These appendices provide additional background and elaborate on some of the finer points in the main text. In Appendix A we illustrate that cumulants have typically lower variance estimators compared to moments. Technical background on tensor products and tensor sums of Hilbert spaces, and on tensor algebras is provided in Appendix B. We present our proofs in Appendix C. In Appendix D additional details on our numerical experiments are provided. Our V-statistic based estimators are detailed in Appendix E.

A Moments and Cumulants

Already for real-valued random variable X, moments have well-known drawbacks that make cumulants often preferable as statistics. For a detailed introduction to the use of cumulants in statistics we refer to McCullagh (2018). Here we just mention that

1. the moment generating function $f(t) = \mathbb{E}[e^{tX}] = \sum_m \mu_m t^m/m!$ describes the law of X with sequence (μ_m) of moments $\mu_m = \mathbb{E}[X^m] \in \mathbb{R}$. However, since the function $t \mapsto f(t)$ is the expectation of an exponential, one would often expect that f is also "exponential in t", hence $g(t) = \log f(t) = \sum_m \kappa_m \frac{t^m}{m!}$ should be simpler to describe as a power series. For example, for a Gaussian $f(t) = e^{t\mathbb{E}(X)+\frac{t^2}{2}Var(X)}$ and while μ_m can be in this case explicitly calculated and uneven moments vanish, the m-moments are fairly complicated compared to the power series expansion of $g(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2}$ which just consists of κ_1 (mean) and κ_2 (variance).
2. In the moment sequence μ_m, lower moments can dominate higher moments. Hence, a natural idea to compensate for these "different scales" is to systematically subtract lower moments from higher moments. As mentioned in the introduction, this is in particular troublesome if finite samples are available. Even in dimension $d = 1$ the second moment is dominated by the squared mean, that is for a real-valued random variable $X \sim \gamma$
$$\mu^2(\gamma) = (\mu^1(\gamma))^2 + \mathrm{Var}(X),$$
where $\mathrm{Var}(X) := \mathbb{E}[(X - \mu^1(\gamma))^2]$. It is well known that the minimum variance unbiased estimators for the variance are more efficient than that for the second moment: denoting them by $\widehat{\mu^2}$ and $\widehat{\kappa}$ respectively, one can show (Bonnier & Oberhauser, 2020) that given N samples from X, the following holds
$$\mathrm{Var}\left(\widehat{\mu^2}\right) = \mathrm{Var}\left(\widehat{\kappa}\right) + \frac{2}{N}\left[(\mathbb{E}X)^4 - (\mathbb{E}X)^2\,\mathrm{Var}(X) - 2\frac{\mathrm{Var}(X)^2}{N-1}\right].$$
This means that when X has a large mean, it is more efficient to estimate its variance than its second moment since the last term in the above expression dominates. Hence, the variance $\mathrm{Var}(X)$ is typically a much more sensible second-order statistic than $\mu^2(\gamma)$. However, we emphasize that there are many other reasons why cumulants can have better properties as estimators
3. Cumulants characterize laws and the independence of two random variables manifests itself simply as vanishing of cross-cumulants. In view of the above item 2, this means for example that testing independence can be preferable in terms of vanishing cumulants rather than testing if moments factor $\mathbb{E}[X^m Y^n] = \mathbb{E}[X^m]\mathbb{E}[X^n]$, and similarly for testing if distributions are the same.

The caveat to the above points is that it is not true that cumulants are always preferable. For example, there are distributions for which (a) the moment generating function is not naturally exponential in t, (b) lower moments do not dominate higher moments, (c) consequently independence or two-sample testing become worse with cumulants. While one can write down conditions under which for example, the variance of the kernelized cumulants is lower, the use of cumulants among statisticians is to simply regard cumulants as arising from natural motivations which leads to another estimator in their toolbox.

The main idea of our paper is simply that for the same reasons that cumulants can turn out to be powerful for real or vector-valued random variables, cumulants of RKHS-valued random variables are a natural choice of statistics. The situation is more complicated since it requires formalizing moment-and cumulant-generating functions in RKHS but ultimately a kernel trick allows for circumventing the computational bottleneck of working in infinite dimensions and leads to computable estimators for independence and two-sample testing.

Further, we note that although cumulants are classic for vector-valued data, there seems to be not much work done about extending their properties to general structured data. Our kernelized cumulants apply to any set $\mathcal{X}$ where a kernel is given. This includes many practically relevant examples such as strings (Lodhi et al., 2002), graphs (Kriege et al., 2020), or general sequentially ordered data (Király & Oberhauser, 2019; Chevyrev & Oberhauser, 2022); a survey of kernels for structured data is provided by Gärtner (2003).

B Technical Background

In Section B.1 the tensor products ($\bigotimes_{j=1}^d \mathcal{H}_j$) and direct sums of Hilbert spaces ($\bigoplus_{i\in I} \mathcal{H}_i$) are recalled. Section B.2 is about tensor algebras over Hilbert spaces ($\prod_{m\geq 0} \mathcal{H}^{\otimes m}$).

B.1 Tensor Products and Direct Sums of Banach and Hilbert Spaces

Tensor products of Hilbert spaces. For Hilbert spaces $\mathcal{H}, \dots, \mathcal{H}_d$ and $(h_1, \dots, h_d) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$, the multi-linear operator $h_1 \otimes \cdots \otimes h_d \in \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d$ is defined as

$$(h_1 \otimes \cdots \otimes h_d)(f_1, \dots, f_d) = \prod_{j=1}^d \langle h_j, f_j \rangle_{\mathcal{H}_j}$$

for all $(f_1, \dots, f_d) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$. By extending the inner product

$$\langle a_1 \otimes \cdots \otimes a_d, b_1 \otimes \cdots \otimes b_d \rangle_{\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d} := \prod_{j=1}^d \langle a_j, b_j \rangle_{\mathcal{H}_j}$$

to finite linear combinations of $a_1 \otimes \cdots \otimes a_d$-s

$$\left\{ \sum_{i=1}^n c_i \otimes_{j=1}^d a_{i,j} \, : \, c_i \in \mathbb{R}, a_{i,j} \in \mathcal{H}_j, \, n \geq 1 \right\}$$

by linearity, and taking the topological completion one arrives at $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d$. Specifically, if $(\mathcal{H}_1, k_1), \dots, (\mathcal{H}_d, k_d)$ are RKHSs, then so is $\mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_d = \mathcal{H}_{\otimes_{j=1}^d k_j}$ (Berlinet & Thomas-Agnan, 2004, Theorem 13) with the tensor product kernel

$$\left(\otimes_{j=1}^d k_j \right) \left((x_1, \dots, x_d), (x'_1, \dots, x'_d) \right) := \prod_{j=1}^d k_j \left(x_j, x'_j \right)$$

where $(x_1, \dots, x_d), (x'_1, \dots, x'_d) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$.

Tensor products of Banach spaces. For Banach spaces $\mathcal{B}_1, \dots \mathcal{B}_d$, the construction of $\mathcal{B}_1 \otimes \cdots \otimes \mathcal{B}_d$ is a little more involved (Lang, 2002) as one cannot rely on an inner product.

Direct sums of Hilbert and Banach spaces. Let $(\mathcal{H}_i)_{i\in I}$ be Hilbert or Banach spaces where I is some index set. The direct sum of $\mathcal{H}_i$-s— written as $\bigoplus_{i\in I} \mathcal{H}_i$—consists of ordered tuples $h = (h_i)_{i\in I}$ such that $h_i \in \mathcal{H}_i$ for all $i \in I$ and $h_i = 0$ for all but a finite number of $i \in I$. Operations (addition, scalar multiplication) are performed coordinate-wise, and the inner product of $a, b \in \bigoplus_{i\in I} \mathcal{H}_i$ is defined as $\langle a, b \rangle_{\bigoplus_{i\in I} \mathcal{H}_i} = \sum_{i\in I} a_i b_i$.

