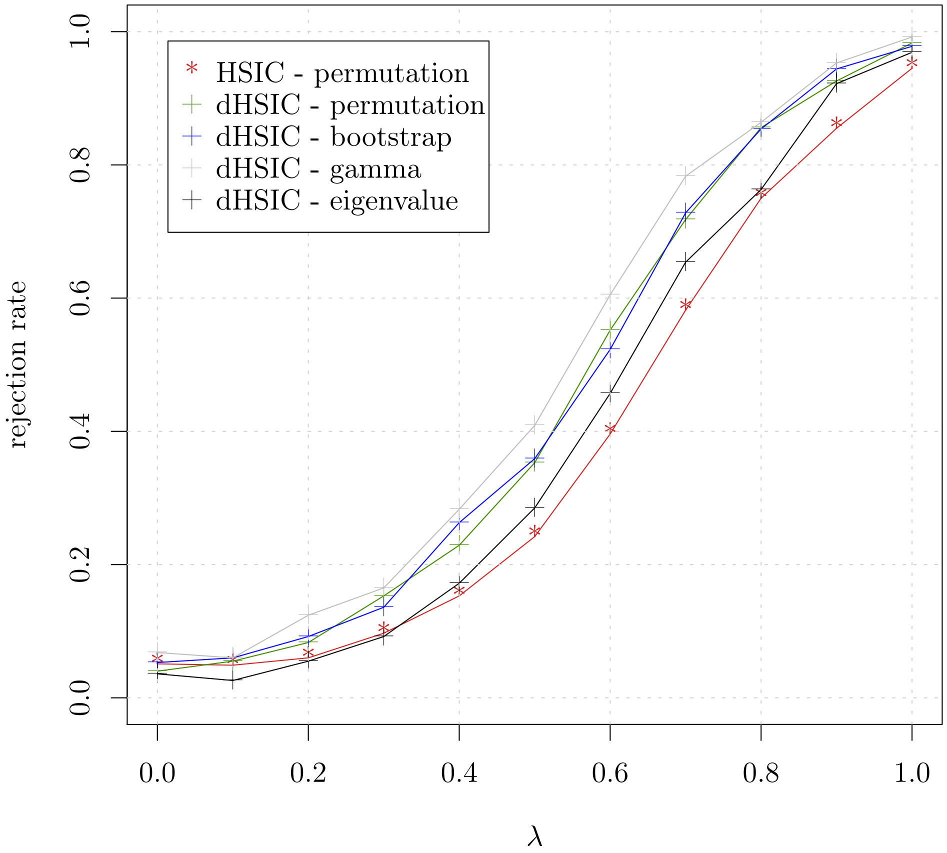


Figure 5.3: Simulation 3 (comparing power): Rejection rates (based on n=1000 repetitions) for each of the four different hypothesis tests based on dHSIC and the competing method based on HSIC. Larger values of $\lambda$ correspond to stronger dependencies between the variables.

## 5.4 Runtime analysis

Finally, we want to compare the runtime of the dHSIC test with the runtime of the multiple testing approach for HSIC.

The computational complexity for the dHSIC test statistic is

$$\mathcal{O}\left(dm^2\right)$$

as can be seen from the considerations in Section 3.4. The multiple testing approach for HSIC computes HSIC $d-1$-times, which appears to result in the same computational complexity. However, since the dimension of the input variables for the HSIC tests are also dependent on $d$ (at least in common settings such as for the Gaussian kernel) due to the process of binding variables together, we in fact end up with a true computational complexity of

$$\mathcal{O}\left(d^2 m^2\right).$$

We numerically test these computational complexities by two simulations. In the first simulation we fix $m$ and let $d$ vary and in the second simulation we fix $d$ and let $m$ very. The results are presented in Figure 5.4 and in Figure 5.5.

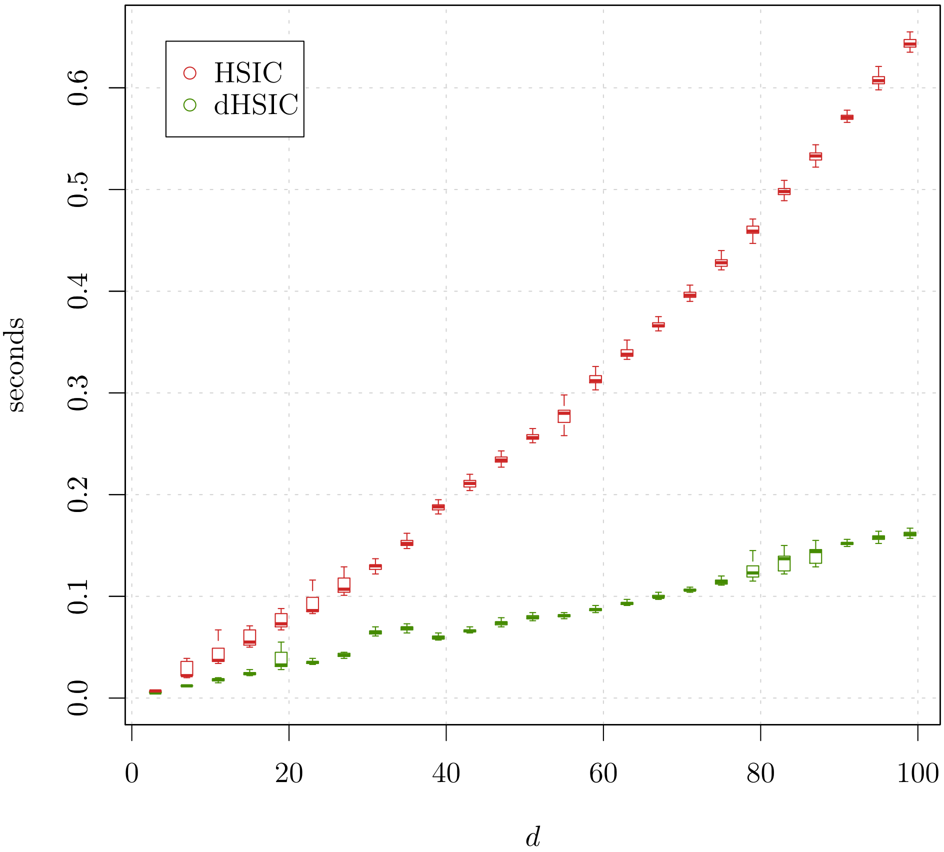


Figure 5.4: runtime analysis with varying variable number and fixed sample size ($m = 100$)

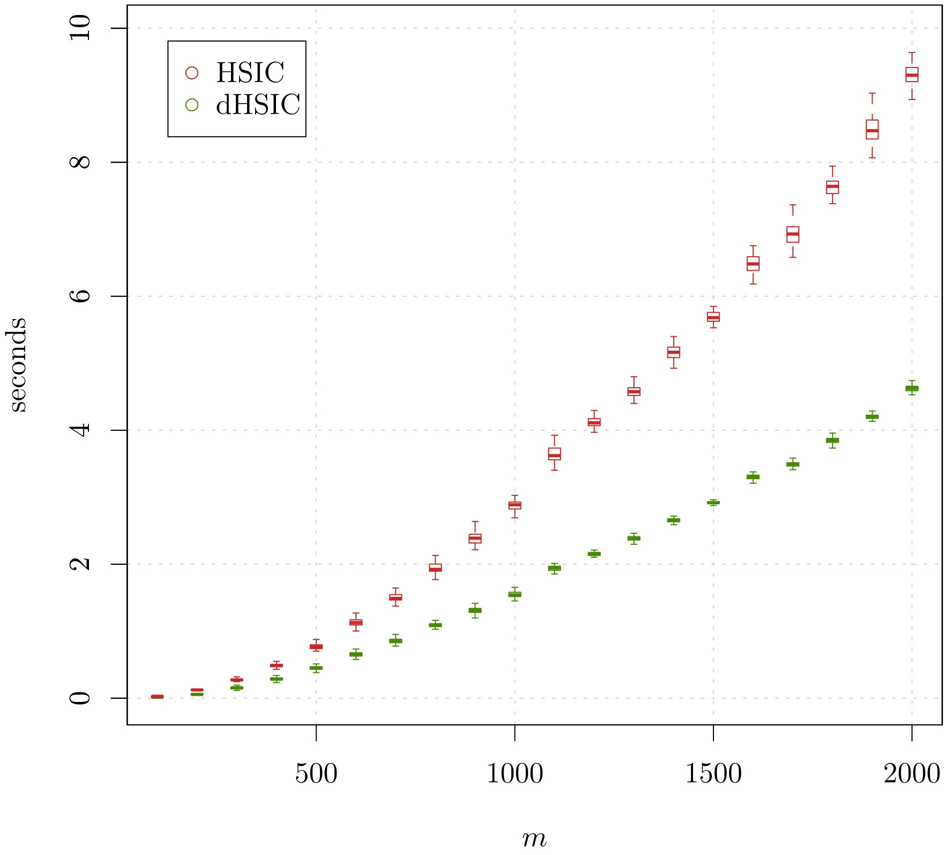


Figure 5.5: runtime analysis with varying sample size and fixed variable number ($d = 10$)

Chapter 6

# Applications to causal inference

Many methods in causal inference rely on independence testing. In this section we illustrate one such application by applying the d-variable Hilbert-Schmidt independence criterion to causal structure learning in additive noise models.

The following material is divided into three parts. First, we introduce additive noise models and state some important remarks related to identifiability. We then apply dHSIC to an simulated data example and conclude the section with a real-world data example.

## 6.1 Additive noise models

The following introduction to additive noise models is based on Peters et al. (2014). The reader is expected to be familiar with causal models, in particular directed acyclic graphs (DAGs), as well as structural equation models (SEMs). A summary of these concepts can be found in Peters et al. (2014, Section 1).

