Conditional Dependence via U-Statistics Pruning

Ferran de Cabrera ⬤, Marc Vilà-Insa ⬤, *Graduate Student Member, IEEE*, and Jaume Riba ⬤, *Senior Member, IEEE*

Abstract—The problem of measuring conditional dependence between two random phenomena arises when a third one (a confounder) has a potential influence on the amount of information between them. A typical issue in this challenging problem is the inversion of ill-conditioned autocorrelation matrices. This paper presents a novel measure of conditional dependence based on the use of *incomplete* unbiased statistics of degree two, which allows to re-interpret independence as uncorrelatedness on a finite-dimensional feature space. This formulation enables to prune data according to observations of the confounder itself, thus avoiding matrix inversions altogether. The proposed approach is articulated as an extension of the Hilbert-Schmidt independence criterion, which becomes expressible through kernels that operate on 4-tuples of data.

Index Terms—Hilbert-Schmidt independence criterion (HSIC), kernel methods, conditional dependence, U-Statistics.

I. INTRODUCTION

CONDITIONAL dependence becomes relevant when a third phenomenon might explain, mediate, or confound the apparent relationship exhibited by an original random pair. Its measurement is an important task in causal discovery and Bayesian network learning [1]–[3, Ch. 7], which emerge in applied fields such as earth system sciences [4] or clinical data analysis [5]. Due to its nature, empirically estimating conditional dependence is a complex problem, given that it also requires inferring conditional marginal distributions. A known approach consists in measuring the Hilbert-Schmidt norm of the normalized conditional cross-covariance operator [6], grounded in the theory of reproducing kernel Hilbert spaces (RKHS). This kernel-based method is capable of measuring statistical dependence by gauging correlation of data implicitly mapped onto an infinite-dimensional space. However, it is also known to have numerical issues concerned with matrix inversions and regularization due to the usually low-rank structure of the involved correlation matrices [6]. This also leads to a fairly complex method that scales cubically with the observed data size, becoming prohibitive for large data sets.

These issues have lately been the subject of interest of numerous lines of research. A kernel-based conditional independence test is proposed in [1] by deriving the asymptotic distribution of the appropriate test statistic under the null hypothesis, which succeeds in avoiding matrix inversions, but whose complexity still increases cubically with the sample size. Other authors have proposed to approximate kernel functions by randomly

selecting features [7], by providing an alternative review to the embedding process onto the RKHS [8], or by leveraging the interpretation that dependence produces correlation of distances [9], *i.e.* close pairs in one data set coincide with close pairs in the other [10]. Nevertheless, none of these methods addresses the aforementioned computational issues, since they all require matrix inversions. An extension of the Hilbert-Schmidt independence criterion (HSIC) [11], a measure of unconditional independence, is introduced in [12] with the objective of inferring causality, a similar endeavor to this letter, although focused on joint independence rather than conditional dependence.

This letter presents a procedure for measuring conditional dependence by statistically conditioning the observed data to a potential confounder in a simple way, bypassing matrix inversion problems. The basis for conditioning is to prune pairwise differences of data under the control of the confounder [13], stemming from distance-based methods [10]. Simultaneously, the estimate is constructed as an alternative derivation of the kernel-based HSIC. While these two approaches are known to be equivalent [14], we propose here an estimate that combines them, albeit for different tasks. To do so, we first provide a novel reinterpretation of the HSIC in Sec. II by translating the problem of statistical dependence into one of correlation after mapping data onto finite-dimensional spaces based on steering vectors [15]. Sec. III briefly reviews the theory of unbiased statistics (U-Statistics) [16], which is then employed in Sec. IV to show that conditional dependence can be obtained by pruning U-Statistics. The obtained measure, named *conditional* HSIC (C-HSIC), is based on a classical signal processing structure but linked to kernel methods, which embraces the HSIC as a particular case when the U-Statistic is complete.

II. MARGINAL DEPENDENCE AS CORRELATION

Let $\mathbf{d} : \mathbb{R} \to \mathbb{C}^M$ be a mapping based on windowed steering vectors. The mth element of $\mathbf{d}$ can be expressed as

$$[\mathbf{d}(\cdot)]_m \triangleq \tfrac{1}{\sqrt[4]{M}} \mathrm{G}\big(\tfrac{m}{\sqrt{M}}\big) e^{\mathrm{j}\pi \frac{m}{\sqrt{M}}}, \tag{1}$$

