Nyström M-Hilbert-Schmidt Independence Criterion

Abstract

Kernel techniques are among the most popular and powerful approaches of data science. Among the key features that make kernels ubiquitous are (i) the number of domains they have been designed for, (ii) the Hilbert structure of the function class associated to kernels facilitating their statistical analysis, and (iii) their ability to represent probability distributions without loss of information. These properties give rise to the immense success of Hilbert-Schmidt independence criterion (HSIC) which is able to capture joint independence of random variables under mild conditions, and permits closed-form estimators with quadratic computational complexity (w.r.t. the sample size). In order to alleviate the quadratic computational bottleneck in large-scale applications, multiple HSIC approximations have been proposed, however these estimators are restricted to $M = 2$ random variables, do not extend naturally to the $M \geqslant 2$ case, and lack theoretical guarantees. In this work, we propose an alternative Nyström-based HSIC estimator which handles the $M \geqslant 2$ case, prove its consistency, and demonstrate its applicability in multiple contexts, including synthetic examples, dependency testing of media annotations, and causal discovery.

1 INTRODUCTION

Kernels methods [Aronszajn, 1950] have been on the forefront of data science for more than 20 years [Schölkopf and Smola, 2002, Steinwart and Christmann, 2008], and they underpin some of the most powerful and principled machine learning techniques currently known. The key idea of kernels is to map the data into a (possibly infinite-dimensional) feature space in which one computes the inner product implicitly by means of a symmetric, positive definite function, the so-called kernel function.

Kernel functions have been designed for strings [Watkins, 1999, Lodhi et al., 2002] or more generally for sequences [Király and Oberhauser, 2019], sets [Haussler, 1999, Gärtner et al., 2002], rankings [Jiao and Vert, 2016], fuzzy domains [Guevara et al., 2017] and graphs [Borgwardt et al., 2020], which renders them broadly applicable. Their extension to the space of probability measures [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007] allows to represent distributions in a reproducing kernel Hilbert space (RKHS) by the so-called mean embedding. Such embeddings form the main building block of maximum mean discrepancy (MMD; Smola et al. [2007], Gretton et al. [2012]), which quantifies the discrepancy of two distributions as the RKHS distance of their respective mean embeddings. MMD is (i) a semi-metric on probability measures, (ii) a metric iff. the kernel is characteristic [Fukumizu et al., 2008, Sriperumbudur et al., 2010], (iii) an instance of integral probability metrics (IPM; Müller [1997], Zolotarev [1983]) when the underlying function class in the IPM is chosen to be the unit ball in an RKHS.

Measuring the discrepancy of a joint distribution to the product of its marginals by MMD gives rise to the Hilbert-Schmidt independence criterion (HSIC; Gretton et al. [2005]). HSIC was shown to be equivalent [Sejdinovic et al., 2013b] to distance covariance [Székely et al., 2007, Székely and Rizzo, 2009, Lyons, 2013]; Sheng and Sriperumbudur [2023] have recently proved a similar result for the conditional case. HSIC is known to capture the independence of $M = 2$ random variables with characteristic $(k_m)_{m=1}^2$ kernels (on the respective domains) as proved by Lyons [2013]; for more than two components ($M > 2$; Quadrianto et al. [2009], Sejdinovic et al. [2013a], Pfister et al. [2018]) universality [Steinwart, 2001, Micchelli et al., 2006] of $(k_m)_{m=1}^M$-s is sufficient [Szabó and Sriperumbudur, 2018]. HSIC has been deployed successfully in a wide range of domains including independence testing [Gretton et al., 2008, Pfister et al., 2018, Albert et al., 2022], feature selection [Camps-Valls et al., 2010, Song et al., 2012, Wang et al., 2022] with applications in biomarker detection [Climente-González et al., 2019] and wind power prediction [Bouche et al., 2022], clustering [Song et al., 2007, Climente-González et al., 2019], and causal discovery [Mooij et al., 2016, Pfister et al., 2018, Chakraborty and Zhang, 2019, Schölkopf et al., 2021].

Various estimators for HSIC and other dependence measures exist in the literature, out of which we summarize the most closely related ones to our work in Table 1. The classical V-statistic based HSIC estimator (V-HSIC; Gretton et al. [2005], Quadrianto et al. [2009], Pfister et al. [2018]) is powerful but its runtime increases quadratically with the number of samples, which limits it applicability in large-scale settings. To tackle this severe computational bottleneck, approximations of HSIC (N-HSIC, RFF-HSIC) have been proposed [Zhang et al., 2017], relying on the Nyström [Williams and Seeger, 2001] and the random Fourier feature (RFF; Rahimi and Recht [2007]) method, respectively. However, these estimators (i) are limited to two components, (ii) their extension to more than two components is not straightforward, and (iii) they lack theoretical guarantees. The RFF-based approach is further restricted to finite-dimensional Euclidean domains and to translation-invariant kernels. The normalized finite set independence criterion (NFSIC; Jitkrittum et al. [2017]) replaces the RKHS norm of HSIC with an L_2 one which allows the construction of linear-time estimators. However, NFSIC is also limited to two components, requires $\mathbb{R}^d$-valued input, and analytic kernels [Chwialkowski et al., 2015]. A novel complementary approach is the kernel partial correlation coefficient (KPCC; [Huang et al., 2022]) but when applied to kernel-enriched domains its runtime complexity is cubic in the sample size.

The restriction of existing HSIC approximations to two components is a severe limitation in recent applications like causal discovery which require independence tests capable of handling more than two components. Furthermore, the emergence of large-scale data sets necessitates algorithms that scale well in the sample size. To alleviate these bottlenecks, we make the following **contributions**.

1. We propose Nyström M-HSIC, an efficient HSIC estimator, which can handle more than two components and has runtime $\mathcal{O}\left(Mn'^3 + Mn'n\right)$, where n denotes the number of samples, $n' \ll n$ stands for the number of Nyström points, and M is the number of random variables whose independence is measured.
2. We provide theoretical guarantees for Nyström M-HSIC: we prove that our estimator converges with rate $\mathcal{O}\left(n^{-1/2}\right)$ for $n' \sim \sqrt{n}$, which matches the convergence of the quadratic-time estimator.
3. We perform an extensive suite of experiments to demonstrate the efficiency of Nyström M-HSIC. These applications include dependency testing of media annotations and causal discovery. In the former, we achieve similar runtime and power as existing HSIC approximations. The latter requires testing joint independence of more than two components, which is beyond the capabilities of existing HSIC accelerations. Here, the proposed algorithm achieves the same performance as the quadratic-time HSIC estimator V-HSIC with a significantly reduced runtime.

The paper is structured as follows. Our notations are introduced in Section 2. The existing Nyström-based HSIC approximation for two components is reviewed in Section 3. Our proposed method, which is capable of handling $M \geqslant 2$ components, is presented in Section 4 together with its theoretical guarantees. In Section 5 we demonstrate the applicability of Nyström M-HSIC. All the proofs of our results are available in the supplementary material.