We consider a $d$-dimensional random vector $\mathbf{X} = (X^1, \dots, X^d)$ with joint distribution $\mathbb{P}^{\mathbf{X}}$ and assume it is generated by an SEM with structural equations $\mathcal{S}$, noise variables $\mathbf{N} = (N^1, \dots, N^d)$ and associated DAG $\mathcal{G}$. An important question in causality is whether the causal structure, in this case $\mathcal{G}$, can be inferred from the observational distribution $\mathbb{P}^{\mathbf{X}}$ alone. In general, this is impossible without additional assumptions on the model class (see Peters et al., 2014, Proposition 9). For graphical models one option is to assume that $\mathbb{P}^{\mathbf{X}}$ is faithful and Markov with respect to the DAG $\mathcal{G}$, then the Markov equivalence class of $\mathcal{G}$ is uniquely determined by $\mathbb{P}^{\mathbf{X}}$. This is, for example, the fundamental assumption behind conditional independence-based methods for causal structure learning. For functional models restricting the model class can be done by specifying the class of allowed functions appearing in the structural equations $\mathcal{S}$. There are several restrictions, under which the graph $\mathcal{G}$ becomes identifiable from $\mathbb{P}^{\mathbf{X}}$. We want to focus on the class of functions which are additive with respect to the noise variables $\mathbf{N} = (N^1, \dots, N^d)$. The resulting SEMs are called continuous additive noise models (see Peters et al., 2014, Definition 16).

**Definition 6.1 (continuous additive noise model (ANM))**
*A continuous additive noise model (ANM) is defined as an SEM $(\mathcal{S}, \mathbb{P}^{\mathbf{N}})$, where $\mathcal{S} = (S^1, \ldots, S^d)$ is a collection of $d$ equations of the form*

$$S^j: \quad X^j = f^j(\mathbf{PA}^j) + N^j, \quad j \in \{1, \ldots, d\},$$

*and where $\mathbf{N} = (N^1, \ldots, N^d)$ is jointly independent and with strictly positive density (with respect to the Lebesgue measure). $\mathbf{PA}^j$ denotes the parents of the node $X^j$ in the graph associated to $(\mathcal{S}, \mathbb{P}^{\mathbf{N}})$.*

In general, continuous additive noise models are not identifiable from $\mathbb{P}^{\mathbf{X}}$. The classical counter-example is if the functions are linear and the noise variables are Gaussian distributed. Models of this type are called linear Gaussian and the next proposition shows that the causal order of such models can be reversed (see Peters et al., 2014, Proposition 13).

**Proposition 6.2 (linear Gaussian model is reversible)**
*Let $X$ and $N$ be two Gaussian random variables with $N \perp\!\!\!\perp X$, let $\alpha \neq 0$ and let*

$$Y = \alpha X + N.$$

*Then there exists $\beta \in \mathbb{R}$ and a Gaussian random variable $\tilde{N}$ with $\tilde{N} \perp\!\!\!\perp Y$, such that*

$$X = \beta Y + \tilde{N}.$$

It turns out that the linear Gaussian case is rather special in this respect. If for example we only allow non-linear functions and Gaussian noise the following theorem shows that the DAG is uniquely identifiable from $\mathbb{P}^{\mathbf{X}}$ (see Peters et al., 2014, Corollary 31). Models of this type are referred to as non-linear Gaussian ANMs.

**Theorem 6.3 (non-linear Gaussian model is identifiable)**
*Let $\mathbb{P}^{\mathbf{X}}$ be generated by the additive noise model $(\mathcal{S}, \mathbb{P}^{\mathbf{N}})$ given by*

$$S^j: \quad X^j = f^j(\mathbf{PA}^j) + N^j, \quad j \in \{1, \ldots, d\},$$

*with normally distributed noise variables $N^j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable functions $f^j$ that are not linear in any component, i.e. if we denote the parents $\mathbf{PA}^j$ of $X^j$ by $X^{k_1}, \ldots, X^{k_l}$, then the function $f^j(x^{k_1}, \ldots, x^{k_{a-1}}, \cdot, x^{k_{a+1}}, \ldots, x^{k_l})$ is assumed to be nonlinear for all $a$ and some $(x^{k_1}, \ldots, x^{k_{a-1}}, x^{k_{a+1}}, \ldots, x^{k_l}) \in \mathbb{R}^{l-1}$. Then we can identify the corresponding DAG from the distribution $\mathbb{P}^{\mathbf{X}}$.*

A simplification of this setting is given by structural equations of the form

$$S^j: \quad X^j = \sum_{k \in \mathbf{PA}_j} f^{j,k}(X^k) + N^j, \quad j \in \{1, \ldots, d\}, \tag{6.1}$$

with normally distributed noise variables $N^j \sim \mathcal{N}(0, \sigma_j^2)$ and three times differentiable, nonlinear functions $f^{j,k}$. Models of this type (non-Gaussian noise included) are called causal additive models (CAM) (see Bühlmann et al., 2014).

Assuming we are given a random variable $\mathbb{P}^{\mathbf{X}}$ generated by a non-linear Gaussian ANM as in Theorem 6.3, we can write the functions $f^j$ in terms of the expectation as

$$f^j = \underset{g \text{ additive}}{\arg\min}\, \mathbb{E}\left(\left(X^j - g\left(\mathbf{PA}^j\right)\right)^2\right),$$

for details see Bühlmann et al. (2014). Theorem 6.3 now ensures that there is exactly one DAG such that

$$X^1 - f^1\left(\mathbf{PA}^1\right), \dots, X^d - f^d\left(\mathbf{PA}^d\right)$$

are independent Gaussian distributed random variables. Using a generalized regression method to estimate $f^j(\mathbf{pa}^j)$ and then concentrating on the independence of the residuals, gives us a method to check whether a given DAG is correct. We make this method explicit for models of the form (6.1) using generalized additive model regression (GAM) (Wood and Augustin, 2002).

**DAG verification method**

Given: observations $\mathbf{X}_1, \dots, \mathbf{X}_m$ and a candidate DAG $\mathcal{G}$

1) Use generalized additive model regression (GAM) to regress each node $X^j$ on all its parents $\mathbf{PA}^j$ and denote the resulting vector of residuals by $\mathbf{res}^j$.
2) Perform a $d$-variable joint independence test (e.g. dHSIC) to test whether $(\mathbf{res}^1, \dots, \mathbf{res}^d)$ is jointly independent.
3) If $(\mathbf{res}^1, \dots, \mathbf{res}^d)$ is jointly independent, then the DAG is not rejected.

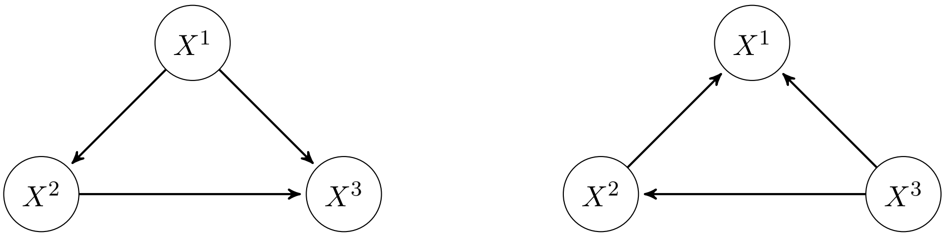
We can also use this DAG verification method to find the correct DAG by performing the check for all possible DAGs with the correct number of nodes. In practice, we do not want to iterate over all possible graphs. A more efficient method, which is based on a similar idea, is the RESIT (regression with subsequent independence test) algorithm described in Peters et al. (2014, Section 4.1) and Bühlmann et al. (2014).

## 6.2 Simulation example

In this subsection, we consider two explicit continuous additive noise models. We then generate data from one of them and use the model verification method described at the end of the previous section to check whether we are able to determine the correct model based only on the data.

We denote the two ANMs by $(\mathcal{S}_1, \mathbb{P}^{\mathbf{N}_1})$ and $(\mathcal{S}_2, \mathbb{P}^{\mathbf{N}_2})$ and assume they satisfy

$$\mathcal{S}_1 \begin{cases} X^1 = N_1^1 \\ X^2 = f_\lambda(X^1) + N_1^2 \\ X^3 = f_\lambda(X^1) + f_\lambda(X^2) + N_1^3 \end{cases} \quad \text{and} \quad \mathcal{S}_2 \begin{cases} X^1 = g(X^3) + h(X^2) + N_2^1 \\ X^2 = j(X^3) + N_2^2 \\ X^3 = N_2^3 \end{cases} \tag{6.2}$$



(a) graph $\mathcal{G}_1$

(b) graph $\mathcal{G}_2$

Figure 6.1: graphical representation of the two ANMs $\mathcal{S}_1$ and $\mathcal{S}_2$ from (6.2)

with $N_i^j \sim \mathcal{N}(0, \sigma_i^j)$ independent normally distributed random variables. The corresponding graphs are given in Figure 6.1.

If $f_\lambda$ is linear we are in the linear Gaussian setting and cannot identify the correct graph due to Proposition 6.2. However, if $f_\lambda$ is non-linear we are in the setting of (6.1) and by Theorem 6.3 are able to determine the correct graph.

For the following simulation we choose the two functions

$$f_\lambda(x) := (1-\lambda)x + \lambda\cos(x),$$

and

$$f_\lambda(x) := (1-\lambda)x + \lambda|x|,$$

with $\lambda \in [0,1]$. The parameter $\lambda$ should be understood as a quantifier of how non-linear $f_\lambda$ is. We then simulate 1000 data samples consisting of $m = 100$ data points from $\mathcal{S}_1$ and perform the model verification method for both graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ to check if they are accepted or rejected.

The result for different values of $\lambda$ is plotted in Figure 6.2 (for the cos non-linearity) and in Figure 6.3 (for the $|\cdot|$ non-linearity). We used the permutation test for dHSIC with $B = 100$ as well as a multiple testing version of the classical two variable HSIC test (also using the permutation test with $B = 100$) for comparison.

This shows that the power of the test depends on the dependence structure of the variables. In particular, one cannot say dHSIC has more power than a multiple testing approach with HSIC or vice versa. The difference becomes more pronounced if we extend the three variable models $\mathcal{S}_1$ and $\mathcal{S}_2$ to 5 variables. Two corresponding plots are given in Figure 6.4 and Figure 6.5.