where[1] $m \in \{-M/2, \dots, M/2 - 1\}$ and $\mathrm{G} : \mathbb{R} \to \mathbb{R}$ is an even and absolutely integrable function with unit L^2-norm and maximum at $\mathrm{G}(0)$. Given two random variables x and y, we define a pair of transformed random vectors[2] $\mathbf{u} \triangleq \mathbf{d}(\mathsf{x})$ and $\mathbf{v} \triangleq \mathbf{d}(\mathsf{y})$. Their cross-covariance matrix $\mathbf{C}_{\mathbf{u},\mathbf{v}} \in \mathbb{C}^{M \times M}$ is defined as $\mathbf{C}_{\mathbf{u},\mathbf{v}} \triangleq \mathrm{E}\big[\mathbf{u}\mathbf{v}^{\mathrm{H}}\big] - \mathrm{E}[\mathbf{u}]\,\mathrm{E}[\mathbf{v}]^{\mathrm{H}}$, where $\mathrm{E}[\cdot]$ denotes the expectation operator. Then, the following implication holds [11]:

$$\lim_{M \to \infty} \|\mathbf{C}_{\mathbf{u},\mathbf{v}}\|_{\mathrm{F}}^2 = 0 \quad \Longleftrightarrow \quad \mathsf{x}, \mathsf{y} \text{ are independent}, \tag{2}$$

F. de Cabrera is an independent researcher (e-mail: ferrandecabrera@gmail.com). M. Vilà-Insa and J. Riba are with the Signal Processing and Communications Group (SPCOM), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain (e-mail: marc.vila.insa@upc.edu; jaume.riba@upc.edu).

[1]M is assumed even for mathematical convenience.

[2]Without loss of generality, and for the sake of simplicity, we let $\mathbf{u}$ and $\mathbf{v}$ have the same dimensionality.

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Terms $\mathrm{E}[\mathbf{d}(\mathsf{x})]$ and $\mathrm{E}[\mathbf{d}(\mathsf{y})]$ become dense samplings of the weighted (by $\mathrm{G}(\cdot)$) marginal characteristic function of x and y, respectively, in the limit of $M \to \infty$. Similarly, $\mathrm{E}[\mathbf{d}(\mathsf{x})\mathbf{d}^{\mathrm{H}}(\mathsf{y})]$ becomes a dense sampling of the weighted joint characteristic function (JCF). Therefore, (2) is effectively checking the separability of the JCF at every sample point (*i.e.* the factorization of $\mathrm{E}[\mathbf{d}(\mathsf{x})\mathbf{d}^{\mathrm{H}}(\mathsf{y})]$ as a product of expectations), which becomes equivalent to an uncorrelatedness property [15], [17], akin to the *distance covariance* in [10].

Consider L i.i.d. samples of x and y, $\{x(l), y(l)\}_{l=1,\ldots,L}$, from which we obtain two transformed complex data matrices $\mathbf{U} \in \mathbb{C}^{M \times L}$ and $\mathbf{V} \in \mathbb{C}^{M \times L}$, defined as follows:

$$\mathbf{U} \triangleq [\mathbf{u}_1, \ldots, \mathbf{u}_L] = [\mathbf{d}(x(1)), \ldots, \mathbf{d}(x(L))] \quad (3\text{a})$$

$$\mathbf{V} \triangleq [\mathbf{v}_1, \ldots, \mathbf{v}_L] = [\mathbf{d}(y(1)), \ldots, \mathbf{d}(y(L))]. \quad (3\text{b})$$

The sample cross-covariance between the transformed vectors is given by

$$\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}} \triangleq \frac{1}{L-1} \sum_{l=1}^{L} \left(\mathbf{u}_l - \frac{1}{L} \sum_{i=1}^{L} \mathbf{u}_i \right) \left(\mathbf{v}_l - \frac{1}{L} \sum_{j=1}^{L} \mathbf{v}_j \right)^{\mathrm{H}}$$
$$= \frac{1}{L-1} \mathbf{U} \mathbf{P} \mathbf{V}^{\mathrm{H}}, \quad (4)$$

where $\mathbf{P} \triangleq \mathbf{I} - \frac{1}{L} \mathbf{1} \mathbf{1}^{\mathrm{T}} \in \mathbb{R}^{L \times L}$ is the *data centering matrix*, a well-known projection matrix in the signal processing and kernel methods literature [18, Appx. B.7]. Using (2), a marginal dependence measure is given by

$$\|\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}\|_{\mathrm{F}}^2 = \mathrm{tr}(\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}^{\mathrm{H}} \widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}) = \mathrm{tr}\left(\frac{1}{(L-1)^2} \mathbf{V} \mathbf{P}^{\mathrm{H}} \mathbf{U}^{\mathrm{H}} \mathbf{U} \mathbf{P} \mathbf{V}^{\mathrm{H}} \right)$$
$$= \frac{1}{(L-1)^2} \mathrm{tr}(\mathbf{P} \mathbf{U}^{\mathrm{H}} \mathbf{U} \mathbf{P} \mathbf{V}^{\mathrm{H}} \mathbf{V}), \quad (5)$$