2 NOTATIONS

This section is dedicated to definitions and to the introduction of our target quantity Hilbert-Schmidt independence criterion (HSIC). In particular, we introduce the **notations** $[M]$, $\langle \mathbf{v}, \mathbf{w} \rangle$, $\|\mathbf{v}\|_2$, $\circ_{m\in[M]}\mathbf{A}_m$, $\operatorname{tr}(\mathbf{A})$, $\mathbf{A}^{-1}$, $\mathbf{A}^{-}$, $\mathbf{A}^{\mathsf{T}}$, $\|\mathbf{A}\|_{\mathrm{F}}$, $\mathbf{1}_d$, $\mathbf{I}_d$, span, $\mathcal{M}_1^+(\mathcal{X})$, $\mathcal{H}_k$, μ_k, MMD_k, $\otimes_{m=1}^M k_m$, $\otimes_{m=1}^M \mathbb{P}_m$, $\mathrm{HSIC}_{\otimes_{m=1}^M k_m}$, C_X, A^{-1}, $\|A\|_{\mathrm{op}}$, $\operatorname{tr}(A)$, $\mathcal{N}_X(\lambda)$, $\mathcal{O}_{\mathrm{P}}\left(r_n\right)$.

For a positive integer M, $[M] := \{1, \dots, M\}$. The Euclidean inner product of vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ is denoted by $\langle \mathbf{v}, \mathbf{w} \rangle$; the Euclidean norm is $\|\mathbf{v}\|_2 := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$. The Hadamard product of matrices $\mathbf{A}_m \in \mathbb{R}^{d_1 \times d_2}$ of equal size ($m \in [M]$) is $\circ_{m\in[M]}\mathbf{A}_m := \left[\prod_{m\in[M]}(\mathbf{A}_m)_{i,j}\right]_{i\in[d_1],j\in[d_2]}$. Matrix multiplication takes precendence over the Hadamard one. For a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\operatorname{tr}(\mathbf{A}) := \sum_{i\in[d]} A_{i,i}$ denotes its trace, $\mathbf{A}^{-1}$ is its inverse (assuming that $\mathbf{A}$ is non-singular), and $\mathbf{A}^{-}$ is its Moore–Penrose inverse. The transpose of a matrix $\mathbf{A} \in \mathbb{R}^{d_1\times d_2}$ is denoted by $\mathbf{A}^{\mathsf{T}}$. The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{d_1\times d_2}$ is $\|\mathbf{A}\|_{\mathrm{F}} := \sqrt{\sum_{i\in[d_1],j\in[d_2]}(A_{i,j})^2}$. The d-dimensional vector of ones is $\mathbf{1}_d$. The $d \times d$-sized identity matrix is denoted by $\mathbf{I}_d$. For a set S in a vector space, $\operatorname{span}(S)$ denotes the linear hull of S. Let $(\mathcal{X}, \tau_{\mathcal{X}})$ be a topological space, and $\mathcal{B}(\tau_{\mathcal{X}})$ the Borel sigma-algebra induced by the topology $\tau_{\mathcal{X}}$. All probability measures in the manuscript are meant with respect to the measurable space $(\mathcal{X}, \mathcal{B}(\tau_{\mathcal{X}}))$, and they are denoted by $\mathcal{M}_1^+(\mathcal{X})$. The RKHS $\mathcal{H}_k$ on $\mathcal{X}$ associated with a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the Hilbert space of functions $h : \mathcal{X} \to \mathbb{R}$ such that $k(\cdot, x) \in \mathcal{H}_k$ and $\langle h, k(\cdot, x) \rangle_{\mathcal{H}_k} = h(x)$ for all $x \in \mathcal{X}$ and $h \in \mathcal{H}_k$.[1] Kernels are assumed to be bounded (in other words, there exists $B \in \mathbb{R}$ such that $\sup_{x,x'\in\mathcal{X}} k(x, x') \leqslant B$) and measurable, and $\mathcal{H}_k$ is assumed to be separable

[1] $k(\cdot, x)$ stands for $x' \in \mathcal{X} \mapsto k(x', x) \in \mathbb{R}$ with $x \in \mathcal{X}$ fixed.

Table 1: Comparison of kernel independence measures: n – number of samples, M – number of components, n' – number of Nyström samples, s – number of random Fourier features, d – data dimensionality.

Independence Measure	Runtime Complexity	M	Domain	Admissible Kernels
V-HSIC [Pfister et al., 2018]	$\mathcal{O}\left(Mn^2\right)$	$M \geqslant 2$	any	universal
NFSIC [Jitkrittum et al., 2017]	$\mathcal{O}\left(n\right)$	$M = 2$	$\mathbb{R}^d$	analytic, characteristic
N-HSIC [Zhang et al., 2017]	$\mathcal{O}\left(n'^3 + nn'^2\right)$	$M = 2$	any	characteristic
RFF-HSIC [Zhang et al., 2017]	$\mathcal{O}\left(s^2 n\right)$	$M = 2$	$\mathbb{R}^d$	translation-invariant, characteristic
KPCC [Huang et al., 2022]	$\mathcal{O}\left(n^3\right)$	$M = 2$	any	characteristic
Nyström M-HSIC (N-MHSIC)	$\mathcal{O}\left(Mn'^3 + Mn'n\right)$	$M \geqslant 2$	any	universal

throughout the paper.[2] The function defined by $\phi_k(x) := k(\cdot, x)$ is the canonical feature map; with this feature map $k(x, x') = \langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}_k} = \langle \phi_k(x), \phi_k(x')\rangle_{\mathcal{H}_k}$ for all $x, x' \in \mathcal{X}$. A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called translation-invariant if there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ such that $k(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. The mean embedding μ_k of a probability measure $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$ is $\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \phi_k(x) \mathrm{d}\mathbb{P}(x)$, where the integral is meant in Bochner's sense. The resulting (semi-)metric is called maximum mean discrepancy (MMD):

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k},$$

for $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$. The injectivity of the mean embedding μ_k is equivalent to MMD_k being a metric; in this case the kernel k is called characteristic. Let $X = (X_m)_{m=1}^M$ denote a random variable with distribution $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$ on the product space $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, where $\mathcal{X}_m$ is enriched with kernel $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$. The distribution of the m-th marginal X_m of X is denoted by $\mathbb{P}_m \in \mathcal{M}_1^+(\mathcal{X}_m)$; the product of these M marginals is $\otimes_{m=1}^M \mathbb{P}_m \in \mathcal{M}_1^+(\mathcal{X})$. The tensor product of the kernels $(k_m)_{m=1}^M$

$$\otimes_{m=1}^M k_m\left((x_m)_{m=1}^M, (x'_m)_{m=1}^M\right) := \prod_{m \in [M]} k_m(x_m, x'_m),$$

with $x_m, x'_m \in \mathcal{X}_m$ ($m \in [M]$), is also a kernel; we will use the shorthand $k = \otimes_{m=1}^M k_m$. The associated RKHS has a simple structure $\mathcal{H}_k = \otimes_{m=1}^M \mathcal{H}_{k_m}$ [Berlinet and Thomas-Agnan, 2004] with the r.h.s. denoting the tensor product of the RKHSs $(\mathcal{H}_{k_m})_{m=1}^M$. Indeed, for $h_m \in \mathcal{H}_{k_m}$, the multi-linear operator $\otimes_{m=1}^M h_m \in \otimes_{m=1}^M \mathcal{H}_{k_m}$ acts as $\otimes_{m=1}^M h_m(v_1, \dots, v_M) = \prod_{m \in [M]} \langle h_m, v_m\rangle_{\mathcal{H}_{k_m}}$, where $h_m, v_m \in \mathcal{H}_{k_m}$. The space $\otimes_{m=1}^M \mathcal{H}_{k_m}$ is the closure of the linear combination of such $\otimes_{m=1}^M h_m$-s:

$$\otimes_{m=1}^M \mathcal{H}_{k_m} = \overline{\operatorname{span}}\left(\otimes_{m=1}^M h_m \,:\, h_m \in \mathcal{H}_{k_m}, m \in [M]\right),$$

where the closure is meant w.r.t. to the (linear extension of the) inner product defined as

$$\begin{aligned}\left\langle \otimes_{m=1}^M a_m, \otimes_{m=1}^M b_m \right\rangle_{\otimes_{m=1}^M \mathcal{H}_{k_m}} &:= \\ &= \prod_{m \in [M]} \langle a_m, b_m\rangle_{\mathcal{H}_{k_m}}, \quad a_m, b_m \in \mathcal{H}_{k_m}. \quad (1)\end{aligned}$$