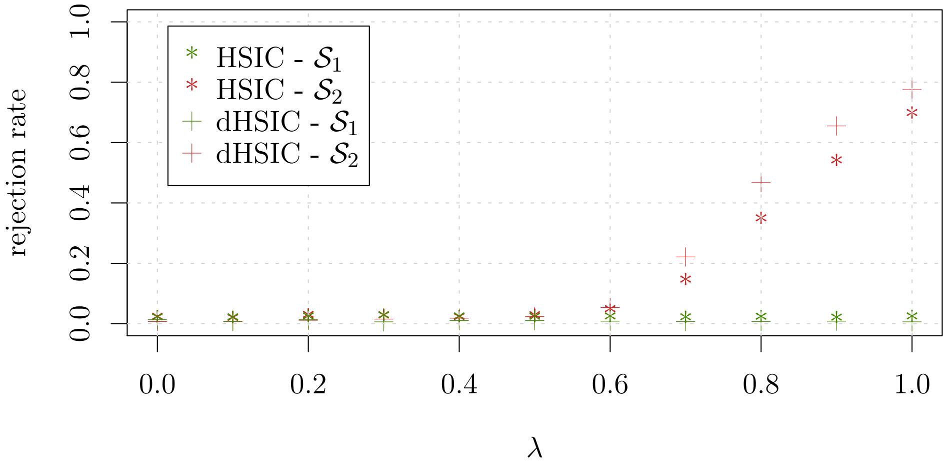


Figure 6.2: Rejection rates (based on $n = 1000$ repetitions), where $\mathcal{S}_1$ is the correct model with 3 variables $f_\lambda(x) = (1 - \lambda)x + \lambda \cos(x)$ and $\mathcal{S}_2$ is the corresponding false model.

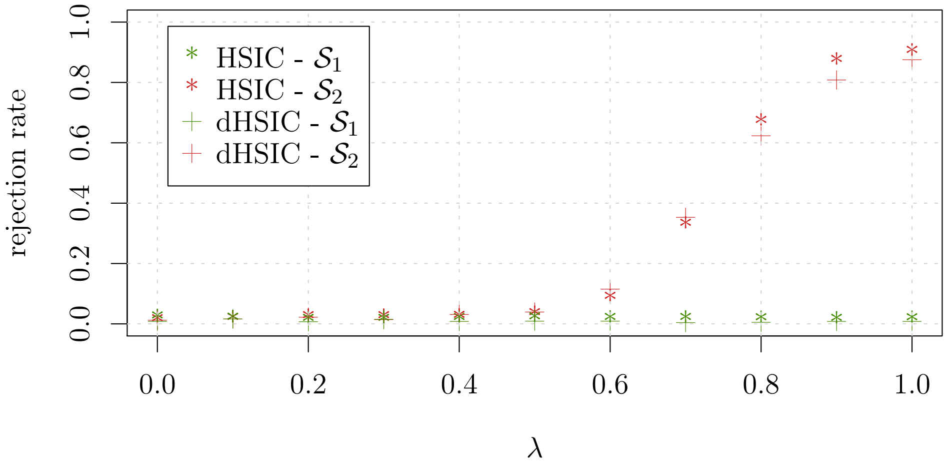


Figure 6.3: Rejection rates (based on $n = 1000$ repetitions), where $\mathcal{S}_1$ is the correct model with 3 variables $f_\lambda(x) = (1 - \lambda)x + \lambda|x|$ and $\mathcal{S}_2$ is the corresponding false model.

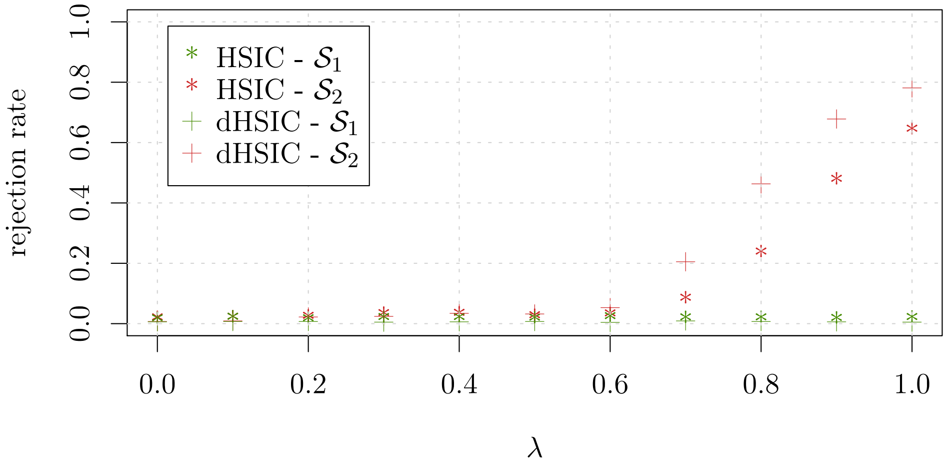


Figure 6.4: Rejection rates (based on $n = 1000$ repetitions), where $\mathcal{S}_1$ is the correct model with 5 variables $f_\lambda(x) = (1 - \lambda)x + \lambda \cos(x)$ and $\mathcal{S}_2$ is the corresponding false model.

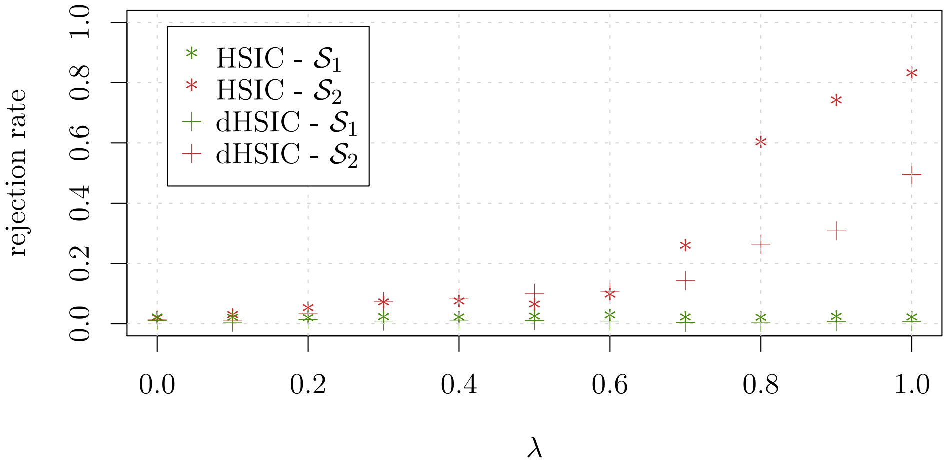


Figure 6.5: Rejection rates (based on $n = 1000$ repetitions), where $\mathcal{S}_1$ is the correct model with 5 variables $f_\lambda(x) = (1 - \lambda)x + \lambda|x|$ and $\mathcal{S}_2$ is the corresponding false model.

## 6.3 Real data example

In this section we want to show that the hypothesis tests we developed can also be applied to real world data. Consider the following causal inference problem: Given 349 measurements of the variables Altitude, Temperature and Sunshine.[1] Can we determine the correct causal ordering.

For 3 variables there exist a total of 25 possible DAGs. The idea is to iterate over all these DAGs and use the DAG verification method introduced at the end of Section 6.1 to check whether a DAG fits the data.

We apply the DAG verification method together with the permutation test for dHSIC (with $B = 1000$) and the multiple testing approach for HSIC (also with a permutation test and $B = 1000$) to every possible DAG and compare the resulting p-values. The result is shown in Figure 6.6.

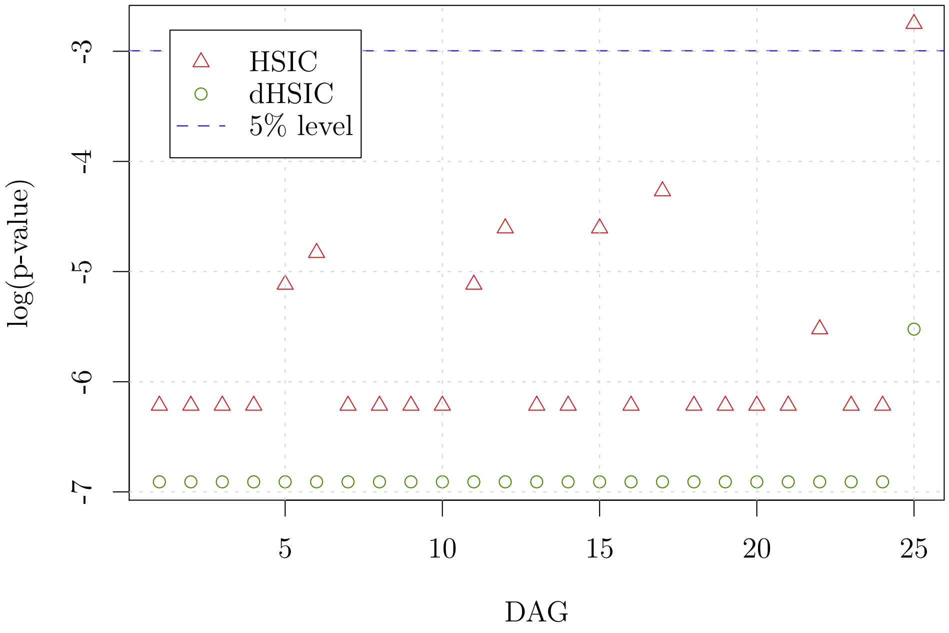


Figure 6.6: Real world data example: log(p-value) for every possible DAG on 3 nodes after applying the DAG verification method. DAG is rejected at at a significance level of 5% if log(p-value) lies below the blue line.

It shows that the dHSIC based test rejects DAG 25 at a 5%-significance level, while the

[1] The dataset is taken from Mooij et al. (2016, pair0001.txt and pair0004.txt).

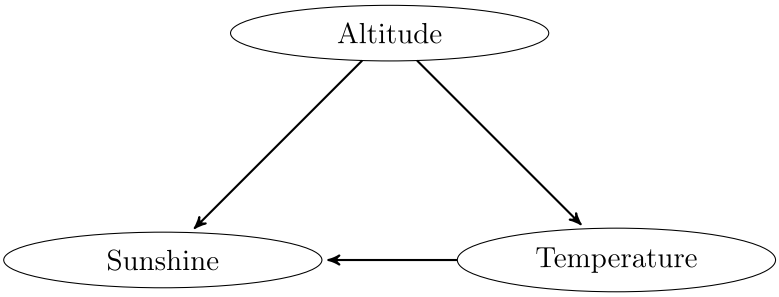


Figure 6.7: DAG 25

HSIC method does not. This implies that for this particular setting dHSIC has more power than HSIC.