B.2 Tensor Algebras

The tensor algebra T_j over a Hilbert space $\mathcal{H}_j$ is defined as the topological completion of the space

$$\bigoplus_{m\geq 0} \mathcal{H}_j^{\otimes m}.$$

Note that it can equivalently be defined as the subset of $(h_0, h_1, h_2, \ldots) \in \prod_{m\geq 0} \mathcal{H}_j^{\otimes m}$ such that $\sum_{m\geq 0} \|h_m\|^2_{\mathcal{H}_j^{\otimes m}} < \infty$, and as such it is a Hilbert space with norm

$$\|(h_0, h_1, h_2, \ldots)\|^2_{\prod_{m\geq 0} \mathcal{H}_j^{\otimes m}} = \sum_{m\geq 0} \|h_m\|^2_{\mathcal{H}_j^{\otimes m}}.$$

T_j is also an algebra, endowed with the tensor product over $\mathcal{H}_j$ as its product. For $a = (a_0, a_1, a_2, a_2 \ldots), b = (b_0, b_1, b_2, b_2 \ldots) \in \mathrm{T}_j$, their product can be written down in coordinates as

$$a \cdot b = \left(\sum_{i=0}^{m} a_i \otimes b_{m-i} \right)_{m\geq 0}.$$

For a sequence $\mathcal{H}_1, \ldots, \mathcal{H}_d$ of Hilbert spaces, we define

$$\mathrm{T} \coloneqq \mathrm{T}_1 \otimes \cdots \otimes \mathrm{T}_d,$$

where $\mathrm{T}_j = \prod_{m\geq 0} \mathcal{H}_j^{\otimes m}$ $(j = 1, \ldots, d)$. Let $\mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_d$, and recall that given a tuple of integers $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$ we define $\mathcal{H}^{\otimes \mathbf{i}} := \mathcal{H}_1^{\otimes i_1} \otimes \cdots \otimes \mathcal{H}_d^{\otimes i_d}$. This allows us to write down a multi-grading for T as

$$\mathrm{T} = \prod_{\mathbf{i}\in\mathbb{N}^d} \mathcal{H}^{\otimes \mathbf{i}}. \tag{8}$$

Note that this gives credence to us using multi-indices $\mathbf{i} \in \mathbb{N}^d$ to describe elements of the tensor algebra, as the multi-indices form its multi-grading.

Furthermore, T is a multi-graded algebra when endowed with the (linear extension of the) following multiplication defined on the components of T

$$\begin{aligned} &\star : \mathcal{H}^{\otimes \mathbf{i}^1} \times \mathcal{H}^{\otimes \mathbf{i}^2} \to \mathcal{H}^{\otimes (\mathbf{i}^1 + \mathbf{i}^2)}, \\ &(x_1 \otimes \cdots \otimes x_d) \star (y_1 \otimes \cdots \otimes y_d) = (x_1 \cdot y_1) \otimes \cdots \otimes (x_d \cdot y_d), \end{aligned} \tag{9}$$

so that for $a = \left(a^{\mathbf{i}}\right)_{\mathbf{i}\in\mathbb{N}^d}, b = \left(b^{\mathbf{i}}\right)_{\mathbf{i}\in\mathbb{N}^d} \in \mathrm{T}$, their product can be written down as

$$(a \star b)^{\mathbf{i}} = \sum_{\mathbf{i}^1 + \mathbf{i}^2 = \mathbf{i}} a^{\mathbf{i}^1} \star b^{\mathbf{i}^2} \tag{10}$$

where addition of tuples $\mathbf{i}^1, \mathbf{i}^2 \in \mathbb{N}^d$ is defined as $\mathbf{i}^1 + \mathbf{i}^2 = \left(i_1^1 + i_1^2, \ldots, i_d^1 + i_d^2\right)$. With the degree of a tuple defined as $\deg(\mathbf{i}) = i_1 + \cdots + i_d$, T is also a graded algebra, with the grading written down as

$$\mathrm{T} = \prod_{m\geq 0} \bigoplus_{\{\mathbf{i}\in\mathbb{N}^d : \deg(\mathbf{i}) = m\}} \mathcal{H}^{\otimes \mathbf{i}},$$

so that if one multiplies two elements together, the degree of their product is the sum of their degree.

Finally we note that T is a unital algebra and the unit has the explicit form

$$(1, 0, 0, \ldots),$$

i.e. the element consisting of only a 1 at degree 0.

C Proofs

This section is dedicated to proofs. The equivalence between the combinatorial expressions of cumulants and the definition via a moment generating function is proved in Section C.2. The derivation of our main results (Theorem 2 and Theorem 3) are detailed in Section C.3.

C.1 Equivalent Definitions of Cumulants in $\mathbb{R}^d$

Here we introduce a classical definition of cumulants via a moment generating function and its equivalence to the combinatorial expressions. If $X = (X_1, \dots, X_d)$ is an $\mathbb{R}^d$-valued random variable distributed according to $X \sim \gamma$, then

$$\mu^{\mathbf{i}} = \mathbb{E}[X_1^{i_1} \cdots X_d^{i_d}] \in \mathbb{R}$$

for $\mathbf{i} = (i_1, \dots, i_d) \in \mathbb{N}^d$. The following definition of the cumulants $\kappa^{\mathbf{i}}(\gamma)$ of γ are equivalent

1. $\sum_{\mathbf{i} \in \mathbb{N}^d} \kappa^{\mathbf{i}}(\gamma) \frac{\boldsymbol{\theta}^{\mathbf{i}}}{\mathbf{i}!} = \log \sum_{\mathbf{i} \in \mathbb{N}^d} \mu^{\mathbf{i}}(\gamma) \frac{\boldsymbol{\theta}^{\mathbf{i}}}{\mathbf{i}!}$,
2. $\kappa^{\mathbf{i}}(\gamma) = \sum_{\pi \in P(d)} c_\pi \prod_{\sigma \in \pi} \mu^{\mathbf{i}}(\sigma)$,

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$, $c_\pi = (-1)^{|\pi|}(|\pi| - 1)!$ and the product $\prod_{\sigma \in \pi}$ is over all the blocks $\sigma \in (\pi_1, \dots, \pi_b)$ in the partition $\pi = (\pi_1, \dots, \pi_b)$ of $\{1, \dots, d\}$. The equivalence between these two definitions of cumulants, via a generating function and via their combinatorial definition, is classic (McCullagh, 2018). This equivalence is also at the heart of many proofs about properties of cumulants since some properties are easier to prove via one or the other definition.

C.2 Equivalent Definitions of Cumulants in RKHS

In the main text, we defined cumulants in RKHS by mimicking the combinatorial definition of cumulants in $\mathbb{R}^d$. It is natural and useful to also have the analogous definition via a "generating function" for RKHS-valued random variables. However, to generalize the definition via the logarithm of the moment generating function to random variables in RKHS, requires to define a logarithm for tensor series of moments. In this part, we show that this can be done and that indeed the two definitions are equivalent.

We use the shorthand $\kappa(\gamma) := \kappa_{k_1,\dots,k_d}(\gamma)$, $\mu(\gamma) := \mu_{k_1,\dots,k_d}(\gamma)$, and we overload the notation $(X_1, \dots, X_d)$ with $(k_1(\cdot, X_1), \dots, k_d(\cdot, X_d))$. With this notation, we show that given coordinates $\mathbf{i} \in \mathbb{N}^d$, one may express the generalized cumulant $\kappa^{\mathbf{i}}(\gamma)$ as either a combinatorial sum over moments indexed by partitions, or by using the cumulant generating function.