Specifically, (1) implies that

$$\left\|\otimes_{m=1}^M a_m\right\|_{\otimes_{m=1}^M \mathcal{H}_{k_m}} = \prod_{m \in [M]} \|a_m\|_{\mathcal{H}_{k_m}}. \quad (2)$$

One can define an independence measure, the so-called Hilbert-Schmidt independence criterion based on k as

$$\mathrm{HSIC}_k(\mathbb{P}) := \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right) = \|C_X\|_{\mathcal{H}_k}, \quad (3)$$

where $C_X := \mu_k(\mathbb{P}) - \mu_k\left(\otimes_{m=1}^M \mathbb{P}_m\right)$ is the centered cross-covariance operator.

Let $A : \mathcal{H}_k \to \mathcal{H}_k$ be a bounded linear operator. Its inverse (provided that it exists) $A^{-1} : \mathcal{H}_k \to \mathcal{H}_k$ is also bounded linear. The operator norm of A is defined as $\|A\|_{\mathrm{op}} := \sup_{\|h\|_{\mathcal{H}_k} = 1} \|Ah\|_{\mathcal{H}_k}$. As $\mathcal{H}_k$ is separable, it has a countable orthonormal basis $(e_j)_{j \in J}$. A is called trace-class if $\sum_{j \in J} \left\langle (A^*A)^{\frac{1}{2}} e_j, e_j \right\rangle_{\mathcal{H}_k} < \infty$ where $(\cdot)^*$ denotes the adjoint, and in this case $\mathrm{tr}(A) := \sum_{j \in J} \langle Ae_j, e_j\rangle_{\mathcal{H}_k} < \infty$ is called the trace of A. For $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$, kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\lambda > 0$, the uncentered covariance operator is $\mu_{k \otimes k}(\mathbb{P}) := \int_{\mathcal{X}} k(\cdot, x) \otimes k(\cdot, x) \mathrm{d}\mathbb{P}(x)$ and its regularized variant is $\mu_{k \otimes k, \lambda}(\mathbb{P}) := \mu_{k \otimes k}(\mathbb{P}) + \lambda I$, respectively, where I denotes the identity operator. Let $\mathcal{N}_x(\lambda) = \left\langle \phi_k(x), \mu_{k \otimes k, \lambda}^{-1}(\mathbb{P}) \phi_k(x) \right\rangle_{\mathcal{H}_k}$. The effective dimension of $X \sim \mathbb{P}$ is defined as $\mathcal{N}_X(\lambda) := \mathbb{E}_{x \sim \mathbb{P}}\left[\mathcal{N}_x(\lambda)\right] = \mathrm{tr}\left(\mu_{k \otimes k}(\mathbb{P}) \mu_{k \otimes k, \lambda}^{-1}(\mathbb{P})\right)$. For a sequence of $r_n > 0$-s and a sequence of real-valued random variables X_n, $X_n = \mathcal{O}_{\mathrm{P}}(r_n)$ denotes that $\frac{X_n}{r_n}$ is bounded in probability.

3 EXISTING HSIC ESTIMATORS

We recall the existing HSIC estimator V-HSIC in Section 3.1, and its Nyström approximation for two compo-

[2] The separability of $\mathcal{H}_k$ can be guaranteed on a separable topological space $\mathcal{X}$ by taking a continuous kernel k [Steinwart and Christmann, 2008, Lemma 4.33].

nents in Section 3.2. We present our proposed Nyström approximation for more than two components in Section 4.

3.1 CLASSICAL HSIC ESTIMATOR (V-HSIC)

Given an i.i.d. sample of M-tuples of size n

$$\hat{\mathbb{P}}_n := \left\{\left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right)\right\} \subset \mathcal{X} \tag{4}$$

drawn from $\mathbb{P}$, the corresponding empirical estimate of the squared HSIC, obtained by replacing the population means with the sample means, gives rise to the V-statistic based estimator

$$\begin{aligned} 0 \leqslant \operatorname{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) &:= \frac{1}{n^2}\, \mathbf{1}_n^{\mathsf{T}} \left(\circ_{m\in[M]} \mathbf{K}_{k_m}\right) \mathbf{1}_n \\ &+ \frac{1}{n^{2M}} \prod_{m\in[M]} \mathbf{1}_n^{\mathsf{T}} \mathbf{K}_{k_m} \mathbf{1}_n - \frac{2}{n^{M+1}} \mathbf{1}_n^{\mathsf{T}} \left(\circ_{m\in[M]} \mathbf{K}_{k_m} \mathbf{1}_n\right) \end{aligned} \tag{5}$$

with Gram matrices

$$\mathbf{K}_{k_m} = \left[k_m\left(x_m^i, x_m^j\right)\right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}, \tag{6}$$

which can be computed in $O(n^2 M)$. [3] This prohibitive runtime inspired the development of HSIC approximations [Zhang et al., 2017] using the Nyström method and random Fourier features. We review the Nyström-based construction in Section 3.2 and explain why the technique is restricted to $M = 2$ components, before presenting our alternative approximation scheme of HSIC in Section 4 which is capable of handling $M \geqslant 2$ components.

3.2 NYSTRÖM METHOD

In this section, we recall the existing Nyström approximation, which can handle $M = 2$ components.

The expression (5) can be rewritten [Gretton et al., 2005] for $M = 2$ components as

$$\operatorname{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2} \operatorname{tr}\left(\mathbf{H}\mathbf{K}_{k_1}\mathbf{H}\mathbf{K}_{k_2}\right), \tag{7}$$

with the centering matrix $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}} \in \mathbb{R}^{n\times n}$, Gram matrices $\mathbf{K}_{k_1}$, $\mathbf{K}_{k_2}$ defined in (6), and sample $\hat{\mathbb{P}}_n := \left\{(x_1^1, x_2^1), \ldots, (x_1^n, x_2^n)\right\}$ as in (4) with $M = 2$. The naive computation of (7) costs $\mathcal{O}\left(n^3\right)$. However, noticing that $\operatorname{tr}(\mathbf{A}^{\mathsf{T}}\mathbf{B}) = \sum_{i,j\in[n]} A_{i,j} B_{i,j}$, the computational complexity reduces to $\mathcal{O}\left(n^2\right)$. The quadratic complexity can be further reduced by the Nyström approximation[3] [Zhang et al., 2017]

$$\begin{aligned} \operatorname{HSIC}_{k,\mathrm{N}_0}^2\left(\hat{\mathbb{P}}_n\right) &= \frac{1}{n^2} \operatorname{tr}\left(\mathbf{H}\mathbf{K}_{k_1}^{\mathrm{Nys}}\mathbf{H}\mathbf{K}_{k_2}^{\mathrm{Nys}}\right) \\ &\overset{(*)}{=} \frac{1}{n^2} \left\| \left(\mathbf{H}\phi_{k_1}^{\mathrm{Nys}}\right)^{\mathsf{T}} \mathbf{H}\phi_{k_2}^{\mathrm{Nys}} \right\|_{\mathsf{F}}^2, \end{aligned} \tag{8}$$