Even though DAG 25 is rejected at a 5%-significance level by the dHSIC based test it still appears to be a good fit, because both tests indicate that DAG 25 is the best fit. The graph of DAG 25 is given in Figure 6.7. This causal structure makes sense from our physical understanding, as we would expect altitude to effect both sunshine and temperature. The effect of temperature on sunshine could be explained by intermediate latent variables such as cloud or fog.

The discrepancy between best fit and rejected model is a common issue that occurs when working with real data. The reason is that we can almost never expect to have the "true" physical model, which implies that given more and more data observations we will eventually reject any model. In the setting of real data it is therefore often beneficial to ask, which model in a given set of models fits the data best, instead of asking whether a given model fits the data.

Chapter 7

# Summary

We introduced a measure of joint dependence between $d$-variables, which we called the d-variable Hilbert-Schmidt independence criterion (dHSIC). This work extends the two-variable Hilbert-Schmidt independence criterion (HSIC). As in the HSIC case, we were able to estimate dHSIC empirically using a V-statistic and derive some important properties of the asymptotic distribution of this estimator. This allowed us to construct four different hypothesis tests, which all use this empirical estimator as a test statistic; the permutation test based on dHSIC (Definition 4.4), the bootstrap test based on dHSIC (Definition 4.6), the gamma approximation based test (Definition 4.13) and the eigenvalue approach based test (Definition 4.18).

For the permutation test we showed that it achieved level in Proposition 4.5. In particular, we also showed that this property carries over to the Monte-Carlo approximated version of the permutation test. This is a very strong property as the Monte-Carlo approximation based permutation test is computationally feasible even for moderately large sample sizes. The bootstrap test was defined very similar to the permutation test. The slight differences, however, allowed us to show that it is connected to the empirical product distribution, via the bootstrapping property stated in Proposition 4.8. This in turn was the central element in proving that it has pointwise asymptotic level (Proposition 4.9) and is consistent (Proposition 4.10). Using the theory on V-statistics, we were able to compute the mean and variance of the empirical estimator of dHSIC in Lemma 4.11 and Lemma 4.12. These were the essential ingredients in constructing the gamma approximation based test for dHSIC. Although this test has no guarantees on level and consistency, it is computationally very fast and therefore of particular interest for practical applications. Finally, we constructed the eigenvalue approach based test for dHSIC by estimating the eigenvalues in the asymptotic distribution of the test statistic directly. Using tools from functional analysis, we were able to show (up to the last approximation step) that the resulting test achieves pointwise asymptotic level and is consistent (Conjecture 4.20). It is therefore a viable alternative to the slower bootstrap test.

We gave implementation details on dHSIC in Section 3.4 and on each of the four hypothesis tests in Section 4.5. This should enable implementation of the tests in the exact same why as has been done for this thesis, hence allowing reproduction of all results.

Finally, we considered various simulations, which showed that the dHSIC based tests were able to compare with the commonly used multiple testing approach based on HSIC. The power of both approaches varies depending on the dependence under considerations; in some cases the dHSIC tests were more powerful, in other cases the multiple testing approach was more powerful. In terms of runtime, we were, however, able to show that dHSIC is generally faster and in terms of the number of variables even beats the multiple testing approach by an order of magnitude of one. Moreover, we also demonstrated that dHSIC can be successfully applied to causal inference both on simulated and real data.

Appendix A

# Functional Analysis

## A.1 Important theorems

### A.1.1 Mercer's theorem

The following version of Mercer's theorem is taken from Ferreira and Menegatto (2009, Theorem 1.1).

**Theorem A.1 (Mercer's theorem)**
*Let $X$ be a topological Hausdorff space equipped with a finite Borel measure $\mu$. Then for every continuous positive definite kernel $k : X \times X \to \mathbb{C}$ there exist a scalar absolutely summable sequence $(\lambda_n)_{n\in\mathbb{N}}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and an orthonormal system $(\varphi_n)_{n\in\mathbb{N}}$ in $L^2(X, \mu)$ consisting of continuous functions only, such that the expansion*

$$k(x, y) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(x)\overline{\varphi_n(y)}, \quad x, y \in \operatorname{supp}(\mu),$$

*converges uniformly. Furthermore let $\mathcal{K} \in L\left(L^2\left(\mathcal{X}, \mathbb{P}^X\right)\right)$ be the integral operator with the property that for every $f \in L^2\left(\mathcal{X}, \mathbb{P}^X\right)$ and for every $x \in X$ it hols that*

$$\left(\mathcal{K}(f)\right)(x) = \int_{\mathcal{X}} k(x, y) f(y)\, \mu(dy)$$

*Then $(\lambda_n)_{n\in\mathbb{N}}$ and $(\varphi_n)_{n\in\mathbb{N}}$ are eigenvalues respectively eigenfunctions of $\mathcal{K}$.*

### A.1.2 McDiarmid inequality

**Theorem A.2 (McDiarmid inequality)**
*Let $X_1, \ldots, X_n$ be $n$ independent random variables taking values in $\mathcal{X}$ and let $Z = f(X_1, \ldots, X_n)$ where $f$ is such that for all $i \in \{1, \ldots, n\}$ it holds that*

$$\sup_{x_1,\ldots,x_n,x_i'} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

*then*

$$\mathbb{P}(Z - \mathbb{E}(Z) \geq \xi) \leq \exp\left(\frac{-2\xi^2}{\sum_{i=1}^n c_i^2}\right)$$

*and*

$$\mathbb{P}(\mathbb{E}(Z) - Z \geq \xi) \leq \exp\left(\frac{-2\xi^2}{\sum_{i=1}^n c_i^2}\right)$$

## A.2 Tensor products of RKHS

Products of kernels are important when considering multidimensional settings. Therefore we give a very short overview of the most basic facts of tensor products of Hilbert spaces and how they behave in relation to reproducing kernel Hilbert spaces. Most of this section follows Berlinet and Thomas-Agnan (2004). More details on tensor product of Hilbert spaces can also be found in Weidmann (1980).

We begin by defining the tensor product for functions.

**Definition A.3 (tensor product for functions)**
*Let $\mathcal{X}$ and $\mathcal{Y}$ be two set, let $f \in \mathcal{F}(\mathcal{X})$ and let $g \in \mathcal{F}(\mathcal{Y})$, then the tensor product of $f$ and $g$ is defined as the function $f \otimes g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with the property that for all $x \in \mathcal{X}$ and for all $y \in \mathcal{Y}$ it holds that*

$$(f \otimes g)(x, y) = f(x)g(y).$$

Now given two set $\mathcal{X}$ and $\mathcal{Y}$, assume $\mathcal{H}^1$ is a RKHS on $\mathcal{X}$ and $\mathcal{H}^2$ is a RKHS on $\mathcal{Y}$. Then define

$$\mathcal{H}^1 \bar{\otimes} \mathcal{H}^2 = \operatorname{span}\{f \otimes g \mid f \in \mathcal{F}(\mathcal{X}), g \in \mathcal{F}(\mathcal{Y})\}.$$

This corresponds with the standard tensor product for vector spaces. Observe that so far $\mathcal{H}^1 \bar{\otimes} \mathcal{H}^2$ is just a vector space with no additional structure.

In order to add a Hilbert space structure we define $\langle \cdot, \cdot \rangle : \mathcal{H}^1 \bar{\otimes} \mathcal{H}^2 \to \mathbb{R}$ to be the function with the property that for all $f_1 \otimes g_1,\ f_2 \otimes g_2 \in \mathcal{H}^1 \bar{\otimes} \mathcal{H}^2$ it holds that

$$\langle f_1 \otimes g_1, f_2 \otimes g_2 \rangle_{\mathcal{H}^1 \bar{\otimes} \mathcal{H}^2} = \langle f_1, g_1 \rangle_{\mathcal{H}^1} \langle f_2, g_2 \rangle_{\mathcal{H}^2}.$$

It can be shown that this defines a scalar product on $\mathcal{H}^1 \bar{\otimes} \mathcal{H}^2$. Denote by $\mathcal{H}^1 \otimes \mathcal{H}^2$ the completion of $\mathcal{H}^1 \bar{\otimes} \mathcal{H}^2$. Then we call $\mathcal{H}^1 \otimes \mathcal{H}^2$ the tensor product of the Hilbert spaces $\mathcal{H}^1$ and $\mathcal{H}^2$.

Next define a tensor product for kernels.

**Definition A.4 (tensor product for kernels)**
*Let $\mathcal{X}$ and $\mathcal{Y}$ be two set and let $k^1$ and $k^2$ be kernels on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then we define the tensor product of $k^1$ and $k^2$ as the function $k^1 \otimes k^2 : (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}$ with the property that for all $((x_1, y_1), (x_2, y_2)) \in (\mathcal{X} \times \mathcal{Y})^2$ it holds that*

$$(k^1 \otimes k^2)((x_1, y_1), (x_2, y_2)) = k^1(x_1, x_2)k^2(y_1, y_2).$$

It follows immediately that $k^1 \otimes k^2$ is a kernel on $\mathcal{X} \times \mathcal{Y}$ according to Definition 2.12. It is also easy to check that if both $k^1$ and $k^2$ are positive semi-definite then also $k^1 \otimes k^2$ is positive semi-definite.

The following theorem shows that $\mathcal{H}^1 \otimes \mathcal{H}^2$ is a RKHS with reproducing kernel $k^1 \otimes k^2$, it can be found in Berlinet and Thomas-Agnan (2004, Theorem 13).

**Theorem A.5 (tensor RKHS)**
*Let $\mathcal{X}$ and $\mathcal{Y}$ be two sets and let $\mathcal{H}^1$ be a RKHS on $\mathcal{X}$ and $\mathcal{H}^2$ be RKHS on $\mathcal{Y}$. Denote by $k^1$ and $k^2$ the corresponding reproducing kernels. Then $\mathcal{H}^1 \otimes \mathcal{H}^2$ is a RKHS with reproducing kernel $k^1 \otimes k^2$.*

Clearly the same result also holds for an arbitrary finite family of RKHS.

## A.3 Bochner integral

In this section we shortly review the Bochner integral. This presentation is mainly based on the lecture notes by Jentzen (2015) but similar definitions and statements are given in Prévôt and Röckner (2007).