More specifically, we show that the generalized cumulant of a probability measure γ on $\mathcal{H}_1 \times \cdots \times \mathcal{H}_d$ defined as

$$\kappa^{\mathbf{i}}(\gamma) = \sum_{\pi \in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}}(X^{\otimes \mathbf{i}})$$

where $c_\pi = (-1)^{|\pi|-1}(|\pi| - 1)!$ can also be expressed as coordinates in the tensorized logarithm of the moment series. Motivated by the Taylor series expansion of the classic logarithm, we define

$$\log : \mathrm{T} \to \mathrm{T}, \quad x \mapsto \sum_{n \geq 1} \frac{(-1)^{n-1}}{n} (x - 1)^{\star n},$$

where $\star$ denotes the product as defined in (9) and for $t \in \mathrm{T}$, $t^{\star n}$ is defined as

$$t^{\star n} = \underbrace{t \star \cdots \star t}_{\text{n - times}},$$

or coordinate-wise $(t^{\star n})^{\mathbf{i}} = \sum_{\mathbf{i}^1 + \cdots + \mathbf{i}^n = \mathbf{i}} t^{\mathbf{i}^1} \star \cdots \star t^{\mathbf{i}^n}$ for $\mathbf{i} \in \mathbb{N}^d$. Note that unlike the classical logarithm $\log : \mathbb{R}_+ \to \mathbb{R}$, the tensorized logarithm is defined on the whole space as a formal expression.

Generalized Cumulants as Logarithms We want to show that the following holds

$$\kappa^{\mathbf{i}}(\gamma) = \big(\log \mu(\gamma^{\mathbf{i}})\big)^{\mathbf{1}_m}, \tag{11}$$

where $\mathbf{1}_m = (1, \dots, 1) \in \mathbb{N}^m$. By iterating (10) we can express (11) as

$$\sum_{j=1}^{m} \frac{(-1)^{j-1}}{j} \sum_{\mathbf{i}^1 + \cdots + \mathbf{i}^j = \mathbf{1}_m} \mu^{\mathbf{i}^1}(\gamma^{\mathbf{i}}) \star \cdots \star \mu^{\mathbf{i}^j}(\gamma^{\mathbf{i}}),$$

and our goal is to express this as a sum over partitions. We will use the notation $[n] = \{1, \dots, n\}$. We can achieve our goal in two parts:

1. Show that for a fixed $\mathbf{i} \in \mathbb{N}^d$ with $\deg(\mathbf{i}) = m$ we can express (11) as a sum over all surjective functions from $[m]$ to $[j]$.
2. Show that this sum over functions reduces to a sum over partitions.

Part 1. Note that given $\mathbf{i}^1 + \cdots + \mathbf{i}^j = \mathbf{1}_m$ we may define $h : [m] \to [j]$ by the relation $(\mathbf{i}^{h(n)})_n = 1$, that is, we take $h(n)$ to be the index c for which the multi-index $\mathbf{i}^c$ is 1 at n. Note that this function is necessarily surjective since the sum is taken over non-zero multi-indices. Equivalently, for any surjective function $h : [m] \to [j]$ we may define multi-indices by setting

$$(\mathbf{i}^c)_n = \begin{cases} 1 & \text{if } n \in h^{-1}(c) \\ 0 & \text{otherwise} \end{cases}.$$

Note that any such multi-index will be non-zero since the function is assumed to be surjective. With this identification we can write

$$\big(\log \mu(\gamma^{\mathbf{i}})\big)^{\mathbf{1}_m} = \sum_{j=1}^{m} \frac{(-1)^{j-1}}{j} \sum_{h:[m]\to[j]} \mu^{\mathbf{i}^{h^{-1}(1)}}(\gamma^{\mathbf{i}}) \star \cdots \star \mu^{\mathbf{i}^{h^{-1}(j)}}(\gamma^{\mathbf{i}}).$$

Part 2. Recall that given a function $h : [m] \to [j]$ we can associate it to its corresponding partition $\pi_h \in \mathcal{P}(m)$ by considering the set $\{h^{-1}(1), \ldots, h^{-1}(j)\}$, and there are exactly $j!$ different functions corresponding to a given partition, which are given by re-ordering the values $1, \ldots, j$. This reordering of the blocks does not change the summands since the marginals of the partition measure are always copies of each other and hence self-commute, hence a product of moments like $\mu^{\mathbf{i}^{h^{-1}(1)}}(\gamma^{\mathbf{i}}) \star \cdots \star \mu^{\mathbf{i}^{h^{-1}(j)}}(\gamma^{\mathbf{i}})$ can always be written as $\mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi_h})$, the $\mathbf{i}$-th coordinate of the moment sequence of the partition measure $\gamma^{\mathbf{i}}_{\pi_h}$. With this in mind we can write

$$\begin{aligned}\big(\log \mu(\gamma^{\mathbf{i}})\big)^{\mathbf{1}_m} &= \sum_{j=1}^{m} \frac{(-1)^{j-1}}{j} \sum_{h:[m]\to[j]} \mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi_h}) = \sum_{\pi\in P(m)} \frac{(-1)^{|\pi|-1}}{|\pi|} |\pi|! \mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi}) \\ &= \sum_{\pi\in P(m)} c_\pi \mu^{\mathbf{i}}(\gamma^{\mathbf{i}}_{\pi}) = \sum_{\pi\in P(m)} c_\pi \mathbb{E}_{\gamma^{\mathbf{i}}_{\pi}}(X^{\otimes \mathbf{i}}).\end{aligned}$$

From this it immediately follows that for two probability measures γ, η we can write

$$\begin{aligned}\langle \kappa^{\mathbf{i}}(\gamma), \kappa^{\mathbf{i}}(\eta)\rangle_{\mathcal{H}^{\otimes \mathbf{i}}} &= \langle \sum_{\pi\in P(m)} c_\pi \mathbb{E}_{\gamma^{\mathbf{i}}_{\pi}}(X^{\otimes \mathbf{i}}), \sum_{\tau\in P(m)} c_\tau \mathbb{E}_{\eta^{\mathbf{i}}_{\tau}}(Y^{\otimes \mathbf{i}})\rangle_{\mathcal{H}^{\otimes \mathbf{i}}} \\ &= \sum_{\pi,\tau\in P(m)} c_\pi c_\tau \mathbb{E}_{(X,Y)\sim \gamma^{\mathbf{i}}_{\pi}\otimes\eta^{\mathbf{i}}_{\tau}} \langle X^{\otimes \mathbf{i}}, Y^{\otimes \mathbf{i}}\rangle_{\mathcal{H}^{\otimes \mathbf{i}}}.\end{aligned}$$

Lemma 1 then follows from the definition of the tensor products.

C.3 Proof of Theorem 2 and Theorem 3

In this section we present the proofs of Theorem 2 and Theorem 3. We do this in a slightly more abstract setting where the feature maps take values in Banach spaces for clarity, until the end when we again restrict our attention to RKHSs. We start out by showing that polynomial functions of the feature maps characterize measures (Lemma 4). From there it is straightforward to show that cumulants have the same property (Theorem 4), and lastly that this also holds when working directly with the kernels (Proposition 1).

A *monomial* on separable Banach spaces $\mathcal{B}_1, \ldots, \mathcal{B}_d$ is any expression of the form

$$M(x_1, \ldots, x_d) = \prod_{j=1}^{i_1} \langle f^1_j, x_1\rangle \cdots \prod_{j=1}^{i_d} \langle f^d_j, x_d\rangle$$

for some $(i_1, \ldots, i_d) \in \mathbb{N}^d$, where $f^i_j \in \mathcal{B}^\star_i$ are elements of the dual space $\mathcal{B}^\star_i$ and $x_i \in \mathcal{B}_i$.[5] Finite linear combinations of monomials are called the *polynomials*. Recall that a set of functions F on a set S is said to *separate the points* of S if for every $x \neq y \in S$ there exists $f \in F$ such that $f(x) \neq f(y)$.

[5]These monomials naturally extend the classical ones.