which we detail in the following. The Nyström approximation relies on a subsample of size $n' \leqslant n$ of $\hat{\mathbb{P}}_n$, which we denote by $\tilde{\mathbb{P}}_{n'} := \left\{\left(\tilde{x}_1^1, \tilde{x}_2^1\right), \ldots, \left(\tilde{x}_1^{n'}, \tilde{x}_2^{n'}\right)\right\}$; the tilde indicates a relabeling. The subsample allows to define three matrices

$$\begin{aligned} \mathbf{K}_{k_m,n'n'} &= \left[k_m\left(\tilde{x}_m^i, \tilde{x}_m^j\right)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}, \\ \mathbf{K}_{k_m,nn} &= \mathbf{K}_{k_m} \in \mathbb{R}^{n\times n}, \\ \mathbf{K}_{k_m,n'n} &= \left[k_m(\tilde{x}_m^i, x_m^j)\right]_{i\in[n'],j\in[n]} \in \mathbb{R}^{n'\times n}, \end{aligned} \tag{9}$$

where $m \in [2]$ and $\mathbf{K}_{k_m}$ is defined in (6), and let $\mathbf{K}_{k_m,nn'} = \mathbf{K}_{k_m,n'n}^{\mathsf{T}} \in \mathbb{R}^{n\times n'}$. The matrices $\mathbf{K}_{k_m}^{\mathrm{Nys}}$ ($m \in [2]$) as used in (8) are

$$\begin{aligned} \mathbf{K}_{k_m}^{\mathrm{Nys}} &:= \mathbf{K}_{k_m,nn'}\mathbf{K}_{k_m,n'n'}^{-1}\mathbf{K}_{k_m,n'n} \\ &= \underbrace{\mathbf{K}_{k_m,nn'}\mathbf{K}_{k_m,n'n'}^{-\frac{1}{2}}}_{=:\phi_{k_m}^{\mathrm{Nys}} \in\mathbb{R}^{n\times n'}} \Big(\underbrace{\mathbf{K}_{k_m,nn'}\mathbf{K}_{k_m,n'n'}^{-\frac{1}{2}}}_{\phi_{k_m}^{\mathrm{Nys}}}\Big)^{\mathsf{T}} \in \mathbb{R}^{n\times n}, \end{aligned}$$

provided that the inverse $\mathbf{K}_{k_m,n'n'}^{-1}$ exists. In (8) the r.h.s. of $(*)$ has a computational complexity of $\mathcal{O}(n'^3 + nn'^2)$,[4] which is smaller than $\mathcal{O}\left(n^2\right)$ of (7), provided that $n' < \sqrt{n}$; this speeds up the computation. $(*)$ relies on the cyclic invariance property of the trace, and the idempotence of $\mathbf{H}$ (in other words, $\mathbf{H}\mathbf{H} = \mathbf{H}$), limiting the above derivation to $M = 2$ components; the approach does not extend naturally to the case of $M > 2$.

4 PROPOSED HSIC ESTIMATOR

We now elaborate the proposed Nyström HSIC approximation for $M \geqslant 2$ components.

Recall that the centered cross-covariance operator takes the form

$$\begin{aligned} C_X &= \mu_k(\mathbb{P}) - \mu_k\left(\otimes_{m=1}^M \mathbb{P}_m\right) \\ &= \mu_k(\mathbb{P}) - \otimes_{m=1}^M \mu_{k_m}\left(\mathbb{P}_m\right). \end{aligned} \tag{10}$$

There are $M+1$ expectations in this expression; we estimate these mean embeddings separately. This conceptually simple construction, is to the best of our knowledge, the first that handles $M \geqslant 2$ components, and it allows to leverage recent bounds on mean estimators (Lemma 4.1). We first detail the

[3] $\operatorname{HSIC}_k^2(\hat{\mathbb{P}}_n)$ denotes the application of HSIC_k^2 to the empirical measure $\hat{\mathbb{P}}_n$. $\operatorname{HSIC}_{k,\mathrm{N}_0}^2(\hat{\mathbb{P}}_n)$ and $\operatorname{HSIC}_{k,\mathrm{N}}^2(\hat{\mathbb{P}}_n)$ indicate dependence on $\hat{\mathbb{P}}_n$. Similarly, $\mu_\ell(\hat{\mathbb{Q}}_n)$ stands for application, $\mu_\ell(\tilde{\mathbb{Q}}_{n'})$, $\mu_{k_m}(\tilde{\mathbb{P}}_{m,n'})$ and $\mu_k(\tilde{\mathbb{P}}_{n'})$ indicate dependence on the argument.

[4] This follows from the complexity of $O(n'^3)$ of inverting an $n' \times n'$ matrix and the complexity of multiplying both feature representations [Zhang et al., 2017].

general Nyström method for approximating expectations $\int_{\mathcal{Y}} \phi_\ell(y)\mathrm{d}\mathbb{Q}(y)$ associated to a kernel $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and probability distribution $\mathbb{Q} \in \mathcal{M}_1^+(\mathcal{Y})$. One can then choose

$$\begin{aligned} (\mathcal{Y}, \ell, \mathbb{Q}) &= (\mathcal{X}, k, \mathbb{P}), \text{ and} \\ (\mathcal{Y}, \ell, \mathbb{Q}) &= (\mathcal{X}_m, k_m, \mathbb{P}_m), \quad m \in [M], \end{aligned} \tag{11}$$

to achieve our goal.

Let $\tilde{\mathbb{Q}}_{n'} = \left\{\tilde{y}^1, \dots, \tilde{y}^{n'}\right\}$ be a subsample (with replacement) of $\hat{\mathbb{Q}}_n = \left\{y^1, \dots, y^n\right\} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$ referred to as Nyström points; the tilde again indicates relabeling. The usual estimator of the mean embedding replaces the population mean with its empirical counterpart over n samples[3]

$$\mu_\ell(\mathbb{Q}) = \int_{\mathcal{Y}} \phi_\ell(y)\mathrm{d}\mathbb{Q}(y) \approx \frac{1}{n}\sum_{i\in[n]} \phi_\ell(y^i) = \mu_\ell(\hat{\mathbb{Q}}_n).$$

Instead, the Nyström approximation uses a weighted sum with weights $\alpha_i \in \mathbb{R}$ ($i \in [n']$): given n' Nyström points, the estimator takes the form[3]

$$\mu_\ell(\mathbb{Q}) \approx \sum_{i\in[n']} \alpha_i \phi_\ell(\tilde{y}^i) = \mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) \in \mathcal{H}_\ell^{\mathrm{Nys}},$$

where $\mathcal{H}_\ell^{\mathrm{Nys}} := \operatorname{span}\left(\phi_\ell\big(\tilde{y}^i\big) \,:\, i \in [n']\right) \subset \mathcal{H}_\ell$. The coefficients $\boldsymbol{\alpha}_\ell = (\alpha_\ell^1, \dots, \alpha_\ell^{n'})^\mathsf{T} \in \mathbb{R}^{n'}$ are obtained by the minimum norm solution of

$$\min_{\boldsymbol{\alpha}_\ell \in \mathbb{R}^{n'}} \left\| \mu_\ell\left(\hat{\mathbb{Q}}_n\right) - \sum_{i\in[n']} \alpha_i \phi_\ell\left(\tilde{y}^i\right) \right\|_{\mathcal{H}_\ell}^2 . \tag{12}$$

The following lemma describes the solution of (12).