where the circularity of the trace operator has been used. To see the link of (5) with the HSIC, let us examine the limit for $M \to \infty$ of the kernel matrices $\mathbf{K} \triangleq \lim_{M \to \infty} \mathbf{U}^{\mathrm{H}} \mathbf{U} \in \mathbb{R}^{L \times L}$ and $\mathbf{Q} \triangleq \lim_{M \to \infty} \mathbf{V}^{\mathrm{H}} \mathbf{V} \in \mathbb{R}^{L \times L}$. The elements of $\mathbf{K}$ (and analogously those of $\mathbf{Q}$) are the following:

$$[\mathbf{K}]_{l,l'} = \lim_{M \to \infty} \mathbf{d}^{\mathrm{H}}(x(l)) \mathbf{d}(x(l'))$$
$$= \lim_{M \to \infty} \sum_{m=-\frac{M}{2}}^{\frac{M}{2}-1} \frac{1}{\sqrt{M}} \mathrm{G}^2\left(\frac{m}{\sqrt{M}} \right) e^{-\mathrm{j}2\pi(x(l)-x(l'))\frac{m}{\sqrt{M}}} \quad (6)$$
$$= \int_{-\infty}^{\infty} \mathrm{G}^2(f) e^{-\mathrm{j}2\pi(x(l)-x(l'))f} \, \mathrm{d}f \triangleq \kappa\big(x(l) - x(l')\big),$$

where $\kappa(\cdot)$ is the kernel function that results from the Fourier transform[3] of $\mathrm{G}^2(f)$, and the integral is the limit of the Darboux sum in the second line. As a result, the entries of $\mathbf{K}$ and $\mathbf{Q}$ are just the evaluation of $\kappa(\cdot)$ at the difference between two data samples. Then, given (5) and taking the limit $M \to \infty$, we obtain the HSIC [11, Sec. 3.1]:

$$\mathrm{HSIC}(\mathbf{x}; \mathbf{y}) \triangleq \frac{1}{(L-1)^2} \mathrm{tr}(\mathbf{P} \mathbf{K} \mathbf{P} \mathbf{Q}) = \lim_{M \to \infty} \|\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}\|_{\mathrm{F}}^2, \quad (7)$$

with $\mathbf{x} = [x(1), \ldots, x(L)]^{\mathrm{T}}$ and $\mathbf{y} = [y(1), \ldots, y(L)]^{\mathrm{T}}$.

---

[3]From the properties imposed on $\mathrm{G}^2(f)$, $\kappa(\cdot)$ becomes naturally an autocorrelation translation-invariant kernel, such that $\kappa(0) = 1 \geq |\kappa(s)|$ and $\kappa(\infty) = \kappa(-\infty) = 0$. The reader is referred to [19, Sec. 1.4] where these ideas emerge in the light of Bochner's theorem.

In summary, the alternative formulation exposed above shows that the HSIC results from measuring correlation in a finite-dimensional space based on steering vectors, where the kernel formulation arises in a second stage once the dimensionality grows without bound. The interplay between kernels and cross-covariance matrices of mapped data will be used later on to condition the kernel matrices by pruning data. Beforehand, we will reformulate the estimation of cross-covariance matrices leveraging incomplete U-statistics, which will serve as the basis for data pruning.

III. SAMPLE COVARIANCE MATRIX AS AN INCOMPLETE U-STATISTIC

Consider a list containing all the unique $K_{\max} \triangleq L(L-1)/2$ pairs that can be constructed from L different samples of a random variable. This list is ordered arbitrarily such that a single index k identifies both elements of a pair. Let $\mathrm{f}_1(k)$ and $\mathrm{f}_2(k)$ be two functions that return the corresponding indices of each term of the kth pair. Given $K \leq K_{\max}$ indices, we construct the pairwise differences

$$\mathring{\mathbf{u}}_k \triangleq \frac{1}{\sqrt{2}} \big(\mathbf{u}_{\mathrm{f}_1(k)} - \mathbf{u}_{\mathrm{f}_2(k)} \big), \quad \mathring{\mathbf{v}}_k \triangleq \frac{1}{\sqrt{2}} \big(\mathbf{v}_{\mathrm{f}_1(k)} - \mathbf{v}_{\mathrm{f}_2(k)} \big), \quad (8)$$