Lemma 4 (Polynomial functions of feature maps characterize probability measures)**.** *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be Polish spaces, $\mathcal{B}_1 \ldots, \mathcal{B}_d$ separable Banach spaces and $\varphi_i : \mathcal{X}_i \to \mathcal{B}_i$ be continuous, bounded, and injective functions. Then the set of functions on the Borel probability measures $\mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right)$ of $\prod_{i=1}^d \mathcal{X}_i$*

$$\mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right) \to \mathbb{R}, \quad \gamma \mapsto \int_{\prod_{i=1}^d \mathcal{X}_i} p\big(\varphi_1(x_1), \ldots, \varphi_d(x_d)\big) \mathrm{d}\gamma(x_1, \ldots, x_d),$$

where p ranges over all polynomials, separates the points of $\mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right)$.

Proof. We first show that the pushforward map

$$\prod_{i=1}^d \varphi_i : \mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right) \to \mathcal{P}\left(\prod_{i=1}^d \mathcal{B}_i\right)$$

is injective. This is done in two parts, first we show that every Borel measure on $\prod_{i=1}^d \mathcal{X}_i$ is a Radon measure, then we show that the pushforward map is injective on Radon measures. To see the first part, note that since $\mathcal{X}_1, \ldots, \mathcal{X}_d$ are Polish spaces, so is their product space $\prod_{i=1}^d \mathcal{X}_i$ (Dudley 2004, Theorem 2.5.7; Willard 1970, Theorem 16.4c), and since Borel measures on Polish spaces are Radon measures (Bogachev, 2007, Theorem 7.1.7), any $\gamma \in \mathcal{P}(\prod_{i=1}^d \mathcal{X}_i)$ must be a Radon measure.

For the second part, note that

$$\prod_{i=1}^d \varphi_i : \prod_{i=1}^d \mathcal{X}_i \to \prod_{i=1}^d \mathcal{B}_i, \quad \left(\prod_{i=1}^d \varphi_i\right)(x_1, \ldots, x_d) \mapsto \prod_{i=1}^d \varphi_i(x_i)$$

is a norm bounded, continuous injection. Since $\prod_{i=1}^d \mathcal{B}_i$ is a Hausdorff space, $\prod_{i=1}^d \varphi_i$ is a homeomorphism on compacts since continuous injections into Hausdorff spaces are homeomorphisms on compacts (Rudin, 1953, Theorem 4.17). Let $\mu, \nu \in \mathcal{P}\left(\prod_{i=1}^d \mathcal{X}_i\right)$ be two Radon measures such that their pushforwards are the same $\prod_{i=1}^d \varphi_i(\mu) = \prod_{i=1}^d \varphi_i(\nu)$, then for any compact $C \subseteq \prod_{i=1}^d \mathcal{X}_i$ we have $\mu(C) = \nu(C)$ as $\prod_{i=1}^d \varphi_i : C \to \prod_{i=1}^d \varphi_i(C)$ is a homeomorphism. Since Radon measures are characterized by their values on compacts, this implies that $\mu = \nu$. Hence the pushforward map is injective.

Denote by K the image of $\prod_{i=1}^d \mathcal{X}_i$ under the mapping $\prod_{i=1}^d \varphi_i$ in $\prod_{i=1}^d \mathcal{B}_i$. Note that K is a bounded Polish space. It is enough to show that the polynomials separate the points of $\mathcal{P}(K)$. To see this, note that the polynomials form an algebra of continuous functions that separate the points of $\prod_{i=1}^d \mathcal{B}_i$, and when restricted to K they are bounded, since K is norm bounded. Since K is Polish, any Borel measure is Radon, and we can apply the Stone-Weierstrass theorem for Radon measures (Bogachev, 2007, Exercise 7.14.79) to get the assertion. □

In what follows we will use the following index notation for linear functionals. Fix some tuple $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$ with $\deg(\mathbf{i}) = m$. Given separable Banach spaces $\mathcal{B}_1 \ldots, \mathcal{B}_d$ we use the notation

$$\mathcal{B}^{\otimes \mathbf{i}} := \mathcal{B}_1^{\otimes i_1} \otimes \cdots \otimes \mathcal{B}_d^{\otimes i_d}$$

and given an element $x = (x_1, \ldots, x_d) \in \prod_{i=1}^d \mathcal{B}_i$ we write $x^{\mathbf{i}} := x_1^{\otimes i_1} \otimes \cdots \otimes x_d^{\otimes i_d}$ so that $x^{\mathbf{i}} \in \mathcal{B}^{\otimes \mathbf{i}}$. If we have functions $(\varphi_i)_{i=1}^d$ such that $\varphi_i : \mathcal{X}_i \to \mathcal{B}_i$ on some Polish spaces $\mathcal{X}_1, \ldots, \mathcal{X}_d$, then we write

$$\varphi^{\otimes \mathbf{i}} := \varphi_1^{\otimes i_1} \otimes \cdots \otimes \varphi_d^{\otimes i_d}, \quad \varphi^{\otimes \mathbf{i}} : \prod_{i=1}^d \mathcal{X}_i \to \mathcal{B}^{\otimes \mathbf{i}}.$$

Given a collection of linear functionals $F \in \prod_{j=1}^d \left(\mathcal{B}_j^\star\right)^{i_j}$ such that $F = (f_1, \ldots, f_d)$ we write

$$F^{\otimes \mathbf{i}} := f_1 \otimes \cdots \otimes f_d, \quad F^{\otimes \mathbf{i}} \in \left(\mathcal{B}^{\otimes \mathbf{i}}\right)^\star.$$

Note the following trick: the monomials on $\prod_{i=1}^d \mathcal{B}_i$ are exactly functions of the form

$$x \mapsto \langle F^{\otimes \mathbf{i}}, x^{\mathbf{i}} \rangle$$

for $F = (f_1, \ldots, f_d)$, this will be used in the proofs. We can now restate and prove the our theorem. Note that the cumulants here are defined like in Definition 4 which is a sensible definition even if the feature maps are not associated to kernels.

Theorem 4 (Generalization of Theorem 2 and Theorem 3)**.** *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be Polish spaces and $\varphi_i : \mathcal{X}_i \to \mathcal{B}_i$ be continuous, bounded and injective feature maps into separable Banach spaces $\mathcal{B}_i$ for $i = 1, \ldots d$. Let γ and η be probability measures on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. Then*

1. *$\gamma = \eta$ if and only if $\kappa(\gamma) = \kappa(\eta)$.*
2. *$\gamma = \bigotimes_{i=1}^d \gamma|_{\mathcal{X}_i}$ if and only if the cross cumulants vanish, that is $\kappa^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}_+^d$.*

Proof.
• Item 2: We want to show that the cross cumulants vanish if and only if $\gamma = \bigotimes_{i=1}^d \gamma|_{\mathcal{X}_i}$. By Lemma 4 it is enough to show that

$$\mathbb{E}_\gamma \Big[p\big(\varphi_1(X_1), \ldots, \varphi_d(X_d)\big) \Big] = \mathbb{E}_{\bigotimes_{i=1}^d \gamma|_{\mathcal{X}_i}} \Big[p\big(\varphi_1(X_1), \ldots, \varphi_d(X_d)\big) \Big]$$

for any monomial function p. Let us take linear functionals $F = (f_1, \ldots, f_d)$ and note that

$$\langle F^{\mathbf{i}}, \kappa^{\mathbf{i}}(\gamma) \rangle = \sum_{\pi \in P(d)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} \left[f_1(\varphi_1(X_1)) \cdots f_d(\varphi_d(X_d)) \right]$$

which is the classical cumulant of the vector-valued random variable

$$\big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\big),$$

where $(X_1, \ldots, X_d) \sim \gamma$. Hence by classical results (Speed, 1983), all cross cumulants of $\big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\big)$ vanish if and only if the cross moments split, that is to say

$$\mathbb{E}_\gamma \Big[p\big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\big) \Big] = \mathbb{E}_{\bigotimes_{i=1}^d \gamma|_{\mathcal{X}_i}} \Big[p\big((f_1 \circ \varphi_1)(X_1), \ldots, (f_d \circ \varphi_d)(X_d)\big) \Big]$$

for any monomial p on $\mathbb{R}^d$. Since $f_1, \ldots, f_d$ were arbitrary this holds for all monomials, which shows the assertion.