Lemma 4.1 (Nyström mean embedding, Chatalic et al. [2022])**.** *For a kernel ℓ with corresponding feature map ϕ_ℓ, an i.i.d. sample $\hat{\mathbb{Q}}_n$ of distribution $\mathbb{Q}$, and a subsample $\tilde{\mathbb{Q}}_{n'}$ of $\hat{\mathbb{Q}}_n$, the Nyström estimate of $\mu_\ell(\mathbb{Q})$ is given by*

$$\begin{aligned} \mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) &= \sum_{i\in[n']} \alpha_\ell^i \phi_\ell\left(\tilde{y}^i\right), \\ \boldsymbol{\alpha}_\ell &= \frac{1}{n}\left(\mathbf{K}_{\ell,n'n'}\right)^- \mathbf{K}_{\ell,n'n}\mathbf{1}_n, \end{aligned} \tag{13}$$

with Gram matrix $\mathbf{K}_{\ell,n'n'} = \left[\ell(\tilde{x}^i, \tilde{x}^j)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}$, and $\mathbf{K}_{\ell,n'n} = \left[\ell(\tilde{x}^i, x^j)\right]_{i\in[n'],j\in[n]} \in \mathbb{R}^{n'\times n}$.

Let

$$\tilde{\mathbb{P}}_{n'} = \left\{\left(\tilde{x}_1^1, \dots, \tilde{x}_M^1\right), \dots, \left(\tilde{x}_1^{n'}, \dots, \tilde{x}_M^{n'}\right)\right\} \tag{14}$$

be a subsample (with replacement) of $\hat{\mathbb{P}}_n = \left\{\left(x_1^1, \dots, x_M^1\right), \dots, \left(x_1^n, \dots, x_M^n\right)\right\}$ defined in (4), and

$$\tilde{\mathbb{P}}_{m,n'} = \left\{\tilde{x}_m^1, \dots, \tilde{x}_m^{n'}\right\} \tag{15}$$

be the corresponding subsample of the m-th marginal ($m \in [M]$). Using our choice (11) with Lemma 4.1, the estimators for the embeddings of marginal distributions take the form[3]

$$\begin{aligned} \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right) &= \sum_{i\in[n']} \alpha_{k_m}^i \phi_{k_m}\left(\tilde{x}_m^i\right), \\ \boldsymbol{\alpha}_{k_m} &= \frac{1}{n}\left(\mathbf{K}_{k_m,n'n'}\right)^- \mathbf{K}_{k_m,n'n}\mathbf{1}_n, \end{aligned} \tag{16}$$

and the estimator of the mean embedding of the joint distribution is[3]

$$\begin{aligned} \mu_k\left(\tilde{\mathbb{P}}_{n'}\right) &= \sum_{i\in[n']} \alpha_k^i \otimes_{m=1}^M \phi_{k_m}\left(\tilde{x}_m^i\right), \\ \boldsymbol{\alpha}_k &= \frac{1}{n}\left(\mathbf{K}_{k,n'n'}\right)^- \left(\mathbf{K}_{k,n'n}\right)\mathbf{1}_n \\ &\overset{(*)}{=} \frac{1}{n} \overbrace{\Big(\underbrace{\circ_{m\in[M]}\mathbf{K}_{k_m,n'n'}}_{(a)}\Big)^-}^{(c)} \times \\ &\qquad \Big(\underbrace{\circ_{m\in[M]}\mathbf{K}_{k_m,n'n}}_{(b)}\Big)\mathbf{1}_n, \end{aligned} \tag{17}$$

where $(*)$ holds as for the Gram matrix $\mathbf{K}_{k,n'n'}$ associated with the product kernel $k = \otimes_{m\in[M]} k_m$ one has

$$\begin{aligned} \mathbf{K}_{k,n'n'} &= \left[k\left((x_1^i, \dots, x_M^i), (x_1^j, \dots, x_M^j)\right)\right]_{i,j\in[n']} \\ &= \left[\prod_{m\in[M]} k_m(x_m^i, x_m^j)\right]_{i,j\in[n']} = \circ_{m\in[M]}\mathbf{K}_{k_m,n'n'}, \end{aligned}$$

and similarly $\mathbf{K}_{k,n'n} = \circ_{m\in[M]}\mathbf{K}_{k_m,n'n}$, with $\mathbf{K}_{k_m,n'n'}$ and $\mathbf{K}_{k_m,n'n}$ defined in (9).

Combining the $M + 1$ Nyström estimators in (16) and in (17) gives rise to the overall Nyström HSIC estimator, which is elaborated in the following lemma.

Lemma 4.2 (Computation of Nyström M-HSIC)**.** *The Nyström estimator for HSIC can be expressed as*[3]

$$\begin{aligned} &\mathrm{HSIC}_{k,N}^2\left(\hat{\mathbb{P}}_n\right) = \boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m,n'n'}\right)\boldsymbol{\alpha}_k \\ &+ \prod_{m\in[M]} \boldsymbol{\alpha}_{k_m}^\mathsf{T}\mathbf{K}_{k_m,n'n'}\boldsymbol{\alpha}_{k_m} - 2\boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m,n'n'}\boldsymbol{\alpha}_{k_m}\right), \end{aligned} \tag{18}$$

with $\boldsymbol{\alpha}_{k_m}$ and $\boldsymbol{\alpha}_k$ defined in (16) *and* (17)*, respectively, $\mathbf{K}_{k_m,n'n'}$ is defined in* (9)*, and N in the subscript of the estimator refers to Nyström. Note that* (18) *depends on $\hat{\mathbb{P}}_n$ as one must solve* (12)*.*

Remark 1.

- ***Uniform weights, no subsampling.*** *The estimator* (18) *gives back* (5) *when $\boldsymbol{\alpha}_k := \boldsymbol{\alpha}_{k_m} := \frac{1}{n}\mathbf{1}_n$ for all $m \in [M]$, and when there is no subsampling applied.*

- ***Runtime complexity.*** *In order to determine the computational complexity of* (18) *one has to find that of* (17)*; that of* (16) *follows by choosing* $M = 1$ *in* (17). (a) *and* (b) *in* (17) *are Hadamard products; hence their computational complexity is* $\mathcal{O}\left(Mn'^2\right)$ *and* $\mathcal{O}\left(Mnn'\right)$. (c) *in* (17) *is the Moore-Penrose inverse of an* $n' \times n'$ *matrix; thus its complexity is* $\mathcal{O}\left(n'^3\right)$. *Hence, the computation of* $\boldsymbol{\alpha}_k$ *costs* $\mathcal{O}\big(Mn'^2 + n'^3 + Mn'n\big)$, *and that of* $(\boldsymbol{\alpha}_{k_m})_{m=1}^M$ *is* $\mathcal{O}\big(n'^2 + n'^3 + n'n\big)$ *for each* $m \in [M]$. *In* (18) *each term can be computed in* $\mathcal{O}\left(Mn'^2\right)$. *Overall the Nyström* M*-HSIC estimator has complexity* $\mathcal{O}\big(Mn'^2 + Mn'^3 + Mn'n\big) = \mathcal{O}\big(Mn'^3 + Mn'n\big)$.