for $k \in \{1, \ldots, K\}$, corresponding to the samples of two virtual sources $\mathring{\mathsf{u}}$ and $\mathring{\mathsf{v}}$. It is worth noting that these are zero-mean variables for any pair with $\mathrm{f}_1(k) \neq \mathrm{f}_2(k)$, since $\mathbf{u}$ and $\mathbf{v}$ are i.i.d.. Their sample cross-covariance matrix $\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}} \in \mathbb{C}^{M \times M}$ is then the following:

$$\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}} = \frac{1}{K} \mathring{\mathbf{U}} \mathring{\mathbf{V}}^{\mathrm{H}}, \quad (9)$$

where $\mathring{\mathbf{U}} \triangleq [\mathring{\mathbf{u}}_1, \ldots, \mathring{\mathbf{u}}_K] \in \mathbb{C}^{M \times K}$ and $\mathring{\mathbf{V}} \triangleq [\mathring{\mathbf{v}}_1, \ldots, \mathring{\mathbf{v}}_K] \in \mathbb{C}^{M \times K}$. Note that, thanks to the constant factor in (8), $\mathring{\mathsf{u}}$ and $\mathring{\mathsf{v}}$ have the same average covariance matrix as $\mathbf{u}$ and $\mathbf{v}$, respectively. In contrast to (4), $\mathbf{P}$ is missing in (9) as a result of constructing zero-mean virtual data in (8), which will lead to a cleaner implementation and fewer matrix operations.

Equation (9) is in fact an instance of a U-Statistic. For $K = K_{\max}$, *i.e.* when all data pairs are taken, we obtain the equality $\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}} = \widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}$ from U-Statistics theory [16]. In contrast, for $K < K_{\max}$, *i.e.* when (9) is an incomplete U-Statistic, although $\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}}$ remains unbiased, its estimation variance increases due to the pruning of data. Nevertheless, $\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}}$ is still a consistent estimate of $\mathbf{C}_{\mathbf{u},\mathbf{v}}$ provided that $K \to \infty$ as $L \to \infty$.

Remark 1 (Robustness to pruning): A notable property of incomplete U-statistics is that their robustness against data pruning increases the larger L is [13]. To briefly illustrate this property, consider taking only the $K = \lfloor L/2 \rfloor$ data pairs with no indices in common. The number of remaining, unused, data pairs is equal to $L(L-1)/2 - \lfloor L/2 \rfloor$, which increases with $O(L^2)$. To only use the i.i.d., unique, terms is effectively equivalent to computing $\widehat{\mathbf{C}}_{\mathbf{u},\mathbf{v}}$ with half of the available samples [20]. It is then safe to assume that their contribution to the overall accuracy of the sample covariance is higher than those with repeated indices. Therefore, the larger L is, the higher the amount of pairs that can be pruned for some specified degradation in the estimation accuracy of the resulting sample covariance. The implication is that K in the incomplete U-Statistic can be designed to grow with $O(L)$

instead of $O(L^2)$, which provides considerable flexibility for pruning, eases the overall computational complexity, and will be used for choosing the number of pairs in the next section.

Once we have determined the incomplete U-statistic formulation of the cross-covariance matrix, we are now in a position to establish the underlying rule of $f_1(\cdot)$ and $f_2(\cdot)$ by looking at sample pairwise differences of the confounder. Afterwards, we will be employing the pruned cross-covariance matrix to return to kernel methods in the same way as portrayed in (7).

IV. CONDITIONAL DEPENDENCE VIA U-STATISTICS

The conditional cross-covariance matrix between $\mathbf{u}$ and $\mathbf{v}$ is defined as:

$$\mathbf{C}_{\mathbf{u},\mathbf{v}|\mathsf{z}} \triangleq \int_{\mathbb{R}} \mathbf{C}_{\mathbf{u},\mathbf{v}|z=z}\, \mathrm{d}F_{\mathsf{z}}(z) = \int_{\mathbb{R}} \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|z=z}\, \mathrm{d}F_{\mathsf{z}}(z), \quad (10)$$

where $\mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|z=z} = \mathbf{C}_{\mathbf{u},\mathbf{v}|z=z}$ (given (8)) is the cross-covariance matrix conditioned to a specific value z of confounder z, and $F_{\mathsf{z}}(z)$ is its cumulative distribution. With the goal of deriving an estimator of (10), we define the virtual random variable $\mathring{\mathsf{z}} \triangleq \mathsf{z}_1 - \mathsf{z}_2$, where z_1 and z_2 are mutually independent and distributed as z. Integrating over all values of z is equivalent to doing so over the events in which z_1 and z_2 take the same value, *i.e.* $\mathring{\mathsf{z}} = 0$. Therefore, the expectation in (10) can be alternatively expressed as