• Item 1: By assumption $\kappa^{\mathbf{i}}(\gamma) = \kappa^{\mathbf{i}}(\eta)$ for every $\mathbf{i} \in \mathbb{N}^d$; this implies that $\mathbb{E}_\gamma p(\varphi_1, \ldots, \varphi_d) = \mathbb{E}_\eta p(\varphi_1, \ldots, \varphi_d)$ for any polynomial p, so we can apply Lemma 4. $\square$

Proposition 1 (Theorem 2 and Theorem 3)**.** *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be Polish spaces and $k_i : \mathcal{X}_i^2 \to \mathbb{R}$ be a collection of bounded, continuous, point-separating kernels. Let γ and η be be probability measures on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. Then*

1. *$\gamma = \eta$ if and only if $\kappa_{k_1, \ldots, k_d}(\gamma) = \kappa_{k_1, \ldots, k_d}(\eta)$.*
2. *$\gamma = \bigotimes_{i=1}^d \gamma|_{\mathcal{X}_i}$ if and only if $\kappa_{k_1, \ldots, k_d}^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i} \in \mathbb{N}_+^d$.*

Proof. We reduce the proof to the checking of the conditions of Theorem 4. Let φ_i denote the canonical feature map of the kernel k_i, and let $\mathcal{B}_i := \mathcal{H}_{k_i}$ be the RKHS associated to k_i ($i \in \{1, \ldots, d\}$). For all $i \in \{1, \ldots, d\}$, φ_i is (i) bounded by the boundeness of k_i since $\|\varphi_i(x)\|_{\mathcal{H}_{k_i}}^2 = k_i(x, x) \le \sup_{x \in \mathcal{X}_i} |k_i(x, x)| < \infty$, (ii) continuous by the continuity of k_i (Steinwart & Christmann, 2008, Lemma 4.29), (iii) injective by the point-separating property of k_i. The separability of $\mathcal{H}_{k_i}$ follows (Steinwart & Christmann, 2008, Lemma 4.33) from the separability of $\mathcal{X}_i$ and the continuity of k_i ($i \in \{1, \ldots, d\}$). Note: Details on the expected kernel trick part of Theorem 2 and Theorem 3 are provided in Section E. $\square$

D Additional Experiments and Details

Here we give additional details on the experiments that were performed, and discuss some further experiments that did not fit into the main text.

Background on permutation testing. Permutation testing works by bootstrapping the distribution of a test statistic under the null hypothesis. This allows the user to estimate confidence intervals under the null, which is a powerful all-purpose way of doing so when analytic expressions are unavailable. As an example, assume we have two probability measures γ, η on $\mathcal{X}$ with i.i.d. samples $x_1, \ldots, x_N \sim \gamma, y_1, \ldots, y_N \sim \eta$. If the null hypothesis is that $\gamma = \eta$ then we may set

$$(z_1, \ldots, z_{2N}) := (x_1, \ldots, x_N, y_1, \ldots, y_N)$$

so that for any permutation σ on $2N$ elements, we get two different set of of i.i.d. samples from $\gamma = \eta$ by using the empirical measures

$$\tilde{\gamma}_\sigma := (z_{\sigma(1)}, \ldots, z_{\sigma(N)}), \quad \tilde{\eta}_\sigma := (z_{\sigma(N+1)}, \ldots, z_{\sigma(2N)})$$

and for any statistic $S : \mathcal{P}(\mathcal{X})^2 \to \mathbb{R}$, we may estimate $S(\gamma, \eta)$ under the null by sampling from $S(\tilde{\gamma}_\sigma, \tilde{\eta}_\sigma)$. If the null hypothesis were true, we might expect $S(\gamma, \eta)$ to lie in a region with high probability of the permutation estimator, and we can use this as a criteria for rejecting the null. Under fairly weak assumptions, this yields a test at the appropriate level (Chung & Romano, 2013).

Comparing a uniform and a mixture. Any uniform random variable over a symmetric interval will have 0 mean and skewness, so a symmetric mixture only needs to match the variance. If X is a $50/50$ mixture of $U[a, b]$ and $U[-a, -b]$ then

$$\mathrm{Var}(X) = \frac{2}{3}\left(b^2 + ba + a^2\right)$$

so if Y is distributed according to $U[-c, c]$ then we only need to solve

$$b^2 + ba + a^2 = c^2$$

which is straightforward for a given a and c.

Computational complexity of estimators. The V-statistic for $d^{(2)}$ as written in Lemma 2 is bottlenecked by the matrix multiplications. We may note however that for two matrices $\mathbf{A}, \mathbf{B}$ it holds that

$$\mathrm{Tr}(\mathbf{A}^\top \mathbf{B}) = \langle \mathbf{A} \circ \mathbf{B} \rangle,$$

where $\langle \cdot \rangle$ denotes the sum over elements and $\circ$ denotes the Hadamard product. We also note that for for $\mathbf{H}_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ we have $\left(\mathbf{A}\mathbf{H}_n\right)_{i,j} = \frac{1}{n}\sum_{c=1}^n A_{i,c}$. Using both of these tricks we may compute both $d^{(2)}$ and CSIC without any matrix multiplications, which brings the computational complexity down to $O(N^2)$ for both. For a comparison of actual computation time, see Fig. 6 and Fig. 7, where the average computational times for out methods are compared to the KME and and HSIC for N between 50 and 2000.

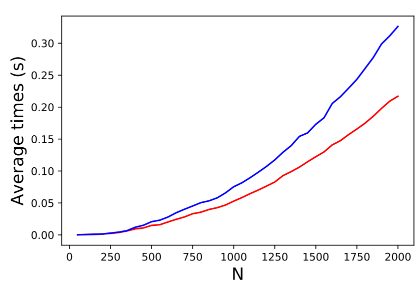

Figure 6: Average computational time in seconds for KME (red) and $d^{(2)}$ (blue) for sample size N between 50 and 2000.

Type I error on the Seoul Bicycle data. The results when comparing the winter data to itself is presented in Fig. 8. As we see the performance is similar for both estimators and lies between 5 and 10%.

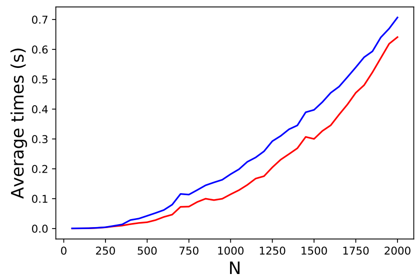

Figure 7: Average computational time in seconds for HSIC (red) and CSIC (blue) for sample size N between 50 and 2000.

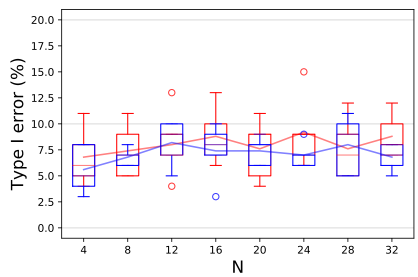

Figure 8: Type I errors using MMD (red) and $d^{(2)}$ (blue) on the Seoul bicycle data set.

Classical vs. kernelized cumulants. Using the same distributions as in the synthetic independence testing experiment, we now compare X with $Y_{0.5}^2$ to contrast independence testing with classical cumulants with their kernelized counterpart. The results are summarized in Table 1 where they are displayed as the median value $\pm$ half the difference between the 75th and 25th percentile. We consider every combination of classical vs. kernelized, variance vs. skewness, and two different sample sizes. One can observe that the classical variance based test performs poorly compared to a classical skewness test, the kernelized variance test is almost as powerful as the kernelized skewness test, and in all cases the kernelized tests deliver higher power.

E Kernel Trick Computations

Here we show how to arrive at the expressions used for the V-statistics used in the experiments.