**Definition A.6 (strongly measurable)**
*Let $\mathcal{B}$ be a Banach space, let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space. Then a function $f : \Omega \to \mathcal{B}$ is called strongly measurable if $f$ is measurable and $f(\Omega)$ is separable.*

In the case where $\mathcal{B}$ is separable, strong measurability coincides with the standard notion of measurability.

**Definition A.7 (simple function)**
*Let $\mathcal{B}$ be a Banach space, let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space. Then a function $f : \Omega \to H$ is called simple if $f(\Omega)$ is finite.*

**Definition A.8 ($\mathcal{L}^p$-space)**
*Let $\mathcal{B}$ be a Banach space, let $p \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space. Then the set*

$$\mathcal{L}^p\left(\mu, \|\cdot\|_{\mathcal{B}}\right) := \Big\{ f : \Omega \to \mathcal{B} \mid f \text{ is strongly measurable and} \\ \|f\|_{\mathcal{L}^p(\mu, \|\cdot\|_{\mathcal{B}})} := \left( \int_\Omega \|f(\omega)\|_{\mathcal{B}}^p \, \mu(d\omega) \right)^{\frac{1}{p}} < \infty \Big\}$$

*is called the space of $p$-integrable functions.*

Since $\|\cdot\|_{\mathcal{L}^p(\mu, \|\cdot\|_{\mathcal{B}})}$ is not definite, it is in particular not a norm. The following class of function spaces fixes this problem.

**Definition A.9 ($L^p$-space)**
*Let $\mathcal{B}$ be a Banach space, let $p \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space and let $\sim$ be the equivalence relation on $\mathcal{L}^p\left(\mu, \|\cdot\|_{\mathcal{B}}\right)$ satisfying for all $f, g \in \mathcal{L}^p\left(\mu, \|\cdot\|_{\mathcal{B}}\right)$ that*

$$f \sim g \quad :\Leftrightarrow \quad f = g \quad \mu\text{-a.s.}.$$

*Then the set*

$$L^p\left(\mu, \|\cdot\|_{\mathcal{B}}\right) := {}^{\mathcal{L}^p\left(\mu, \|\cdot\|_{\mathcal{B}}\right)}\!\big/\!_{\sim}$$

*is called the space of equivalence classes of p-integrable functions.*

Now $\|\cdot\|_{\mathcal{L}^p(\mu, \|\cdot\|_{\mathcal{B}})}$ is a norm on the space $L^p\left(\mu, \|\cdot\|_{\mathcal{B}}\right)$ and it in fact turns out that this forms a Banach space.

**Theorem A.10 (Bochner integral)**
*Let $\mathcal{B}$ be a Banach space, let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space. Then*

- *there exists a unique bounded linear function $I : \mathcal{L}^1\left(\mu, \|\cdot\|_{\mathcal{B}}\right) \to \mathcal{B}$ with the property that for all simple functions $f : \Omega \to \mathcal{B}$ it holds that*

$$I(f) = \sum_{x \in f(\Omega)} \mu\left(f^{-1}(\{x\})\right) x$$

- *and it holds for all $f \in \mathcal{L}^1\left(\mu, \|\cdot\|_{\mathcal{B}}\right)$ that $\|I(f)\|_{\mathcal{B}} \leq \|f\|_{\mathcal{L}^1(\mu, \|\cdot\|_{\mathcal{B}})}$.*

*We call the function $I$ the Bochner integral.*

Generally we use the standard integral notation to write Bochner integrals, i.e. for $f \in \mathcal{L}^1\left(\mu, \|\cdot\|_{\mathcal{B}}\right)$ we write

$$I(f) = \int_\Omega f(\omega)\, \mu(\mathrm{d}\omega).$$

**Proposition A.11 (properties of the Bochner integral)**
*Let $\mathcal{B}$ be a Banach space, let $\mathcal{H}$ be a Hilbert space and let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space, then the following properties of the Bochner integral hold*

*(i) for all $f \in \mathcal{L}^1\left(\mu, \|\cdot\|_{\mathcal{B}}\right)$ it holds that*

$$\left\| \int_\Omega f(\omega)\, \mu(d\omega) \right\|_{\mathcal{B}} \leq \int_\Omega \|f(\omega)\|_{\mathcal{B}}\, \mu(d\omega)$$

*(ii) for all $f \in \mathcal{L}^1\left(\mu, \|\cdot\|_{\mathcal{B}}\right)$ and for all $\varphi \in \mathcal{B}'$ that*

$$\varphi\left( \int_\Omega f(\omega)\, \mu(d\omega) \right) = \int_\Omega \varphi\left(f(\omega)\right)\, \mu(d\omega)$$

*(iii) for all $f \in \mathcal{L}^1\left(\mu, \|\cdot\|_{\mathcal{H}}\right)$ and for all $g \in \mathcal{H}$ that*

$$\left\langle \int_\Omega f(\omega)\, \mu(d\omega), g \right\rangle_{\mathcal{H}} = \int_\Omega \langle f(\omega), g \rangle_{\mathcal{H}}\, \mu(d\omega).$$

Appendix B

# Probability Theory

## B.1 Convergence in distribution

The first theorem shows that convergence in distribution also implies a convergence of the quantiles if the limit distribution is continuous (e.g. Lehmann and Romano, 2005, Lemma 11.2.1).

**Theorem B.1 (convergence of quantiles)**
*Let $(F_n)_{n\in\mathbb{N}}$ be a sequence of distribution functions on the real line converging weakly to a distribution function $F$. Assume $F$ is continuous and strictly increasing at $y = F^{-1}(1-\alpha)$. Then,*

$$\lim_{n\to\infty} F_n^{-1}(1-\alpha) = F^{-1}(1-\alpha).$$

*More generally, let $(\widehat{F}_n)_{n\in\mathbb{N}}$ be a sequence of random distribution functions satisfying $\widehat{F}_n(x) \xrightarrow{\mathbb{P}} F(x)$ as $n \to \infty$ at all $x$ which are continuity points of a fixed distribution function $F$. Assume $F$ is continuous and strictly increasing at $F^{-1}(1-\alpha)$. Then,*

$$\widehat{F}_n^{-1}(1-\alpha) \xrightarrow{\mathbb{P}} F^{-1}(1-\alpha)$$

*as $n \to \infty$.*

A further classical result is Slutsky's theorem (e.g. Lehmann and Romano, 2005, Theorem 11.2.11)

**Theorem B.2 (Slutsky's theorem)**
*Suppose $(X_n)_{n\in\mathbb{N}}$ is a sequence of real-valued random variables such that $X_n \xrightarrow{d} X$. Further, suppose $(A_n)_{n\in\mathbb{N}}$ and $(B_n)_{n\in\mathbb{N}}$ satisfy $A_n \xrightarrow{\mathbb{P}} a$ and $B_n \xrightarrow{\mathbb{P}} b$, where $a$ and $b$ are constants. Then,*

$$A_n X_n + B_n \xrightarrow{d} aX + b.$$

The next result is a corollary of Slutsky's theorem (e.g. Lehmann and Romano, 2005, Corollary 11.2.3).

**Corollary B.3**
*Suppose $(X_n)_{n\in\mathbb{N}}$ is a sequence of real-valued random variables such that $X_n \xrightarrow{d} X$, where $X$ has a continuous cumulative distribution function $F$. If $(C_n)_{n\in\mathbb{N}}$ is a sequence of real-valued random variables satisfying $C_n \xrightarrow{\mathbb{P}} c$, where $c$ is a constant, then*

$$\lim_{n\to\infty} \mathbb{P}\left(X_n \leq C_n\right) = F(c).$$

## B.2 Strong law of large numbers

The following theorem is a slight extension to the basic strong law of large numbers to allow for triangular schemes.

**Theorem B.4 (SLLN for triangular schemes)**
*Let $(X_{m,k})_{k\in\{1,\dots,m\}}$, $m \in \mathbb{N}$ be a triangular scheme of real-valued random variables satisfying*

*(i) for all $m \in \mathbb{N}$, $X_{m,1}, \dots, X_{m,m}$ are iid,*

*(ii) there exists $c \in \mathbb{R}$ such that $\lim_{m\to\infty} \mathbb{E}(X_{m,1}) = c$ and*

*(iii) there exists $K \in \mathbb{R}$ such that $\sup_{m\in\mathbb{N}} \mathbb{E}(X_{m,1}^2) < K$.*

*Then, it holds that*

$$\frac{1}{m}\sum_{k=1}^{m} X_{m,k} \xrightarrow{\mathbb{P}\text{-}a.s.} c$$

*as $m \to \infty$.*

**Proof** We can assume without loss of generality that $X_{m,k} \geq 0$ $\mathbb{P}$-a.s., otherwise we could simply consider the positive and negative part separately (i.e. $X_{m,k}^{+} := \max\{X_{m,k}, 0\}$ and $X_{m,k}^{-} := \max\{-X_{m,k}, 0\}$). Begin by setting $S_m := \sum_{k=1}^{m} X_{m,k}$. Then, by Chebyshev's inequality we get for all $\varepsilon > 0$ that

$$\mathbb{P}\left(\left|\frac{S_m}{m} - c\right| > \varepsilon\right) \leq \frac{\operatorname{Var}\left(\frac{S_m}{m}\right)}{\varepsilon^2} = \frac{\operatorname{Var}\left(S_m\right)}{m^2\varepsilon^2}. \tag{B.1}$$

Using (i) and (iii) we further get

$$\operatorname{Var}\left(S_m\right) = \sum_{k=1}^{m} \operatorname{Var}\left(X_{m,k}\right) \leq \sum_{k}^{m} \mathbb{E}\left(X_{m,k}^2\right) \leq mK. \tag{B.2}$$

Combining (B.1) and (B.2) this implies

$$\sum_{m=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{m^2}}{m^2} - c\right| > \varepsilon\right) \leq \sum_{m=1}^{\infty} \frac{K}{m^2\varepsilon^2} < \infty.$$

This however implies that

$$\frac{S_{m^2}}{m^2} \xrightarrow{\mathbb{P}\text{-a.s.}} c \tag{B.3}$$

as $m \to \infty$. Since we assumed $X_{m,k} \geq 0$ it holds for all $k \in \mathbb{N}$ that $S_k \leq S_{k+1}$ and hence for all $k \in \{m^2, \ldots, (m+1)^2\}$ it holds that

$$\frac{m^2}{(m+1)^2}\frac{S_{m^2}}{m^2} = \frac{S_{m^2}}{(m+1)^2} \leq \frac{S_k}{k} \leq \frac{S_{(m+1)^2}}{m^2} = \frac{S_{(m+1)^2}}{(m+1)^2}\frac{(m+1)^2}{m^2}.$$

Taking limits on both sides and using (B.3) completes the proof of Theorem B.4. □

The following theorem due to Beck and Giesy (1970, Theorem III.13) extends the strong law of large number to Banach space valued random variables. An overview of further options to extend the strong law is given by Beck et al. (1975).