- ***Difference compared to the estimator by Zhang et al. [2017].*** *For* $M = 2$, (18) *reduces to*

$$\mathrm{HSIC}^2_{k,N}\left(\hat{\mathbb{P}}_n\right) = \boldsymbol{\alpha}_k^\top \left(\mathbf{K}_{k_1} \circ \mathbf{K}_{k_2}\right) \boldsymbol{\alpha}_k \qquad (19)$$
$$+ \prod_{i \in [2]} \alpha_{k_i}^\top \mathbf{K}_{k_i} \boldsymbol{\alpha}_{k_i} - 2\boldsymbol{\alpha}_k^\top \left(\mathbf{K}_{k_1}\boldsymbol{\alpha}_{k_1} \circ \mathbf{K}_{k_2}\boldsymbol{\alpha}_{k_2}\right).$$

Using the equivalence of (5) *and* (7) *in case* $M = 2$ *gives*

$$\mathrm{tr}\left(\mathbf{H}\mathbf{K}_{k_1}\mathbf{H}\mathbf{K}_{k_2}\right) = \frac{1}{n^2}\mathbf{1}_n^\top \left(\mathbf{K}_{k_1} \circ \mathbf{K}_{k_2}\right)\mathbf{1}_n$$
$$+ \frac{1}{n^4}\prod_{i \in [2]} \mathbf{1}_n^\top \mathbf{K}_{k_i}\mathbf{1}_n - \frac{2}{n^3}\mathbf{1}_n^\top \left(\mathbf{K}_{k_1}\mathbf{1}_n \circ \mathbf{K}_{k_2}\mathbf{1}_n\right),$$

hence (8) *becomes*

$$\mathrm{HSIC}^2_{k,N_0}\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\mathbf{1}_n^\top \left(\mathbf{K}_{k_1}^{Nys} \circ \mathbf{K}_{k_2}^{Nys}\right)\mathbf{1}_n \qquad (20)$$
$$+ \frac{1}{n^4}\prod_{i \in [2]} \mathbf{1}_n^\top \mathbf{K}_{k_i}^{Nys}\mathbf{1}_n - \frac{2}{n^3}\mathbf{1}_n^\top \left(\mathbf{K}_{k_1}^{Nys}\mathbf{1}_n \circ \mathbf{K}_{k_2}^{Nys}\mathbf{1}_n\right).$$

The estimators (19) *and* (20) *are identical if* $\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k_m} = \frac{1}{n}\mathbf{1}_n$ *for all* $m \in [M]$ *and when there is no subsampling; in the general case they do not coincide. In* (8) *the dominant term in the complexity is* $(n')^2\, n$ *(since* $n' < n$*), this reduces to* $n'n$ *in our proposed estimator* (18).

Key to showing the consistency of the proposed Nyström M-HSIC estimator (18) (Proposition 4.1) is our next lemma, which describes how the Nyström approximation error of the mean embeddings of the components (d_{k_m} below) can be propagated through tensor products.

Lemma 4.3 (Error propagation on tensor products)**.** *Let* $X = (X_m)_{m=1}^M \in \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ *bounded kernels* ($\exists a_{k_m} \in (0, \infty)$ *such that* $\sup_{x_m \in \mathcal{X}_m} \sqrt{k_m(x_m, x_m)} \leqslant a_{k_m}$, $m \in [M]$*)*, $k = \otimes_{m=1}^M k_m$, $\mathcal{H}_k$ *the RKHS associated to* k, $X \sim \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$, $\mathbb{P}_m$ *the* m*-th marginal of* $\mathbb{P}$ ($m \in [M]$)*,* $n' \leqslant n$, *and* $\tilde{\mathbb{P}}_{m,n'}$ *defined according to* (15)*. Then*

$$\left\|\otimes_{m=1}^M \mu_{k_m}\left(\mathbb{P}_m\right) - \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_k} \leqslant$$
$$\leqslant \prod_{m \in [M]} \left(a_{k_m} + d_{k_m}\right) - \prod_{m \in [M]} a_{k_m},$$

where $d_{k_m} = \left\|\mu_{k_m}\left(\mathbb{P}_m\right) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right)\right\|_{\mathcal{H}_{k_m}}$.

Our resulting Nyström M-HSIC performance guarantee is as follows.

Proposition 4.1 (Error bound for Nyström M-HSIC)**.** *Let* $X = (X_m)_{m=1}^M \in \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $X \sim \mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$, $(\mathcal{X}_m)_{m \in [M]}$ *locally compact, second-countable topological spaces,* $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ *bounded kernels, i.e.,* $\exists a_{k_m} \in (0, \infty)$ *such that* $\sup_{x_m \in \mathcal{X}_m} \sqrt{k_m(x_m, x_m)} \leqslant a_{k_m}$ *for all* $m \in [M]$, $k = \otimes_{m \in [M]} k_m$, $a_k = \prod_{m=1}^M a_{k_m}$, $\phi_{k_m}(x_m) = k_m(\cdot, x_m)$ *for all* $x_m \in \mathcal{X}_m$, $\phi_k = \otimes_{m=1}^M \phi_{k_m}$, $C_k = \mathbb{E}\left[\phi_k(X) \otimes \phi_k(X)\right]$, $C_{k_m} = \mathbb{E}\left[\phi_{k_m}(X_m) \otimes \phi_{k_m}(X_m)\right]$, *the number of Nyström points* $n' \leqslant n$, $\hat{\mathbb{P}}_n$ *defined according to* (4)*. Then, for any* $\delta \in \left(0, \frac{1}{M+1}\right)$

$$\left|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_{k,N}\left(\hat{\mathbb{P}}_n\right)\right| \leqslant \underbrace{\frac{c_{k,1}}{\sqrt{n}}}_{t_{k,1}} + \underbrace{\frac{c_{k,2}}{n'}}_{t_{k,2}} +$$
$$+ \underbrace{\frac{c_{k,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_X\left(\frac{12a_k^2\log(n'/\delta)}{n'}\right)}}_{t_{k,3}} +$$
$$+ \prod_{m \in [M]}\Bigg[a_{k_m} + \underbrace{\frac{c_{k_m,1}}{\sqrt{n}}}_{t_{k_m,1}} + \underbrace{\frac{c_{k_m,2}}{n'}}_{t_{k_m,2}} +$$
$$+ \underbrace{\frac{c_{k_m,3}\sqrt{\log(n'/\delta)}}{n'}\sqrt{\mathcal{N}_{X_m}\left(\frac{12a_{k_m}^2\log(n'/\delta)}{n'}\right)}}_{t_{k_m,3}}\Bigg]$$
$$- \prod_{m \in [M]} a_{k_m}$$

holds with probability at least $1 - (M+1)\delta$*, provided that*

$$n' \geqslant \max_{m \in [M]}\left(67, 12a_k^2\left\|C_k\right\|_{\mathrm{op}}^{-1}, 12a_{k_m}^2\left\|C_{k_m}\right\|_{\mathrm{op}}^{-1}\right)\log\frac{n'}{\delta},$$

where $c_{k,1} = 2a_k\sqrt{2\log(6/\delta)}$, $c_{k,2} = 4\sqrt{3}a_k\log(12/\delta)$, $c_{k,3} = 12\sqrt{3\log(12/\delta)}a_k$, $c_{k_m,1} = 2a_{k_m}\sqrt{2\log(6/\delta)}$, $c_{k_m,2} = 4\sqrt{3}a_{k_m}\log(12/\delta)$, $c_{k_m,3} = 12\sqrt{3\log(12/\delta)}a_{k_m}$ *for* $m \in [M]$.

As a baseline, to interpret the result (see the second bullet point in Remark 2), one could consider the V-statistic based HSIC estimator (5) for $M \geqslant 2$, which according to our following lemma has a convergence rate of $\mathcal{O}_{\mathrm{P}}\left(\frac{1}{\sqrt{n}}\right)$.