Given a real analytic function $f(x, \dots, x_d) = \sum_{\mathbf{i} \in \mathbb{N}^d} f_{\mathbf{i}} x^{\mathbf{i}}$ in m variables with nonzero radius of convergence and Hilbert spaces $\mathcal{H}_1, \dots, \mathcal{H}_d$ we may (formally) extend f to a function

$$f_{\otimes} : \prod_{i=1}^{d} \mathcal{H}_i \to \mathrm{T}, \quad f_{\otimes}(x_1, \dots, x_d) = \prod_{\mathbf{i} \in \mathbb{N}^d} f_{\mathbf{i}} x^{\otimes \mathbf{i}}.$$

Moreover, if the Hilbert spaces are RKHSs then we have the following result.

Lemma 5 (Nonlinear kernel trick)**.** *For any collection of RKHSs $\mathcal{H}_1, \dots, \mathcal{H}_d$ with feature maps $\varphi_i : \mathcal{X}_i \to \mathcal{H}_i$, assume that f and g are real analytic functions with radii of convergence $r(f)$ and*

Table 1: Comparison of classical and kernelized cumulants for independence testing with both variance and skewness.

N=20	Variance	Skewness
Classical	$19\% \pm 3.0\%$	$56\% \pm 3.5\%$
Rbf kernel	$39\% \pm 4.5\%$	$59\% \pm 3.0\%$

N=30	Variance	Skewness
Classical	$17\% \pm 0.5\%$	$68\% \pm 1.0\%$
Rbf kernel	$65\% \pm 3.5\%$	$79\% \pm 1.5\%$

$r(g)$ *such that* $\max_{1\le i\le d}\sup_{x\in\mathcal{X}_i}|\varphi_i(x)| < \min(r(f), r(g))$. *Then*

$$\langle f_\otimes\big(\varphi_1(x_1),\dots,\varphi_d(x_d)\big), g_\otimes\big(\varphi_1(y_1),\dots,\varphi_d(y_d)\big)\rangle_{\mathrm{T}} = \sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}} g_{\mathbf{i}} k_1(x_1,y_1)^{i_1}\dots k_d(x_d,y_d)^{i_d}.$$

Proof. Since the image of the φ_is lie inside the radius of convergence of $f_\otimes$ and $g_\otimes$ the power series converge absolutely and we can write

$$\begin{aligned}\langle f_\otimes\big(\varphi^{\otimes\mathbf{i}}(x^{\mathbf{i}})\big), g_\otimes\big(\varphi^{\otimes\mathbf{i}}(y^{\mathbf{i}})\big)\rangle_{\mathrm{T}} &= \langle \sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}}\varphi^{\otimes\mathbf{i}}(x^{\mathbf{i}}), \sum_{\mathbf{i}\in\mathbb{N}^d} g_{\mathbf{i}}\varphi^{\otimes\mathbf{i}}(y^{\mathbf{i}})\rangle_{\mathrm{T}} \\ &= \sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}} g_{\mathbf{i}}\langle\varphi^{\otimes\mathbf{i}}(x^{\mathbf{i}}),\varphi^{\otimes\mathbf{i}}(y^{\mathbf{i}})\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} = \sum_{\mathbf{i}\in\mathbb{N}^d} f_{\mathbf{i}} g_{\mathbf{i}} k_1(x_1,y_1)^{i_1}\dots k_d(x_d,y_d)^{i_d},\end{aligned}$$

where $\mathcal{H} = \mathcal{H}_1\times\dots\times\mathcal{H}_d$. □

Using Lemma 5, we can choose kernels $k_i : \mathcal{X}_i^2 \to \mathbb{R}$ with associated RKHSs $\mathcal{H}_i$ and feature maps φ_i and some $\mathbf{i}\in\mathbb{N}^d$ with $\deg(\mathbf{i}) = m$. We make the observation that with $X = (X_1,\dots,X_d)\sim\gamma$, $Y = (Y_1,\dots,Y_d)\sim\eta$ and $k^{\otimes\mathbf{i}}$ and $\mathcal{H}^{\otimes\mathbf{i}}$ as in (4), one has

$$\begin{aligned}\langle\kappa^{\mathbf{i}}(\gamma),\kappa^{\mathbf{i}}(\eta)\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} &= \langle \sum_{\pi\in P(m)} c_\pi \mathbb{E}_{\gamma^{\mathbf{i}}_\pi}\varphi^{\otimes\mathbf{i}}(X^{\mathbf{i}}), \sum_{\tau\in P(m)} c_\tau\mathbb{E}_{\eta^{\mathbf{i}}_\tau}\varphi^{\otimes\mathbf{i}}(Y^{\mathbf{i}})\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} \\ &= \sum_{\pi,\tau\in P(m)} c_\pi c_\tau\langle\mathbb{E}_{\gamma^{\mathbf{i}}_\pi}\varphi^{\otimes\mathbf{i}}(X^{\mathbf{i}}),\mathbb{E}_{\eta^{\mathbf{i}}_\tau}\varphi^{\otimes\mathbf{i}}(Y^{\mathbf{i}})\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} \\ &= \sum_{\pi,\tau\in P(m)} c_\pi c_\tau\mathbb{E}_{\gamma^{\mathbf{i}}_\pi\otimes\eta^{\mathbf{i}}_\tau}\langle\varphi^{\otimes\mathbf{i}}(X^{\mathbf{i}}),\varphi^{\otimes\mathbf{i}}(Y^{\mathbf{i}})\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} \\ &= \sum_{\pi,\tau\in P(m)} c_\pi c_\tau\mathbb{E}_{\gamma^{\mathbf{i}}_\pi\otimes\eta^{\mathbf{i}}_\tau} k^{\otimes\mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m)),\end{aligned}$$

Since

$$\begin{aligned}\|\kappa^{\mathbf{i}}(\gamma)\|^2_{\mathcal{H}^{\otimes\mathbf{i}}} &= \langle\kappa^{\mathbf{i}}(\gamma),\kappa^{\mathbf{i}}(\gamma)\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} \\ \|\kappa^{\mathbf{i}}(\gamma)-\kappa^{\mathbf{i}}(\eta)\|^2_{\mathcal{H}^{\otimes\mathbf{i}}} &= \langle\kappa^{\mathbf{i}}(\gamma),\kappa^{\mathbf{i}}(\gamma)\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} + \langle\kappa^{\mathbf{i}}(\eta),\kappa^{\mathbf{i}}(\eta)\rangle_{\mathcal{H}^{\otimes\mathbf{i}}} - 2\langle\kappa^{\mathbf{i}}(\gamma),\kappa^{\mathbf{i}}(\eta)\rangle_{\mathcal{H}^{\otimes\mathbf{i}}}\end{aligned}$$

one gets the expected kernel trick statements of Theorem 2 and Theorem 3.

We are now interested in explicitly computing the expression $\|\kappa^{(1,2)}_{k,\ell}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1}\otimes\mathcal{H}_\ell^{\otimes 2}}$, $\|\kappa^{(2)}_k(\gamma)-\kappa^{(2)}_k(\eta)\|^2_{\mathcal{H}^{(1,1)}}$ and $\|\kappa^{(3)}_k(\gamma)-\kappa^{(3)}_k(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$, and their corresponding V-statistics. Recall that for a (w.l.o.g.) symmetric, measurable function $h(z_1,\dots,z_m)$, the V-statistic of h with N samples $Z_1,\dots,Z_N$ is defined as

$$\mathrm{V}(h;Z_1,\dots,Z_N) \coloneqq N^{-m}\sum_{i_1,\dots,i_m=1}^{N} h(Z_{i_1},\dots,Z_{i_m}).$$

Under fairly general conditions, the V-statistic converges in distribution to $\mathbb{E}[h(Z_1,\dots,Z_m)]$ and a well-developed theory describes this convergence Van der Waart (2000); Serfling (1980); Arcones & Giné (1992).