**Theorem B.5 (Extension of SLLN)**
*Let $\mathcal{B}$ be an arbitrary Banach space and let $(X_k)_{k\in\mathbb{N}}$ be a sequence of independent $\mathcal{B}$-valued random variables such that for all $k \in \mathbb{N}$ it holds that $\mathbb{E}\left(X_k\right) = 0$. If either*

*(i)* $\sum_{k=1}^{\infty} \frac{\mathbb{E}\left(\|X_k\|_{\mathcal{B}}^2\right)}{k^2} < \infty$ *and* $\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left(\|X_k\|_{\mathcal{B}}^2\right) \to 0$ *as* $n \to \infty$ *or*

*(ii)* $\frac{1}{n}\sum_{k=1}^{n} \operatorname{esssup}\|X_k\|_{\mathcal{B}} \to 0$ *as* $n \to \infty$,

*then it holds that*

$$\left\|\frac{1}{n}\sum_{k=1}^{n} X_k\right\|_{\mathcal{B}} \overset{\mathbb{P}-a.s.}{\longrightarrow} 0$$

*as* $n \to \infty$.

The following Theorem is due to Mourier and can be found in Beck et al. (1975).

**Theorem B.6 (Extension of SLLN (i.i.d. setting))**
*Let $\mathcal{B}$ be a separable Banach space and let $(X_k)_{k\in\mathbb{N}}$ be a sequence of independent and identically distributed $\mathcal{B}$-valued random variables such that for all $k \in \mathbb{N}$ it holds that $\mathbb{E}\left(X_k\right) = 0$. Then it holds that*

$$\left\|\frac{1}{n}\sum_{k=1}^{n} X_k\right\|_{\mathcal{B}} \overset{\mathbb{P}-a.s.}{\longrightarrow} 0$$

*as* $n \to \infty$.

Appendix C

# Auxiliary results and proofs

## C.1 Notation

In order to make the calculations in this section more readable we use the following conventions.

- for all $j \in \{1, \ldots, d\}$ and for all $i_1, i_2 \in \{1, \ldots, m\}$ we set

$$k^j_{i_1,i_2} := k^j(X^j_{i_1}, X^j_{i_2})$$

- for all $q, n \in \mathbb{N}$, for all functions $g : \boldsymbol{\mathcal{X}}^n \to \mathbb{R}$ and for all $i_1, \ldots, i_q, j_1, \ldots, j_n \in \{1, \ldots, m\}$ we set

$$\mathbb{E}_{i_1,\ldots,i_q}\left(g(\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_n})\right) = \int_{\boldsymbol{\mathcal{X}}} \cdots \int_{\boldsymbol{\mathcal{X}}} g(\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_n})\, \mathbb{P}^{\mathbf{X}}(\mathrm{d}\mathbf{X}_{i_1}) \cdots \mathbb{P}^{\mathbf{X}}(\mathrm{d}\mathbf{X}_{i_q})$$

## C.2 Expansions of $h_1$ and $h_2$

**Lemma C.1 (expansion of $h_1$)**
*Assume Setting 3.1. Then it holds for all* $\mathbf{z} \in \boldsymbol{\mathcal{X}}$ *that,*

$$\begin{aligned}
h_1(\mathbf{z}) = \frac{1}{d}&\left[\mathbb{E}\left(\prod_{j=1}^{d} k^j(z^j, X^j_1)\right) - \mathbb{E}\left(\prod_{j=1}^{d} k^j(z^j, X^j_j)\right)\right] \\
&+ \frac{d-1}{d}\left[\mathbb{E}\left(\prod_{j=1}^{d} k^j(X^j_1, X^j_2)\right) - \mathbb{E}\left(\prod_{j=1}^{d} k^j(X^j_1, X^j_{j+1})\right)\right] \\
&+ \frac{1}{d}\left[\sum_{r=1}^{d} \mathbb{E}\left(\left(\prod_{j\neq r}^{d} k^j(X^j_{2j-1}, X^j_{2j})\right) k^r(z^r, X^r_{2r})\right)\right. \\
&\qquad \left. - \sum_{r=1}^{d} \mathbb{E}\left(\left(\prod_{j\neq r}^{d} k^j(X^j_1, X^j_{j+1})\right) k^r(z^r, X^r_{r+1})\right)\right]
\end{aligned}$$

**Proof** Recall that

$$h_1(\mathbf{z}) = \mathbb{E}\left(h(\mathbf{z}, \mathbf{X}_1, \dots, \mathbf{X}_{2d-1})\right).$$

Next we separate $h$ into 3 terms as follows.

$$\begin{aligned} h(\mathbf{z}_1, \dots, \mathbf{z}_{2d}) = {} & \frac{1}{(2d)!} \sum_{\pi \in S_{2d}} \left[ \prod_{j=1}^{d} k^j \left( z^j_{\pi(1)}, z^j_{\pi(2)} \right) \right] (=: b_1) \\ & + \frac{1}{(2d)!} \sum_{\pi \in S_{2d}} \left[ \prod_{j=1}^{d} k^j \left( z^j_{\pi(2j-1)}, z^j_{\pi(2j)} \right) \right] (=: b_2) \\ & - \frac{2}{(2d)!} \sum_{\pi \in S_{2d}} \left[ \prod_{j=1}^{d} k^j \left( z^j_{\pi(1)}, z^j_{\pi(j+1)} \right) \right] (=: b_3). \end{aligned}$$

Now we calculate $\mathbb{E}_{2,\dots,2d}\left(h(\mathbf{X}_1, \dots, \mathbf{X}_{2d})\right)$ by considering these three terms separately.

$\mathbf{b_1}$: Begin by letting $\pi \in S_{2d}$, then

$$\mathbb{E}_{2,\dots,2d}\left( \prod_{j=1}^{d} k^j_{\pi(1),\pi(2)} \right) = \begin{cases} \mathbb{E}_{2,3}\left( \prod_{j=1}^{d} k^j_{2,3} \right) & \text{if } \pi(1) \neq 1 \wedge \pi(2) \neq 1 \\ \mathbb{E}_{2}\left( \prod_{j=1}^{d} k^j_{1,2} \right) & \text{if } \pi(1) = 1 \vee \pi(2) = 1. \end{cases}$$

Counting how often each of these cases can occur for $\pi \in S_{2d}$ leads to

$$\begin{aligned} & \frac{1}{(2d)!} \sum_{\pi \in S_{2d}} \mathbb{E}_{2,\dots,2d}\left( \prod_{j=1}^{d} k^j_{\pi(1),\pi(2)} \right) \\ & = \frac{(2d-2)(2d-1)!}{(2d)!} \mathbb{E}_{2,3}\left( \prod_{j=1}^{d} k^j_{2,3} \right) + \frac{2(2d-1)!}{(2d)!} \mathbb{E}_{2}\left( \prod_{j=1}^{d} k^j_{1,2} \right) \\ & = \frac{d-1}{p} \mathbb{E}_{2,3}\left( \prod_{j=1}^{d} k^j_{2,3} \right) + \frac{1}{p} \mathbb{E}_{2}\left( \prod_{j=1}^{d} k^j_{1,2} \right) \qquad \text{(C.1)} \end{aligned}$$

$\mathbf{b_2}$: Begin by letting $\pi \in S_{2d}$, $r \in \{1, \dots, p\}$ such that $\pi(2r-1) = 1$ or $\pi(2r) = 1$ then

$$\mathbb{E}_{2,\dots,2d}\left( \prod_{j=1}^{d} k^j_{\pi(2j-1),\pi(2j)} \right) = \mathbb{E}_{2,\dots,2d+1}\left( \left( \prod_{j \neq r}^{d} k^j_{2j,2j+1} \right) k^r_{1,2r} \right)$$

Counting how many combinations are possible for each $r$ and adding all different combinations up gives us

$$\begin{aligned}
&\frac{1}{(2d)!}\sum_{\pi\in S_{2d}}\mathbb{E}_{2,\dots,2d}\left(\prod_{j=1}^{d}k^{j}_{\pi(2j-1),\pi(2j)}\right)\\
&\quad=\frac{2(2d-1)!}{(2d)!}\sum_{r=1}^{d}\mathbb{E}_{2,\dots,2d+1}\left(\left(\prod_{j\neq r}^{d}k^{j}_{2j,2j+1}\right)k^{r}_{1,2r}\right)\\
&\quad=\frac{1}{d}\sum_{r=1}^{d}\mathbb{E}_{2,\dots,2d+1}\left(\left(\prod_{j\neq r}^{d}k^{j}_{2j,2j+1}\right)k^{r}_{1,2r}\right) \qquad \text{(C.2)}
\end{aligned}$$

**b$_3$**: Begin by letting $\pi \in S_{2d}$, then

$$\mathbb{E}_{2,\dots,2d}\left(\prod_{j=1}^{d}k^{j}_{\pi(1),\pi(j+1)}\right)=\begin{cases}\mathbb{E}_{2,\dots,d+2}\left(\prod_{j=1}^{d}k^{j}_{2,j+2}\right) & \text{if } \pi(1)\neq 1\wedge\cdots\wedge\pi(d+1)\neq 1\\ \mathbb{E}_{2,\dots,d+1}\left(\prod_{j=1}^{d}k^{j}_{1,j+1}\right) & \text{if } \pi(1)=1\\ \mathbb{E}_{2,\dots,d+2}\left(\prod_{j\neq r}^{d}k^{j}_{2,j+2}k^{r}_{1,2}\right) & \text{if } \pi(r+1)=1 \text{ for } r\in\{1,\dots,d\}\end{cases}$$