$$\mathbf{C}_{\mathbf{u},\mathbf{v}|\mathsf{z}} = \iint_{\mathbb{R}^2} \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{\mathsf{z}}=0}\, \mathrm{d}F_{\mathsf{z}}(z_1)\, \mathrm{d}F_{\mathsf{z}}(z_2)$$
$$= \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{\mathsf{z}}=0} \iint_{\mathbb{R}^2} \mathrm{d}F_{\mathsf{z}}(z_1)\, \mathrm{d}F_{\mathsf{z}}(z_2) = \mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{\mathsf{z}}=0}, \quad (11)$$

where $\iint_{\mathbb{R}^2} \mathrm{d}F_{\mathsf{z}}(z_1)\, \mathrm{d}F_{\mathsf{z}}(z_2) = 1$, and $\mathbf{C}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathring{\mathsf{z}}=0}$ does not depend on the specific values of z_1 and z_2 but rather on them being equal. In consequence, conditioning with respect to z is equal to conditioning with respect to $\mathring{\mathsf{z}} = 0$. Therefore, (11) suggests to let the pruning of the incomplete U-Statistics in (9) be handled by the potential confounder. However, since $\mathring{\mathsf{z}} = 0$ is an event of zero probability for continuous random variables, data control should be based on merely small values of $|\mathring{\mathsf{z}}|$ [13].

With the intention of choosing data pairs to prune according to $|\mathring{\mathsf{z}}|$, we define the samples of $\mathring{\mathsf{z}}$ as follows:

$$\mathring{z}_{l,l'} \triangleq z(l) - z(l'), \qquad l \neq l', \quad (12)$$

where $z(l)$ and $z(l')$ are i.i.d. samples drawn from z. Then, we let the sorting of $|\mathring{z}_{l,l'}|$ (in ascending order) be the one that determines the ordering of the index pairs provided by $f_1(\cdot)$ and $f_2(\cdot)$ in Sec. III. Moreover, in view of *Remark 1*, the amount of pairs K is set to grow as $O(L)$ with

$$K = \left\lfloor \frac{L\alpha}{2} \right\rfloor, \quad (13)$$

being $1 \leq \alpha \ll (L-1)$ a tuning hyper-parameter. While $\alpha = 1$ ensures that only very small values of $|\mathring{z}_{l,l'}|$ are considered, the U-Statistic becomes complete and there is no conditioning at all for $\alpha = L - 1$, thus yielding the HSIC as a particular case as in (7). Then a natural trade-off emerges: low values of α are desirable to provide strong conditioning of data, but may also lead to excessive pruning with low statistical accuracy in (9). Remarkably, the selection of α becomes a minor issue provided that L is sufficiently large, as shown in [13] under the

correlation measure framework between a pair of vectors. This will be further discussed in Sec. V with a numerical example.

A. Conditional HSIC

Now that we have determined the sorting and pruning of data pairs according to the confounder z, let us write a conditional dependence measure as the Frobenius norm of (9):

$$\mathrm{tr}\big(\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathsf{z}}^{\mathrm{H}} \widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathsf{z}}\big) = \frac{1}{K^2}\, \mathrm{tr}\big(\mathring{\mathbf{U}}^{\mathrm{H}}\mathring{\mathbf{U}}\mathring{\mathbf{V}}^{\mathrm{H}}\mathring{\mathbf{V}}\big), \quad (14)$$

where the circularity of the trace has been used. To link the previous expression with kernel-based methods, let us rewrite the zero-mean virtual data matrix $\mathring{\mathbf{U}}$ as follows:

$$\mathring{\mathbf{U}} = \frac{1}{\sqrt{2}}(\mathbf{U}_1 - \mathbf{U}_2), \quad \mathbf{U}_a \triangleq [\mathbf{u}_{f_a(1)}, \ldots, \mathbf{u}_{f_a(K)}], \quad (15)$$

for $a = \{1, 2\}$. The same is done for $\mathring{\mathbf{V}}$. Accordingly, (14) is then rewritten as

$$\mathrm{tr}\big(\widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathsf{z}}^{\mathrm{H}} \widehat{\mathbf{C}}_{\mathring{\mathbf{u}},\mathring{\mathbf{v}}|\mathsf{z}}\big) = \quad\quad\quad\quad (16)$$
$$\frac{1}{4K^2}\, \mathrm{tr}\big((\mathbf{U}_1 - \mathbf{U}_2)^{\mathrm{H}}(\mathbf{U}_1 - \mathbf{U}_2)(\mathbf{V}_1 - \mathbf{V}_2)^{\mathrm{H}}(\mathbf{V}_1 - \mathbf{V}_2)\big).$$