Lemma 4.4 (Deviation bound for V-statistic based HSIC estimator)**.** *Let* $\mathrm{HSIC}_k(\hat{\mathbb{P}}_n)$ *be as in* (5) *on a metric space* $\mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, *and* $\mathrm{HSIC}_k(\mathbb{P}) > 0$. *Then*

$$\left|\mathrm{HSIC}_k(\mathbb{P}) - \mathrm{HSIC}_k\left(\hat{\mathbb{P}}_n\right)\right| = \mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right).$$

Remark 2.

- *From the terms $t_{k,1}, t_{k,2}, t_{k_m,1}, t_{k_m,2}, m \in [M]$ it follows that for $n' < \sqrt{n}$ the respective second term dominates, thus increasing the error; for $n' > \sqrt{n}$ the respective first term dominates and the computational complexity increases. The effective dimension $(t_{k,3}, t_{k_m,3})$ controls the trade off between the two terms and can be related [Chatalic et al., 2022] to the decay of the eigenvalues of the respective covariance operator. A convergence rate of $n^{-1/2}$ for the sums $t_{k,1} + t_{k,2} + t_{k,3}$ and $t_{k_m,1} + t_{k_m,2} + t_{k_m,3}$ can be achieved if*
  - $\max_{m \in [M]} \left(\mathcal{N}_X(\lambda), \mathcal{N}_{X_m}(\lambda)\right) \leqslant c\lambda^{-\gamma}$ *for some* $c > 0$ *and* $\gamma \in (0, 1]$ *with* $n' = n^{1/(2-\gamma)} \log(n/\delta)$*, or*
  - $\max_{m \in [M]} \left(\mathcal{N}_X(\lambda), \mathcal{N}_{X_m}(\lambda)\right) \leqslant \log(1 + c/\lambda)/\beta$ *for some* $c > 0$*,* $\beta > 0$*, and* $n' = \sqrt{n} \log\left(\sqrt{n} \max_{m \in [M]} \left(\frac{1}{\delta}, \frac{c}{6a_k^2}, \frac{c}{6a_{k_m}^2}\right)\right)$.

  This rate of convergence propagates through the product.
- *Lemma 4.4 establishes that the V-statistic based estimator of HSIC converges with rate $n^{-1/2}$. Hence, setting $n' \sim \sqrt{n}$ in Nyström M-HSIC allows to obtain the same rate of convergence while decreasing runtime. Assumption* $\text{HSIC}_k(\mathbb{P}) > 0$ *in Lemma 4.4 protects one from attaining a convergence rate of n^{-1} of* $\text{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right)$.

5 EXPERIMENTS

In this section, we demonstrate the efficiency of the proposed method (N-MHSIC) against the baselines NFSIC, RFF-HSIC, N-HSIC and the quadratic-time V-statistic based HSIC estimator (V-HSIC) in the context of independence testing. Hence, the null hypothesis H_0 is that the joint distribution factorizes to the product of the marginals, the alternative H_1 is that this is not the case. The experiments study both synthetic (Section 5.1) and real-world (Section 5.2) examples, in terms of power and runtime.[5]

We use the Gaussian kernel $k_m(\mathbf{x}_m, \mathbf{x}'_m) = \exp\left(-\gamma_{k_m} \|\mathbf{x}_m - \mathbf{x}'_m\|_2^2\right)$ for all experiments, with γ_{k_m} chosen according to the median heuristic. For a fair comparison of the test power, we approximate the null distribution of each test statistic by the permutation approach with 250 samples. We then perform a one-sided test with an acceptance region of 5% ($\alpha = 0.05$), which we repeat, for all power experiments, on 100 independent draws of the data; the runtime results include these. We set each algorithm's parameters as recommended by the respective authors: For NFSIC, we set the number of test locations $J = 5$; the number of Fourier features (RFF-HSIC) and Nyström samples (N-HSIC) is set to $\sqrt{n}$. The number of Nyström samples of N-MHSIC is indicated within the experiment description. The opaque area in the figures indicates the 0.95-quantile obtained over 5 runs. All experiments were performed on a PC with Ubuntu 20.04, 124GB RAM, and 32 cores with 2GHz each.

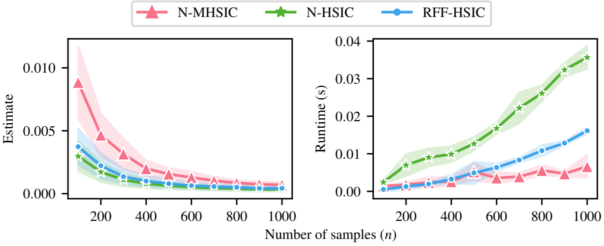

Figure 1: Estimation accuracy for $M = 2$ components; the theoretical HSIC value is zero.

5.1 SYNTHETIC DATA

We examine three toy problems in the following, illustrating runtime and statistical power.

Comparison of HSIC approximations under H_0. First, for $M = 2$ components, we compare our proposed method to the existing accelerated HSIC estimators (N-HSIC, RFF-HSIC) on independent data to assess convergence w.r.t. runtime. Specifically, we set $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. The theoretical value of HSIC is thus zero. Figure 1 shows the estimates for sample sizes from 100 to 1000; the number of Nyström samples for N-MHSIC is set to $n' = 2\sqrt{n}$. All approaches converge to zero, with N-MHSIC converging a bit slower than the exisiting HSIC approximations. However, we note that the gap is on the order of 10^{-3} so it is close to the theoretical value also for small sample sizes. The runtime scales as predicted by the complexity analysis, with the proposed approach running faster than both N-HSIC and RFF-HSIC starting from $n = 500$ samples.

Dependent Data (H_1 holds). To evaluate the statistical power on $M = 2$ components, we set $X_1 \sim \mathcal{N}(0, 1)$, $X_2 = X_1 + \epsilon$, and $\epsilon \sim \mathcal{N}(0, 1)$, with n' set as before. Figure 2 shows that N-MHSIC achieves a power of one for $n \approx 100$ and that it is slightly worse than the existing HSIC approximations for small sample sizes. V-HSIC has the highest power but also the highest runtime. Even though NFSIC has linear runtime complexity it is slower than all other statistics on small sample sizes.

Causal Discovery. The experiments until now considered $M = 2$ components. However, N-MHSIC allows for handling $M \geqslant 2$ components and thus can estimate the directed acyclic graph (DAG) governing causality if one assumes an additive noise model.

Specifically, we sample from the structural equations $X_i = \sum_{j \in \text{PA}_i} f^{i,j}(X_j) + \epsilon_i$ for $i \in [d]$, of a randomly selected

[5] The code of our experiments is available in the supplement.

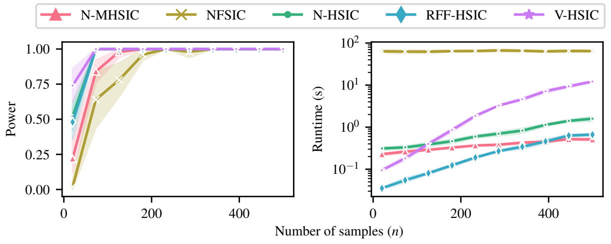

Figure 2: Power on dependent data. Runtime on log scale.

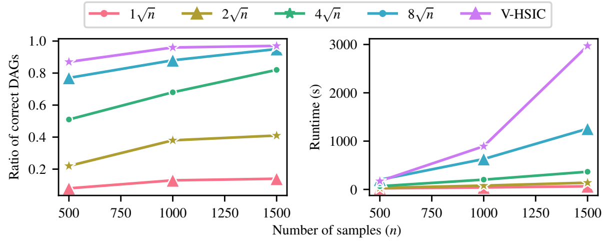

Figure 3: Ratio of correctly identified DAGs with 4 nodes.

fully connected DAG with four nodes ($d = 4$), of which there are 24. In the equation, PA_i denotes the parents of i in the associated DAG, and the ϵ_i are normally distributed and jointly independent, with a variance sampled independently from the uniform distribution $\mathcal{U}\left(1, \sqrt{2}\right)$.