Counting how often each of these cases can occur for different $\pi \in S_{2d}$ and adding all cases up results in

$$\begin{aligned}
&\frac{1}{(2d)!}\sum_{\pi\in S_{2d}}\mathbb{E}_{2,\dots,2d}\left(\prod_{j=1}^{d}k^{j}_{\pi(1),\pi(j+1)}\right)\\
&\quad=\frac{d-1}{2d}\mathbb{E}_{2,\dots,p+2}\left(\prod_{j=1}^{d}k^{j}_{2,j+2}\right)+\frac{1}{2d}\mathbb{E}_{2,\dots,p+1}\left(\prod_{j=1}^{d}k^{j}_{1,j+1}\right)\\
&\qquad+\frac{1}{2d}\sum_{r=1}^{d}\mathbb{E}_{2,\dots,p+2}\left(\prod_{j\neq r}^{d}k^{j}_{2,j+2}k^{r}_{1,2}\right) \qquad \text{(C.3)}
\end{aligned}$$

Finally combining (C.1), (C.1) and (C.1) completes the proof of Lemma C.1. □

**Lemma C.2 (expansion of $h_2$ under $H_0$)**
*Assume Setting 3.1. Then under $H_0$ it holds for all $\mathbf{z}_1, \mathbf{z}_2 \in \boldsymbol{\mathcal{X}}$ that,*

$$\begin{aligned}
\binom{2d}{2} h_2(\mathbf{z}_1, \mathbf{z}_2) = {} & \prod_{r=1}^{d} k^r(z_1^r, z_2^r) \quad (=: a_1) \\
& + (d-1)^2 \prod_{r=1}^{d} \mathbb{E}\left(k^r(X_1^r, X_2^r)\right) \quad (=: a_2) \\
& + (d-1) \prod_{r=1}^{d} \mathbb{E}\left(k^r(z_1^r, X_1^r)\right) \quad (=: a_3) \\
& + (d-1) \prod_{r=1}^{d} \mathbb{E}\left(k^r(z_2^r, X_1^r)\right) \quad (=: a_4) \\
& + \sum_{r=1}^{d} k^r(z_1^r, z_2^r) \prod_{l \neq r} \mathbb{E}\left(k^l(X_1^l, X_2^l)\right) \quad (=: a_5) \\
& - \sum_{r=1}^{d} k^r(z_1^r, z_2^r) \prod_{l \neq r} \mathbb{E}\left(k^l(z_1^l, X_1^l)\right) \quad (=: a_6) \\
& - \sum_{r=1}^{d} k^r(z_1^r, z_2^r) \prod_{l \neq r} \mathbb{E}\left(k^l(z_2^r, X_1^r)\right) \quad (=: a_7) \\
& + \sum_{r \neq s} \mathbb{E}\left(k^r(z_1^r, X_1^r)\right) \mathbb{E}\left(k^s(z_2^s, X_1^s)\right) \prod_{l \neq r,s} \mathbb{E}\left(k^l(X_1^l, X_2^l)\right) \quad (=: a_8) \\
& - (d-1) \sum_{r=1}^{d} \mathbb{E}\left(k^r(z_1^r, X_1^r)\right) \prod_{l \neq r} \mathbb{E}\left(k^l(X_1^l, X_2^l)\right) \quad (=: a_9) \\
& - (d-1) \sum_{r=1}^{d} \mathbb{E}\left(k^r(z_2^r, X_1^r)\right) \prod_{l \neq r} \mathbb{E}\left(k^l(X_1^l, X_2^l)\right) \quad (=: a_{10}).
\end{aligned}$$

**Proof** Begin by setting,

$$\begin{aligned}
A &:= \sum_{\pi \in S_{2d}} \mathbb{E}_{3,\dots,2d}\left(\prod_{j=1}^{d} k^j_{\pi(1),\pi(2)}\right) \\
B &:= \sum_{\pi \in S_{2d}} \mathbb{E}_{3,\dots,2d}\left(\prod_{j=1}^{d} k^j_{\pi(2j-1),\pi(2j)}\right) \\
C &:= \sum_{\pi \in S_{2d}} \mathbb{E}_{3,\dots,2d}\left(\prod_{j=1}^{d} k^j_{\pi(1),\pi(j+1)}\right).
\end{aligned}$$

Then it holds that,

$$h_2(\mathbf{X}_1, \mathbf{X}_2) = \mathbb{E}_{3,\dots,2d}\left(h(\mathbf{X}_1, \dots, \mathbf{X}_{2d})\right) = \frac{1}{(2d)!}\left(A + B - 2C\right). \tag{C.4}$$

Under the null hypothesis $H_0$ the terms $A$,$B$ and $C$ can be simplified using combinatorial arguments (similar to the ones used in the proof of Lemma C.1).

$$\begin{aligned}
A = {} & 2(2d-2)! \prod_{r=1}^{d} k_{1,2}^r \\
& + (2d-2)(2d-3)(2d-2)! \prod_{r=1}^{d} \mathbb{E}_{3,4}\left(k_{3,4}^r\right) \\
& + 2(2d-2)(2d-2)! \prod_{r=1}^{d} \mathbb{E}_{3}\left(k_{1,3}^r\right) \\
& + 2(2d-2)(2d-2)! \prod_{r=1}^{d} \mathbb{E}_{3}\left(k_{2,3}^r\right)
\end{aligned}$$

$$\begin{aligned}
B = {} & 2(2d-2)! \sum_{r=1}^{d} k_{1,2}^r \prod_{l \neq r} \mathbb{E}_{3,4}\left(k_{3,4}^l\right) \\
& + 4(2d-2)! \sum_{r \neq s} \mathbb{E}_{3}\left(k_{1,3}^r\right) \mathbb{E}_{3}\left(k_{2,3}^s\right) \prod_{l \neq r,s} \mathbb{E}_{3,4}\left(k_{3,4}^l\right)
\end{aligned}$$

$$\begin{aligned}
C = {} & 2(2d-2)! \sum_{r=1}^{d} k_{1,2}^r \prod_{l \neq r} \mathbb{E}_{3}\left(k_{1,3}^l\right) \\
& + (d-1)(2d-2)! \prod_{r=1}^{d} \mathbb{E}_{3}\left(k_{1,3}^r\right) + (d-1)(2d-2)! \prod_{r=1}^{d} \mathbb{E}_{3}\left(k_{2,3}^r\right) \\
& + (d-1)(d-2)(2d-2)! \prod_{r=1}^{d} \mathbb{E}_{3,4}\left(k_{3,4}^r\right) \\
& + (2d-2)! \sum_{r \neq s} \mathbb{E}_{3}\left(k_{1,3}^r\right) \mathbb{E}_{3}\left(k_{2,3}^s\right) \prod_{l \neq r,s} \mathbb{E}_{3,4}\left(k_{3,4}^l\right) \\
& + (d-1)(2d-2)! \sum_{r=1}^{d} \mathbb{E}_{3}\left(k_{1,3}^r\right) \prod_{l \neq r} \mathbb{E}_{3,4}\left(k_{3,4}^l\right) \\
& + (d-1)(2d-2)! \sum_{r=1}^{d} \mathbb{E}_{3}\left(k_{2,3}^r\right) \prod_{l \neq r} \mathbb{E}_{3,4}\left(k_{3,4}^l\right)
\end{aligned}$$

Plugging these expressions for $A$, $B$ and $C$ into (C.4) completes the proof of Lemma C.2. □

**Lemma C.3 (degeneracy under $H_0$)**
*Assume Setting 3.1. Then under $H_0$ it holds for all $\mathbf{z} \in \boldsymbol{\mathcal{X}}$ that*

$$h_1(\mathbf{z}) = 0,$$

*and therefore in particular that $\xi_1(h) = 0$.*

**Proof** Observe that under $H_0$ it holds for all $\mathbf{z} \in \boldsymbol{\mathcal{X}}$ that

- $$\mathbb{E}\left(\prod_{j=1}^{d} k^j(z^j, X_1^j)\right) = \mathbb{E}\left(\prod_{j=1}^{d} k^j(z^j, X_j^j)\right) \tag{C.5}$$
- $$\mathbb{E}\left(\prod_{j=1}^{d} k^j(X_1^j, X_2^j)\right) = \mathbb{E}\left(\prod_{j=1}^{d} k^j(X_1^j, X_{j+1}^j)\right) \tag{C.6}$$
- $$\begin{aligned}\mathbb{E}&\left(\left(\prod_{j\neq r}^{d} k^j(X_{2j-1}^j, X_{2j}^j)\right) k^r(z^r, X_{2r}^r)\right)\\ &= \mathbb{E}\left(\left(\prod_{j\neq r}^{d} k^j(X_1^j, X_{j+1}^j)\right) k^r(z^r, X_{r+1}^r)\right).\end{aligned} \tag{C.7}$$

Plugging (C.5), (C.6) and (C.7) into the explicit form of $h_1$ given in Lemma C.1 results in

$$h_1(\mathbf{z}) = 0.$$

This completes the proof of Lemma C.3. □

**Lemma C.4 (condition for non-degeneracy of $\xi_1(h)$ under $H_A$)**
*Assume Setting 3.1 and let $d = 2$. Then under $H_A$ it holds that*

$$\xi_1(h) > 0 \quad \Leftrightarrow \quad \left\|\Pi\left(\mathbb{P}^{\mathbf{X}}\right)\right\|_{\boldsymbol{\mathcal{H}}}^2 \neq \left\|\Pi\left(\mathbb{P}^{X^1} \otimes \mathbb{P}^{X^2}\right)\right\|_{\boldsymbol{\mathcal{H}}}^2.$$