Taking the limit of $M \to \infty$, kernel matrices are obtained from the products among $\mathbf{U}$ and $\mathbf{V}$ in (16):

$$\mathbf{K}_{a,a'} \triangleq \lim_{M \to \infty} \mathbf{U}_a^{\mathrm{H}} \mathbf{U}_{a'}, \qquad \mathbf{Q}_{a,a'} \triangleq \lim_{M \to \infty} \mathbf{V}_a^{\mathrm{H}} \mathbf{V}_{a'}, \quad (17)$$

whose elements are

$$[\mathbf{K}_{a,a'}]_{k,k'} = \kappa\big(x(f_a(k)) - x(f_{a'}(k'))\big), \quad (18a)$$
$$[\mathbf{Q}_{a,a'}]_{k,k'} = \kappa\big(y(f_a(k)) - y(f_{a'}(k'))\big), \quad (18b)$$

where $\kappa(\cdot)$ is the kernel function as in (6). Finally, the resulting C-HSIC can be expressed as follows:

$$\text{C-HSIC}_\alpha(\mathbf{x}; \mathbf{y}) \triangleq \frac{1}{4K^2}\, \mathrm{tr}\big(\mathbf{\breve{K}}\mathbf{\breve{Q}}\big) \quad (19)$$

with $\mathbf{\breve{K}} \triangleq \mathbf{K}_{1,1} + \mathbf{K}_{2,2} - \mathbf{K}_{1,2} - \mathbf{K}_{1,2}^{\mathrm{H}}$ and $\mathbf{\breve{Q}} \triangleq \mathbf{Q}_{1,1} + \mathbf{Q}_{2,2} - \mathbf{Q}_{1,2} - \mathbf{Q}_{1,2}^{\mathrm{H}}$. Note that, as a result of the U-Statistics implementation, each entry of the new kernel-based matrices involves four data samples of the same source (4-tuples), in contrast to only the pairs typically involved in classical kernel methods. This fact, along with the lack of $\mathbf{P}$, are the main distinctive features of the C-HSIC (19) vs. HSIC (7). In contrast to other conditional dependence measures, (19) does not require any matrix inversion. Since K is set to grow linearly with L in (13), the computational complexity is $O(L^2)$.

V. NUMERICAL ILLUSTRATIONS

To test the proposed method, we aim at generating uncorrelated data with a controlled amount of conditional dependence. Two scenarios are studied, $\mathcal{M}^+$ and $\mathcal{M}^-$, defined as follows:

$$\mathcal{M}^+ : \begin{cases} \mathsf{x} = \sqrt{\gamma}\mathsf{a}\mathsf{p} + \mathsf{v} \\ \mathsf{y} = \sqrt{\gamma}\mathsf{a}\mathsf{q} + \mathsf{w} \\ \mathsf{z} = \mathsf{a} \end{cases}, \quad \mathcal{M}^- : \begin{cases} \mathsf{x} = \sqrt{\gamma}\mathsf{b}\mathsf{p} + \mathsf{v} \\ \mathsf{y} = \sqrt{\gamma}\mathsf{c}\mathsf{q} + \mathsf{w} \\ \mathsf{z} = \mathsf{b} - \mathsf{c} \end{cases}. \quad (20)$$

The internal i.i.d. random variables are distributed as $\mathsf{a}, \mathsf{b}, \mathsf{c} \sim \mathcal{U}(0, \sqrt{3})$ (uniform), $\mathsf{v}, \mathsf{w} \sim \mathcal{N}(0, 1)$ (normal), and $\mathsf{p}, \mathsf{q} \sim \text{Bern}_{1/2}\{-1, 1\}$ (equiprobable Bernoulli). In $\mathcal{M}^+$, the pair $\{\mathsf{x}, \mathsf{y}\}$ are dependent, *i.e.* their mutual information is greater than zero, but they are conditionally independent, since knowing
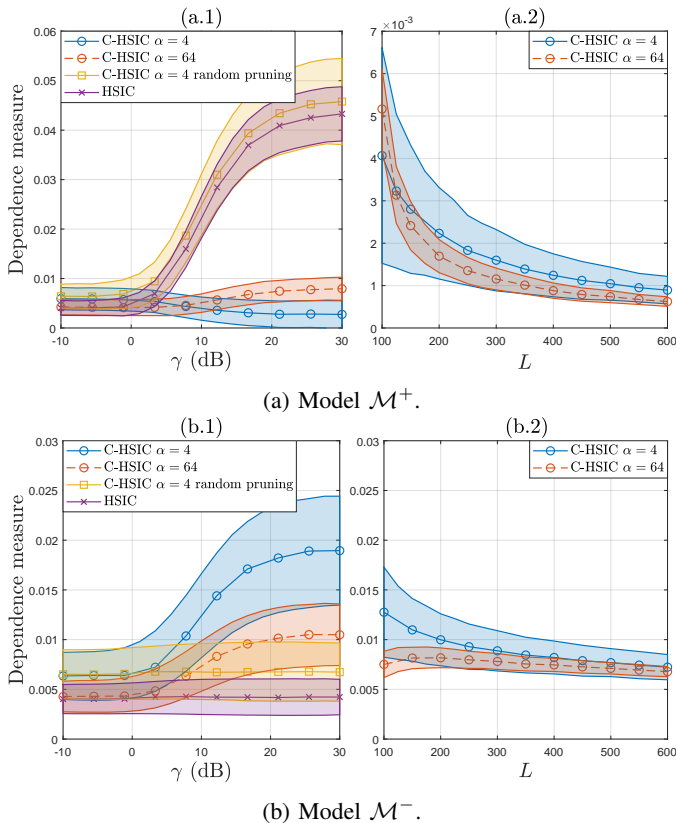
(a) Model $\mathcal{M}^+$.