To now test whether a particular DAG fits the data, Pfister et al. [2018] propose to use generalized additive model regression to find the residuals when regressing each node onto all its parents and to reject the DAG if the residuals are not jointly independent. If these are independent, we accept the causal structure. In this application, one is only interested in the relative p-values when performing the procedure for all possible DAGs with the correct number of nodes.

V-HSIC has the best performance in [Pfister et al., 2018], so we only compare against V-HSIC; it is also the only other approach which allows testing joint independence of more than two components. Figure 3 shows how often N-MHSIC and V-HSIC identify the correct DAG in 100 samples. V-HSIC has higher power than N-MHSIC and more often identifies the correct DAG for small sample sizes. However, as the r.h.s. of Figure 3 shows, the proposed algorithm runs even for $n' = 8\sqrt{n}$ and $n = 1500$ twice as fast as V-HSIC while producing the same result quality. Due to their different runtime complexities, the gap in runtime widens further with increasing sample size.

5.2 REAL-WORLD DATA

This section is dedicated to benchmarks on real-world data.

Million Song Data. The Million Song Data [Bertin-Mahieux et al., 2011] contains approximately 500,000 songs.

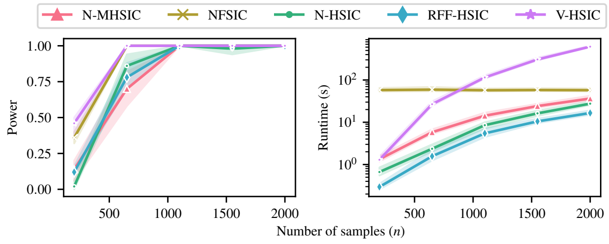

Figure 4: Test power vs. runtime on the Million Song Data.

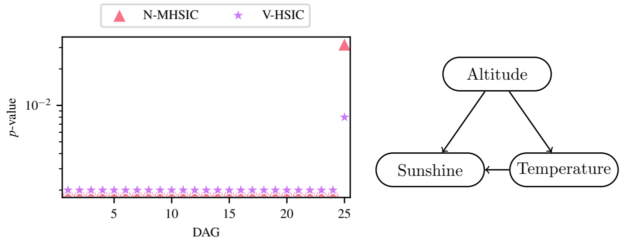

Figure 5: Testing for joint independence on the residuals of DAGs with three nodes (left) and the DAG with the largest p-value (right). The p-values agree on DAGs 1 to 24.

Each has 90 features (X) together with its year of release, which ranges from 1922 to 2011 (Y). The algorithms must detect the dependence between the features and the year of release. To approximate the power, we draw 100 independent samples of the whole data set. Figure 4 shows the results, for level $\alpha = 0.01$. In contrast to a similar experiment of Jitkrittum et al. [2017], we use a permutation approach for all two-sample tests and increase the number of Nyström samples (random Fourier features) as a function of n, obtaining higher power throughout. The problem is sufficiently challenging, so that we set the number of Nyström samples to $8\sqrt{n}$ for N-MHSIC. V-HSIC and NFSIC achieve maximum power from $n = 650$. N-MHSIC features similar runtime and power as the existing HSIC approximations N-HSIC and RFF-HSIC but can handle more than two components.

Weather Causal Discovery. Here, we aim to infer the correct causality DAG from real-world data, namely the data set of Mooij et al. [2016] which contains 349 measurements consisting of altitude, temperature and sunshine. The goal is to infer the most plausible DAG with three nodes ($d = 3$) out of the 25 possible DAGs ($3^3 - 2 = 25$; two graphs have a cycle). We assume the structural equations discussed before. Figure 5 shows the p-values with the estimated DAG (with index 25) having the largest p-value. Again, we compare our results to V-HSIC and find that both successfully identify the most plausible DAG [Pfister et al., 2018].

These experiments demonstrate the efficiency of the proposed Nyström M-HSIC method.

References

Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879, 2022.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–596, 2011.

Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Riec. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712, 2020.

Dimitri Bouche, Rémi Flamary, Florence d'Alché Buc, Riwal Plougonven, Marianne Clausel, Jordi Badosa, and Philippe Drobinski. Wind power predictions from nowcasts to 4-hour forecasts: a learning approach with variable selection. Technical report, 2022. (`https://arxiv.org/abs/2204.09362`).

Gustavo Camps-Valls, Joris M. Mooij, and Bernhard Schölkopf. Remote sensing feature selection by kernel dependence measures. *IEEE Geoscience and Remote Sensing Letters*, 7(3):587–591, 2010.

Shubhadeep Chakraborty and Xianyang Zhang. Distance metrics for measuring joint dependence with application to causal inference. *Journal of the American Statistical Association*, 114(528):1638–1650, 2019.

Antoine Chatalic, Nicolas Schreuder, Alessandro Rudi, and Lorenzo Rosasco. Nyström kernel mean embeddings. In *International Conference on Machine Learning (ICML)*, pages 3006–3024, 2022.

Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1972–1980, 2015.

Héctor Climente-González, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. *Bioinformatics*, 35(14):i427–i435, 2019.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496, 2008.

Thomas Gärtner, Peter Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *International Conference on Machine Learning (ICML)*, pages 179–186, 2002.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–78, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592, 2008.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, pages 723–773, 2012.

Jorge Guevara, Roberto Hirata, and Stéphane Canu. Cross product kernels for fuzzy set similarity. In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, 2017.

David Haussler. Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz, 1999. (`http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf`).

Zhen Huang, Nabarun Deb, and Bodhisattva Sen. Kernel partial correlation coefficient – a measure of conditional dependence. *Journal of Machine Learning Research*, 23 (216):1–58, 2022.

Yunlong Jiao and Jean-Philippe Vert. The Kendall and Mallows kernels for permutations. In *International Conference on Machine Learning (ICML)*, pages 2982–2990, 2016.

Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning (ICML)*, pages 1742–1751, 2017.

Franz J. Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20:1–45, 2019.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2: 419–444, 2002.

Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41:3284–3305, 2013.

Charles Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7: 2651–2667, 2006.

Joris Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 5–31, 2018.

Novi Quadrianto, Le Song, and Alex Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296, 2009.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184, 2007.

Bernhard Schölkopf and Alexander Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Dino Sejdinovic, Arthur Gretton, and Wicher Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132, 2013a.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013b.

Tianhong Sheng and Bharath K. Sriperumbudur. On distance and kernel measures of conditional independence. *Journal of Machine Learning Research*, 24(7): 1–16, 2023.

Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

Le Song, Alexander J. Smola, Arthur Gretton, and Karsten M. Borgwardt. A dependence maximization view of clustering. In *International Conference on Machine Learning (ICML)*, pages 815—-822, 2007.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, pages 1517–1561, 2010.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, pages 67–93, 2001.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3:1236–1265, 2009.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794, 2007.

Andi Wang, Juan Du, Xi Zhang, and Jianjun Shi. Ranking features to promote diversity: An approach based on sparse distance correlation. *Technometrics*, 64(3):384–395, 2022.

Chris Watkins. Dynamic alignment kernels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50, 1999.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 682–688, 2001.

Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, pages 1–18, 2017.

V. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.