Example E.1 (Estimating $\|\kappa^{(2)}_k(\gamma)-\kappa^{(2)}_k(\eta)\|^2_{\mathcal{H}^{(1,1)}}$)**.** *Let* X, X', X'', X''' *denote independent copies of* γ *and* Y, Y', Y'', Y''' *denote independent copies of* η*. The full expression for* $\|\kappa^{(2)}_k(\gamma)-\kappa^{(2)}_k(\eta)\|^2_{\mathcal{H}^{(1,1)}}$ *is*

$$\begin{aligned}\|\kappa^{(2)}_k(\gamma)-\kappa^{(2)}_k(\eta)\|^2_{\mathcal{H}^{(1,1)}} = {} & \mathbb{E}k(X,X')k(X'',X''') + \mathbb{E}k(Y,Y')k(Y'',Y''') \\ & + \mathbb{E}k(X,X')^2 + \mathbb{E}k(Y,Y')^2 \\ & + 2\mathbb{E}k(X,Y)k(X',Y) + 2\mathbb{E}k(X,Y)k(X,Y') \\ & - 2\mathbb{E}k(X,Y)k(X',Y') - 2\mathbb{E}k(X,Y)^2 \\ & - 2\mathbb{E}k(X,X')k(X,X'') - 2\mathbb{E}k(Y,Y')k(Y,Y'').\end{aligned} \tag{12}$$

Given samples $(x_i)_{i=1}^N$, $(y_i)_{i=1}^M$ from γ and η respectively the corresponding V statistic is

$$\begin{aligned}
&\frac{1}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_l) + \frac{1}{M^4}\sum_{i,j,\kappa,l=1}^{M} k(y_i,y_j)k(y_\kappa,y_l) \qquad (13)\\
&+ \frac{1}{N^2}\sum_{i,j=1}^{N} k(x_i,x_j)^2 + \frac{1}{M^2}\sum_{i,j=1}^{M} k(y_i,y_j)^2\\
&+ \frac{2}{N^2M}\sum_{i,\kappa=1}^{N}\sum_{j=1}^{M} k(x_i,y_j)k(x_\kappa,y_j) + \frac{2}{NM^2}\sum_{i=1}^{N}\sum_{j,\kappa=1}^{M} k(x_i,y_j)k(x_i,y_\kappa)\\
&- \frac{2}{N^2M^2}\sum_{i,l=1}^{N}\sum_{j,\kappa=1}^{M} k(x_i,y_j)k(x_\kappa,y_l) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(x_i,y_j)^2\\
&- \frac{2}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,x_j)k(x_i,x_\kappa) - \frac{2}{M^3}\sum_{i,j,\kappa=1}^{M} k(y_i,y_j)k(y_i,y_\kappa).
\end{aligned}$$

Let us define the Gram matrices $\mathbf{K}_x = [k(x_i,x_j)]_{i,j=1}^N \in \mathbb{R}^{N\times N}$, $\mathbf{K}_y = [k(y_i,y_j)]_{i,j=1}^M \in \mathbb{R}^{M\times M}$, $\mathbf{K}_{x,y} = [k(x_i,y_j)]_{i,j=1}^{N,M}$ and let $\mathbf{H}_N = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \in \mathbb{R}^{N\times N}$, $\mathbf{H}_M = \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top \in \mathbb{R}^{M\times M}$ be the centering, then (13) *can be rewritten as*

$$\begin{aligned}
&\frac{1}{N^2}\mathrm{Tr}(\mathbf{H}_N\mathbf{K}_x\mathbf{H}_N\mathbf{K}_x) + \frac{1}{M^2}\mathrm{Tr}(\mathbf{H}_M\mathbf{K}_y\mathbf{H}_M\mathbf{K}_y) + \frac{1}{N^2}\mathrm{Tr}(\mathbf{K}_x^2) + \frac{1}{M^2}\mathrm{Tr}(\mathbf{K}_y^2)\\
&+ \frac{2}{NM}\mathrm{Tr}(\mathbf{K}_{xy}\mathbf{H}_N\mathbf{K}_{xy}) + \frac{2}{NM}\mathrm{Tr}(\mathbf{K}_{xy}\mathbf{H}_M\mathbf{K}_{xy}^\top) - \frac{2}{NM}\mathrm{Tr}(\mathbf{H}_M\mathbf{K}_{xy}^\top\mathbf{H}_N\mathbf{K}_{xy}) - \frac{2}{NM}\mathrm{Tr}(\mathbf{K}_{xy}^2)\\
&- \frac{2}{N^2}\mathrm{Tr}(\mathbf{K}_x\mathbf{H}_N\mathbf{K}_x) - \frac{2}{M^2}\mathrm{Tr}(\mathbf{K}_y\mathbf{H}_M\mathbf{K}_y)
\end{aligned}$$

which simplifies to

$$\frac{1}{N^2}\mathrm{Tr}\big[(\mathbf{K}_x(\mathbf{I}-\mathbf{H}_N))^2\big] + \frac{1}{M^2}\mathrm{Tr}\big[(\mathbf{K}_y(\mathbf{I}-\mathbf{H}_M))^2\big] - \frac{2}{NM}\mathrm{Tr}\big[\mathbf{K}_{xy}(\mathbf{I}-\mathbf{H}_M)\mathbf{K}_{xy}^\top(\mathbf{I}-\mathbf{H}_N)\big].$$

This estimator can be computed in quadratic time.

Example E.2 (Estimating $\|\kappa_{k,\ell}^{(1,2)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1}\otimes\mathcal{H}_\ell^{\otimes 2}}$)**.** *Let k denote the kernel on $\mathcal{X}_1$ and ℓ denote the kernel on $\mathcal{X}_2$. Let $(X,Y),(X',Y'),(X'',Y''),(X^{(3)},Y^{(3)}),(X^{(4)},Y^{(4)}),(X^{(5)},Y^{(5)})$ denote independent copies of $\gamma \in \mathcal{P}(\mathcal{X}_1\times\mathcal{X}_2)$. The full expression for $\|\kappa_{k,\ell}^{(1,2)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 1}\otimes\mathcal{H}_\ell^{\otimes 2}}$ is*

$$\begin{aligned}
&\mathbb{E}k(X,X')k(X,X')\ell(Y,Y') - 4\mathbb{E}k(X,X')k(X,X'')\ell(Y,Y')\\
&- 2\mathbb{E}k(X,X')k(X,X')\ell(Y,Y'') + 4\mathbb{E}k(X,X')k(X,X'')\ell(Y,Y^{(3)})\\
&+ 2\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y,Y') + 2\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y,Y^{(3)})\\
&+ 4\mathbb{E}k(X,X')k(X'',X')\ell(Y,Y^{(3)}) + \mathbb{E}k(X,X')k(X,X')\ell(Y'',Y^{(3)})\\
&- 8\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y^{(4)},Y') - 4\mathbb{E}k(X,X')k(X'',X')\ell(Y^{(4)},Y^{(3)})\\
&+ 4\mathbb{E}k(X,X')k(X'',X^{(3)})\ell(Y^{(4)},Y^{(5)}).
\end{aligned}$$

Given samples $(x_i, y_i)_{i=1}^N$ from γ the corresponding V-statistic for this expression is

$$
\begin{aligned}
&\frac{1}{N^2}\sum_{i,j=1}^{N} k(x_i,x_j)k(x_i,x_j)\ell(y_i,y_j) - \frac{4}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,x_j)k(x_i,x_\kappa)\ell(y_i,y_j)\\
&-\frac{2}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,x_j)k(x_i,x_j)\ell(y_i,y_\kappa) + \frac{4}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_i,x_\kappa)\ell(y_i,y_l)\\
&+\frac{2}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_i,y_j) + \frac{2}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_i,y_l)\\
&+\frac{4}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_\kappa,x_j)\ell(y_i,y_l) + \frac{1}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,x_j)k(x_i,x_j)\ell(y_\kappa,y_l)\\
&-\frac{8}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_m,y_j) - \frac{4}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,x_j)k(x_\kappa,x_j)\ell(y_m,y_l)\\
&+\frac{4}{N^6}\sum_{i,j,\kappa,l,m,n=1}^{N} k(x_i,x_j)k(x_\kappa,x_l)\ell(y_m,y_n).
\end{aligned}
$$