(b) Model $\mathcal{M}^-$.

Fig. 1: Dependence measures as a function of γ (with $L = 100$) and L (with $\gamma = 10$dB). C-HSIC measures conditional dependence while HSIC measures unconditional dependence. Markers denote the empirical average and bands indicate the standard deviation.

z implies that x and y become solely driven by independent phenomena (v and w). By contrast, x and y are marginally independent in $\mathcal{M}^-$, but become conditionally dependent, since knowledge of z correlates the possible joint values of b and c. Parameter γ is the signal-to-noise ratio associated with the measurements and controls the total amount of unconditional dependence in $\mathcal{M}^+$ and conditional dependence in $\mathcal{M}^-$. In both models, x, y and z are mutually uncorrelated due to the multiplicative effect of mutually independent variables p and q, so correlation measures are unable to discover any data association. Similar ideas on modeling conditional dependencies can be found in [21], which are inspired by co-information [22], [23].

The universal Gaussian kernel is used as the kernel function of choice, which yields $\kappa(s) = \exp(-(\frac{s}{\hat{\sigma} L^{-1/5}})^2)$ being $\hat{\sigma}$ the sample standard deviation. We use this expression because of its association with kernel density estimation, known to be related to kernel methods when estimating certain dependence measures [24, Ch. 2]. This connection also justifies the adoption of the power law $O(L^{-1/5})$ [25, Ch. 3], [15, Appx. D].

Fig. 1 shows measures of dependence for both models with two α values from (13): a moderate $\alpha = 4$ and a sixteenfold increase $\alpha = 64$. This yields 4% and 64.5%, respectively, of the total data pairs for $L = 100$, or 0.7% and 10.7% for $L = 600$ (and even less for higher L). For $\mathcal{M}^+$ (Fig. 1-(a)), C-HSIC

correctly depicts conditional independence for an increasing value of γ, while HSIC confirms that marginal dependence is high as γ increases. Conversely, Fig. 1-(b) exhibits the capability of C-HSIC to discover conditional dependence for moderate and high γ values in $\mathcal{M}^-$, while marginal dependence is confirmed to be small by HSIC at any value of γ.

To get further insights on the proposed ideas, Fig. 1-(a.1) and Fig. 1-(b.1) also show the performance of C-HSIC when pruning data pairs is performed randomly in the U-statistic, *i.e.* the pair given by $f_1(\cdot)$ and $f_2(\cdot)$ is not controlled by the confounder. As a result of not being properly conditioned, C-HSIC produces a measure of marginal dependence similar to HSIC, but with increased variance due to the pruning itself. It is also worth noting that random and orderly pruning perform similarly for low values of γ, corresponding to the regime where the confounder has no influence on the pair $\{x, y\}$, and start to deviate as γ increases. Additionally, it can be seen that the expected value deteriorates for $\alpha = 64$, increasing in $\mathcal{M}^+$ and decreasing in $\mathcal{M}^-$. This is a consequence of a mild pruning for $L = 100$ (where $\alpha = 64$ is intentionally chosen to illustrate this effect), worsening the conditioning process and starting to behave as an unconditioned measure.

Finally, Fig. 1-(a.2) and Fig. 1-(b.2) illustrate the behavior of C-HSIC for multiple values of L at $\gamma = 10$dB. Model $\mathcal{M}^+$, which is conditionally independent, shows that C-HSIC correctly approaches zero as L increases. Conversely, it tends to some nonzero value in $\mathcal{M}^-$. As mentioned above, $\alpha = 64$ is using too many data pairs for small L values and provides a degraded empirical average. In both models, $\alpha = 4$ shows an increased variance with respect to $\alpha = 64$ due to a more thorough pruning. However, this is reduced for sufficiently large data sizes due to *Remark 1*. Similarly, the gap between expected values for $\alpha = 4$ and $\alpha = 64$ is reduced as L increases, albeit the former has the advantage of a lower computational complexity. These results show that α is a noncritical hyper-parameter, provided it is reasonably small.