**Proof** By Lemma C.1 it holds that

$$h_1(\mathbf{z}) = \frac{1}{2}\left[\mathbb{E}\left(k^1(z^1, X_1^1)k^2(z^2, X_1^2)\right) - \mathbb{E}\left(k^1(z^1, X_1^1)\right)\mathbb{E}\left(k^2(z^2, X_1^2)\right)\right] + c, \tag{C.8}$$

where

$$c = \frac{1}{2}\left[\mathbb{E}\left(k^1(X_1^1, X_2^1)k^2(X_1^2, X_2^2)\right) - \mathbb{E}\left(k^1(X_1^1, X_2^1)k^2(X_1^2, X_3^2)\right)\right]$$

is a constant. Jensen's inequality implies that

$$\xi_1(h) = \mathbb{E}\left([h_1(\mathbf{X}_1)]^2\right) - \theta_h^2 \geq \mathbb{E}\left(h_1(\mathbf{X}_1)\right)^2 - \theta_h^2 = 0$$

with equality if and only if $h_1(\mathbf{X}_1)$ is degenerate (i.e. $\mathbb{P}$-a.s. constant). Assume $h_1(\mathbf{X}_1)$ is degenerate, then by (C.8) it holds that

$$\mathbb{E}\left(k^1(X_1^1, X_2^1)k^2(X_1^2, X_2^2)\right) = \mathbb{E}\left(k^1(X_1^1, X_2^1)\right)\mathbb{E}\left(k^2(X_1^1, X_2^2)\right).$$

This can be written in terms of the mean embedding as follows

$$\left\|\Pi\left(\mathbb{P}^{\mathbf{X}}\right)\right\|_{\boldsymbol{\mathcal{H}}}^2 = \left\|\Pi\left(\mathbb{P}^{X^1}\otimes\mathbb{P}^{X^2}\right)\right\|_{\boldsymbol{\mathcal{H}}}^2. \qquad \square$$

**Example C.5 (counter example for non-degeneracy of $\boldsymbol{\xi_1(h)}$ under $\boldsymbol{H_A}$)** *Assume setting 3.1. Let $d = 3$, let $\boldsymbol{\mathcal{X}} = \mathbb{R}^3$, let $k^1(x, y) = k^2(x, y) = k^3(x, y) = e^{-\frac{(x-y)^2}{2}}$, let $X : \Omega \to \{0, 1\}$ be a random variable with $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$, let $Y = X$ and let $Z$ be an independent copy of $X$.*

*Define $\mathbf{X} \coloneqq (X, Y, Z)$, we will show that even though $X, Y, Z$ are not jointly independent it holds that $h_1(\mathbf{X})$ is a degenerate random variable. This in turn implies that $\xi_1(h) = 0$. By Lemma C.1 it holds that*

$$\begin{aligned}
h_1(\mathbf{x}) = &\tfrac{1}{3}\mathbb{E}\left(k(x, X_1)k(y, Y_1)k(z, Z_1)\right)\\
&- \tfrac{1}{3}\mathbb{E}\left(k(x, X_1)k(y, Y_2)k(z, Z_3)\right)\\
&+ \tfrac{1}{3}\mathbb{E}\left(k(Y_3, Y_4)k(Z_5, Z_6)k(x, X_2)\right)\\
&+ \tfrac{1}{3}\mathbb{E}\left(k(X_1, X_2)k(Z_1, Z_2)k(y, Y_4)\right)\\
&+ \tfrac{1}{3}\mathbb{E}\left(k(X_1, X_2)k(Y_3, Y_4)k(z, Z_6)\right)\\
&- \tfrac{1}{3}\mathbb{E}\left(k(Y_1, Y_3)k(Z_1, Z_4)k(x, X_2)\right)\\
&- \tfrac{1}{3}\mathbb{E}\left(k(X_1, X_2)k(Z_1, Z_4)k(y, Y_3)\right)\\
&- \tfrac{1}{3}\mathbb{E}\left(k(X_1, X_2)k(Y_1, Y_3)k(z, Z_4)\right) + c
\end{aligned}$$

*where*

$$c = \tfrac{2}{3}\mathbb{E}\left(k(X_1, X_2)k(Y_1, Y_2)k(Z_1, Z_2)\right) - \tfrac{2}{3}\mathbb{E}\left(k(X_1, X_2)k(Y_1, Y_3)k(Z_1, Z_4)\right)$$

*is a constant. Making use of independence and the fact that $Y = X$ leads to*

$$\begin{aligned}
h_1(\mathbf{x}) = &\tfrac{1}{3}\mathbb{E}\left(k(x, X_1)k(y, X_1)\right)\mathbb{E}\left(k(z, X_1)\right)\\
&- \tfrac{1}{3}\mathbb{E}\left(k(x, X_1)\right)\mathbb{E}\left(k(y, X_1)\right)\mathbb{E}\left(k(z, X_1)\right)\\
&+ \tfrac{1}{3}\mathbb{E}\left(k(X_1, X_2)\right)\mathbb{E}\left(k(X_1, X_2)\right)\mathbb{E}\left(k(z, X_1)\right)\\
&- \tfrac{1}{3}\mathbb{E}\left(k(X_1, X_2)k(X_1, X_3)\right)\mathbb{E}\left(k(z, X_1)\right) + c
\end{aligned}$$

*Due to the simplicity of the random variable $\mathbf{X}$ it is straight forward to calculate the expectations explicitly,*

- $\mathbb{E}\left(k(w, X_1)\right) = \frac{1}{2}e^{-\frac{w^2}{2}} + \frac{1}{2}e^{-\frac{(w-1)^2}{2}}$
- $\mathbb{E}\left(k(x, X_1)k(y, X_1)\right) = \frac{1}{2}e^{-\frac{x^2+y^2}{2}} + \frac{1}{2}e^{-\frac{(x-1)^2+(y-1)^2}{2}}$
- $\mathbb{E}\left(k(X_1, X_2)\right) = \frac{1}{2} + \frac{1}{2}e^{-\frac{1}{2}}$
- $\mathbb{E}\left(k(X_1, X_2)k(X_1, X_3)\right) = \frac{1}{4} + \frac{1}{2}e^{-\frac{1}{2}} + \frac{1}{4}e^{-1}$

*Using these expressions we can explicitly calculate $h_1$ as follows,*

$$\begin{aligned} h(\mathbf{x}) = \tfrac{1}{24}\Bigg[ & e^{-\frac{x^2+y^2+z^2}{2}} + e^{-\frac{(x-1)^2+(y-1)^2+(z-1)^2}{2}} \\ & + e^{-\frac{x^2+y^2+(z-1)^2}{2}} + e^{-\frac{(x-1)^2+(y-1)^2+z^2}{2}} \\ & - e^{-\frac{x^2+(y-1)^2+(z-1)^2}{2}} - e^{-\frac{(x-1)^2+y^2+(z-1)^2}{2}} \\ & - e^{-\frac{(x-1)^2+y^2+z^2}{2}} - e^{-\frac{x^2+(y-1)^2+z^2}{2}} \Bigg] + c \end{aligned}$$

*The support of* $\mathbb{P}^{\mathbf{X}}$ *is* $\operatorname{supp}(\mathbb{P}^{\mathbf{X}}) = \{(0,0,0),(0,0,1),(1,1,0),(1,1,1)\}$. *It can be checked directly that $h_1$ is constant on* $\operatorname{supp}(\mathbb{P}^{\mathbf{X}})$*. Hence we have shown that $h_1(\mathbf{X})$ is degenerate.*

# Bibliography

Beck, A. and D. P. Giesy (1970). P-uniform convergence and a vector-valued strong law of large numbers. *Transactions of the American Mathematical Society 147*(2), 541–559.

Beck, A., D. P. Giesy, and P. Warren (1975). Recent developments in the theory of strong laws of large numbers for vector-valued random variables. *Theory of Probability and its Applications 20*(1), 127–134.

Berlinet, A. and C. Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publishers.

Billingsley, P. (2008). *Convergence of Probability Measures.* John Wiley and Sons.

Blanchard, G., O. Bousquet, and L. Zwald (2007). Statistical properties of kernel principal component analysis. *Machine Learning 66*(2-3), 259–294.

Bühlmann, P., J. Peters, and J. Ernest (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics 42*(6), 2526–2556.

Da Prato, G. and J. Zabczyk (2014). *Stochastic Equations in Infinite Dimensions.* Cambridge University Press. Cambridge Books Online.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

Dunford, N. and J. T. Schwartz (1963). *Linear Operators, Part 2.* John Wiley and Sons.

Ferreira, J. and V. Menegatto (2009). Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory 64*(1), 61–81.

Gretton, A., O. Bousquet, A. Smola, and B. Schölkopf (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, pp. 63–77. Springer-Verlag.

Gretton, A., K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur (2009). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems (NIPS 22)*, pp. 673–681.

Gretton, A., K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NIPS 20)*, pp. 585–592.

Hoeffding, W. (1948a). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics 19*(3), 293–325.

Hoeffding, W. (1948b). A non-parametric test of independence. *The Annals of Mathematical Statistics 19*(4), 546–557.

Jentzen, A. (2015). Numerical analysis of stochastic partial differential equations. Lecture notes, ETH Zürich.

Klenke, A. (2014). *Probability Theory.* Springer-Verlag.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation.* Springer-Verlag.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses.* Springer-Verlag.

Leucht, A. and M. H. Neumann (2009). Consistency of general bootstrap methods for degenerate U-type and V-type statistics. *Journal of Multivariate Analysis 100*(8), 1622–1633.

Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research, to appear. ArXiv e-prints (1412.3773).*

Peters, J. (2008). Asymmetries of time series under inverting their direction. Master's thesis, Ruprecht-Karls-Universität Heidelberg, Germany.

Peters, J., J. M. Mooij, D. Janzing, and B. Schölkopf (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research 15*(1), 2009–2053.

Phipson, B. and G. K. Smyth (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology 9*(1), 1–16.

Prévôt, C. and M. Röckner (2007). *A Concise Course on Stochastic Partial Differential Equations.* Springer-Verlag.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 2*(6), 110–114.

Sejdinovic, D., A. Gretton, and W. Bergsma (2013). A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems (NIPS 26)*, pp. 1124–1132.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley and Sons.

Smola, A., A. Gretton, L. Song, and B. Schölkopf (2007). A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, Volume 4754 of *Lecture Notes in Computer Science*, pp. 13–31. Springer-Verlag.

Sriperumbudur, B. K., A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf (2008). Injective Hilbert space embeddings of probability measures. In *Conference on Learning Theory (COLT 21)*, pp. 111–122.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press. Cambridge Books Online.

Weidmann, J. (1980). *Linear Operators in Hilbert Spaces.* Springer-Verlag.

Werner, D. (2011). *Funktionalanalysis.* Springer-Verlag.

Wood, S. N. and N. H. Augustin (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling 157*(2–3), 157–177.