Using the shorthand notation $\mathbf{K} = \mathbf{K}_x, \mathbf{L} = \mathbf{L}_y$ and $\mathbf{H} = \mathbf{H}_N$ and denoting by $\circ$ the Hadamard product $[\mathbf{A} \circ \mathbf{B}]_{i,j} = A_{i,j}B_{i,j}$ and $\langle\cdot\rangle$ the sum over all elements of a matrix $\langle \mathbf{A}\rangle = \sum_{i,j=1}^N A_{i,j}$, the V-statistic above can be written in the simpler form

$$
\begin{aligned}
\frac{1}{N^2}\Bigg\langle & \mathbf{K}\circ\mathbf{K}\circ\mathbf{L} - 4\mathbf{K}\circ\mathbf{KH}\circ\mathbf{L} - 2\mathbf{K}\circ\mathbf{K}\circ\mathbf{LH}\\
&+4\mathbf{KH}\circ\mathbf{K}\circ\mathbf{LH} + 2\mathbf{K}\circ\mathbf{L}\left\langle\frac{\mathbf{K}}{N^2}\right\rangle + 2\mathbf{KH}\circ\mathbf{HK}\circ\mathbf{L}\\
&+4\mathbf{K}\circ\mathbf{HK}\circ\mathbf{LH} + \mathbf{K}\circ\mathbf{K}\left\langle\frac{\mathbf{L}}{N^2}\right\rangle - 8\mathbf{K}\circ\mathbf{LH}\left\langle\frac{\mathbf{K}}{N^2}\right\rangle\\
&-4\mathbf{K}\circ\mathbf{HK}\left\langle\frac{\mathbf{L}}{N^2}\right\rangle + 4\left\langle\frac{\mathbf{K}}{N^2}\right\rangle^2\mathbf{L}\Bigg\rangle.
\end{aligned}
$$

Again this estimator can be computed in quadratic time.

Example E.3 (Estimating $\|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}$)**.** *In order to estimate $d^{(3)}(\gamma,\eta)$ we note that one can write*

$$
\begin{aligned}
\|\kappa_k^{(3)}(\gamma) - \kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}} &= \|\kappa_k^{(3)}(\gamma)\|^2_{\mathcal{H}_k^{\otimes 3}} + \|\kappa_k^{(3)}(\eta)\|^2_{\mathcal{H}_k^{\otimes 3}}\\
&\quad - 2\langle\kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}}.
\end{aligned}
$$

We can estimate the first two terms like in Example E.2, and the third term can be expressed as

$$
\begin{aligned}
\langle\kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta)\rangle_{\mathcal{H}_k^{\otimes 3}} &= \mathbb{E}k(X,Y)^3 - 3\mathbb{E}k(X,Y)^2k(X,Y')^2\\
&- 3\mathbb{E}k(X,Y)^2k(X',Y)^2 + 6\mathbb{E}k(X,Y)k(X,Y')k(X',Y)\\
&+ 3\mathbb{E}k(X,Y)^2k(X',Y') + 2\mathbb{E}k(X,Y)k(X',Y)k(X'',Y)\\
&+ 2\mathbb{E}k(X,Y)k(X,Y')k(X,Y'') - 6\mathbb{E}k(X,Y)k(X,Y')k(X',Y'')\\
&- 6\mathbb{E}k(X,Y)k(X',Y)k(X'',Y') + 4\mathbb{E}k(X,Y)k(X',Y')k(X'',Y'').
\end{aligned}
$$

For simplicity we will assume that we have an equal number of samples (N) from both measures $(x_i)_{i=1}^N \in \gamma$ and $(y_i)_{i=1}^N \in \eta$. The V-statistic for $\langle \kappa_k^{(3)}(\gamma), \kappa_k^{(3)}(\eta) \rangle_{\mathcal{H}_k^{\otimes 3}}$ can be expressed as

$$\begin{aligned}
&\frac{1}{N^2}\sum_{i,j=1}^{N} k(x_i,y_j)^3 - \frac{3}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,y_j)^2 k(x_i,y_\kappa) \\
&- \frac{3}{N^3}\sum_{i,j,\kappa=1}^{N} k(x_i,y_j)^2 k(x_\kappa,y_i) + \frac{6}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,y_j)k(x_i,y_\kappa)k(x_l,y_j) \\
&+ \frac{3}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,y_j)^2 k(x_\kappa,y_l) + \frac{2}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,y_j)k(x_\kappa,y_j)k(x_l,y_j) \\
&+ \frac{2}{N^4}\sum_{i,j,\kappa,l=1}^{N} k(x_i,y_j)k(x_i,y_\kappa)k(x_i,y_l) - \frac{6}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,y_j)k(x_i,y_\kappa)k(x_l,y_m) \\
&- \frac{6}{N^5}\sum_{i,j,\kappa,l,m=1}^{N} k(x_i,y_j)k(x_\kappa,y_j)k(x_l,y_m) + \frac{4}{N^6}\sum_{i,j,\kappa,l,m,n=1}^{N} k(x_i,y_j)k(x_\kappa,y_l)k(x_m,y_n).
\end{aligned}$$

Using the notation $\mathbf{K}_{xy} = [k(x_i,y_j)]_{i,j=1}^N$, this estimator simplifies to

$$\begin{aligned}
\frac{1}{N^2}\Bigg\langle &\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{K}_{xy} - 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \\
&- 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} + 6\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \circ \mathbf{H}\mathbf{K}_{xy} \\
&+ 3\mathbf{K}_{xy} \circ \mathbf{K}_{xy} \left\langle \frac{\mathbf{K}_{xy}}{N^2} \right\rangle + 2\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \\
&+ 2\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \circ \mathbf{K}_{xy}\mathbf{H} - 6\mathbf{K}_{xy} \circ \mathbf{K}_{xy}\mathbf{H} \left\langle \frac{\mathbf{K}_{xy}}{N^2} \right\rangle \\
&- 6\mathbf{K}_{xy} \circ \mathbf{H}\mathbf{K}_{xy} \left\langle \frac{\mathbf{K}_{xy}}{N^2} \right\rangle + 4\left\langle \frac{\mathbf{K}}{N^2} \right\rangle^2 \mathbf{K}_{xy} \Bigg\rangle.
\end{aligned}$$

We mention also that the first two terms $\|\kappa_k^{(3)}(\gamma)\|_{\mathcal{H}_k^{\otimes 3}}^2, \|\kappa_k^{(3)}(\eta)\|_{\mathcal{H}_k^{\otimes 3}}^2$ can be computed a little more simply than in Example E.2 since the expressions have more symmetry, using the notation $\mathbf{K}_x = [k(x_i,x_j)]_{i,j=1}^N$ we can write down the V-statistic for $\|\kappa_k^{(3)}(\gamma)\|_{\mathcal{H}_k^{\otimes 3}}^2$ as

$$\begin{aligned}
\frac{1}{N^2}\Bigg\langle &\mathbf{K}_x \circ \mathbf{K}_x \circ \mathbf{K}_x - 6\mathbf{K}_x \circ \mathbf{K}_x\mathbf{H} \circ \mathbf{K}_x \\
&+ 4\mathbf{K}_x\mathbf{H} \circ \mathbf{K}_x \circ \mathbf{K}_x\mathbf{H} + 3\mathbf{K}_x \circ \mathbf{K}_x \left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle \\
&+ 6\mathbf{K}_x\mathbf{H} \circ \mathbf{H}\mathbf{K}_x \circ \mathbf{K}_x - 12\mathbf{K}_x \circ \mathbf{H}\mathbf{K}_x \left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle \\
&+ 4\left\langle \frac{\mathbf{K}_x}{N^2} \right\rangle^2 \mathbf{K}_x \Bigg\rangle
\end{aligned}$$

with a similar expression for $\|\kappa_k^{(3)}(\eta)\|_{\mathcal{H}_k^{\otimes 3}}^2$. The estimator can be computed in quadratic time.