VI. CONCLUSIONS

This work has proposed a proof of concept for a modification on the classical HSIC under the conditional dependence framework, named C-HSIC. Its interpretation as a correlation measure on a finite but high-dimensional space allows an insightful connection with kernels by letting the dimension grow without bound. Moreover, it opens the possibility of leveraging U-Statistics for this task. Thanks to this formulation, we can provide a novel approach for performing statistical conditioning by pruning data pairs based on the pairwise differences of a potential confounder. Furthermore, the proposed measure of conditional dependence does not require matrix inversions, which has the advantage of reduced computational complexity and the avoidance of addressing ill-conditioned matrices. Numerical illustrations have shown that C-HSIC is capable of measuring both conditional dependence and independence in two different scenarios. Further research should study the potential of the proposed method with richer data sets, as well as provide a thorough comparison with other methods for measuring conditional dependence.

REFERENCES

[1] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, *Kernel-based Conditional Independence Test and Application in Causal Discovery.* Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI), 2011, pp. 804–813.

[2] J. Zhang, "Causal reasoning with ancestral graphs," *Journal of Machine Learning Research*, vol. 9, no. 7, 2008.

[3] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms.* Cambridge, MA, USA: The MIT Press, 2017.

[4] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, 2019.

[5] X. Shen, S. Ma, P. Vemuri, and G. Simon, "Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology," *Scientific reports*, vol. 10, no. 1, 2020.

[6] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[7] E. V. Strobl, K. Zhang, and S. Visweswaran, "Approximate kernel-based conditional independence tests for fast non-parametric causal discovery," *Journal of Causal Inference*, vol. 7, no. 1, p. 20180017, 2019.

[8] J. Park and K. Muandet, "A measure-theoretic approach to kernel conditional mean embeddings," *Advances in neural information processing systems*, vol. 33, pp. 21 247–21 259, 2020.

[9] X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang, "Conditional distance correlation," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1726–1734, 2015.

[10] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769 – 2794, 2007.

[11] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," *International conference on algorithmic learning theory*, pp. 63–77, 2005.

[12] N. Pfister, P. Bühlmann, B. Schölkopf, and J. Peters, "Kernel-based tests for joint independence," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 80, no. 1, pp. 5–31, 2018.

[13] M. Vilà and J. Riba, "A test for conditional correlation between random vectors based on weighted U-statistics," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5792–5796.

[14] T. Sheng and B. K. Sriperumbudur, "On distance and kernel measures of conditional dependence," *Journal of Machine Learning Research*, vol. 24, no. 7, pp. 1–16, 2023.

[15] F. de Cabrera and J. Riba, "Regularized estimation of information via canonical correlation analysis on a finite-dimensional feature space," *IEEE Transactions on Information Theory*, vol. 69, no. 8, pp. 5135–5150, 2023.

[16] A. J. Lee, *U-Statistics: Theory and Practice.* Boca Raton, FL, USA: Routledge, 1990.

[17] F. de Cabrera and J. Riba, "A novel formulation of independence detection based on the sample characteristic function," in *26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2608–2612.

[18] D. Ramírez, I. Santamaría, and L. Scharf, *Coherence: In Signal Processing and Machine Learning.* Cham, Switzerland: Springer Nature, 2023.

[19] W. Rudin, *Fourier Analysis on Groups.* Hoboken, NJ, USA: John Wiley & Sons, Ltd, 1990.

[20] X. Chen and K. Kato, "Randomized incomplete U-statistics in high dimensions," *The Annals of Statistics*, vol. 47, no. 6, pp. 3127 – 3156, 2019.

[21] B. Póczos and J. Schneider, "Nonparametric estimation of conditional information and divergences," in *Proceedings of 15th International Conference on Artificial Intelligence and Statistics (AISTATS '12)*, April 2012, pp. 914 – 923.

[22] A. J. Bell, "The co-information lattice," in *Proceedings of the fifth international workshop on independent component analysis and blind signal separation: ICA*, vol. 2003, 2003.

[23] A. Ghassami and N. Kiyavash, "Interaction information for causal inference: The case of directed triangle," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 1326–1330.

[24] J. C. Príncipe, *Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives.* New York, NY, USA: Springer, 2010.

[25] B. Silverman, *Density estimation for statistics and data analysis.* London, UK: Chapman and Hall, 1986.