Scale Invariant Conditional Dependence Measures

Sashank J. Reddi SJAKKAMR@CS.CMU.EDU
Machine Learning Department, School of Computer Science, Carnegie Mellon University

Barnabás Póczos BAPOCZOS@CS.CMU.EDU
Machine Learning Department, School of Computer Science, Carnegie Mellon University

Abstract

In this paper we develop new dependence and conditional dependence measures and provide their estimators. An attractive property of these measures and estimators is that they are invariant to any monotone increasing transformations of the random variables, which is important in many applications including feature selection. Under certain conditions we show the consistency of these estimators, derive upper bounds on their convergence rates, and show that the estimators do not suffer from the curse of dimensionality. However, when the conditions are less restrictive, we derive a lower bound which proves that in the worst case the convergence can be arbitrarily slow similarly to some other estimators. Numerical illustrations demonstrate the applicability of our method.

1. Introduction

Measuring dependencies and conditional dependencies are of great importance in many scientific fields including machine learning, and statistics. There are numerous problems where we want to know how large the dependence is between random variables, and how this dependence changes if we observe other random variables. Correlated random variables might become independent when we observe a third random variable, and the opposite situation is also possible where independent variables become dependent after observing other random variables.

Thanks to the wide potential application range (e.g., in bioinformatics, pharmacoinformatics, epidemiology, psychology, econometrics), finding efficient dependence and conditional dependence measures has been an active research area for decades. These measures have been used, for example, in causality detection, feature selection, active learning, structure learning, boosting, image registration, independent component and subspace analysis. Although the theory of these estimators is actively researched, there are still several fundamental open questions.

The estimation of certain dependence and conditional dependence measures is easy in a few cases: For example, (i) when the random variables have discrete distributions with finitely many possible values, (ii) when there is a known simple relationship between them (e.g., a linear model describes their behavior), or (iii) if they have joint distributions that belong to a parametric family that is easy to estimate (e.g. normal distributions). In this paper we consider the more challenging nonparametric estimation problem when the random variables have continuous distributions, and we do not have any other information about them.

Numerical experiments indicate, that a recently proposed kernel measure based on normalized cross-covariance operators (D_{HS}) appears to be very powerful in measuring both dependencies and conditional dependencies (Fukumizu et al., 2008). Nonetheless, even this method has a few drawbacks and several open fundamental questions. In particular, lower and upper bounds on the convergence rates of the estimators are not known. It is also not clear if the D_{HS} measure or its existing estimator $\widehat{D}_{HS}$ are invariant to any invertible transformations of the random variables. This kind of invariance property is so important that Schweizer & Wolff (1981) even put it into the axioms of dependence. One reason for this is that in many scenarios we need to compare the estimated dependencies. If certain variables are measured on different scales, the dependence can be much different in absence of this invariance property. As a result, it

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

might happen that in a dependence based feature selection algorithm different features would be selected if we measured a quantity e.g. in grams, kilograms, or if we used log-scale. This is an odd situation that can be avoided with dependence measures that are invariant to invertible transformations of the variables.

Main contributions: The goal of this paper is to provide new theoretical insights in this field. Our contributions can be summarized in the following points: (i) We prove that the dependence and conditional dependence measures (D_{HS}) are invariant to any invertible transformations, but its estimator $\widehat{D}_{HS}$ does not have this property. (ii) Under some conditions we derive an upper bound on the rate of convergence of $\widehat{D}_{HS}$. (iii) We show that if we apply $\widehat{D}_{HS}$ on the empirical copula transformed points, then the resulting estimator $\widehat{D}_C$ will be invariant to any monotone increasing transformations. (iv) We show that under some conditions the estimator is consistent and derive an upper bound on the rate. (v) We prove that if the conditions are less restrictive, then convergence of both $\widehat{D}_{HS}$ and $\widehat{D}_C$ can be arbitrarily slow. (vi) We also generalize these dependence measures as well as their estimators to sets of random variables and provide an upper bound on the convergence rate of the estimators.

Related work: Since the literature on dependence measure is huge, we only mention few prominent examples here. The most well-known dependence measure is probably the Shannon mutual information, which has been generalized to the Rényi-α (Rényi, 1961) and Tsallis-α mutual information (Tsallis, 1988). Other interesting dependence measures are the maximal correlation coefficient (Rényi, 1959), kernel mutual information (Gretton et al., 2003), the generalized variance and kernel canonical correlation analysis (Bach, 2002), the Hilbert-Schmidt independence criterion (Gretton et al., 2005), the Schweizer-Wolff measure (Schweizer & Wolff, 1981), maximum-mean discrepancy (MMD) (Borgwardt et al., 2006; Fortet & Mourier, 1953), Copula-MMD (Póczos et al., 2012), and the distance based correlation (Székely et al., 2007). Some of these measures, e.g. the Shannon mutual information, are invariant to any invertible transformations of the random variables. The Copula-MMD and Schweizer-Wolff measures are invariant to monotone increasing transformations, while the MMD and many other dependence measures are not invariant to any of these transformations. The conditional dependence estimation is an even more challenging problem, and only very few dependence measures have been generalized to the conditional case (Fukumizu et al., 2008; Poczos & Schneider, 2012).

Notation: We use $X \sim P$ to denote that the random variable X has probability distribution P. The symbol $X \perp\!\!\!\perp Y|Z$ indicates the conditional independence of X and Y given Z. Let $X_{i:j}$ denote the tuple $(X_i, \ldots, X_j)$. $\mathbb{E}[X]$ stands for the expectation of random variable X. The symbols $\mu_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{X}}$ denote measure and Borel $\sigma-$field on $\mathcal{X}$ respectively. We use $D(X_i, \ldots, X_j)$ to denote the dependence measure of set of random variables $\{X_i, \ldots, X_j\}$. With slight abuse of notation, we will use $D(X, Y)$ to denote the dependence measure between sets of the random variables X and Y. $\{A_j\}$ denotes the set $\{A_1, \ldots, A_k\}$ where k will be clear from the context. The null space and range of an operator L are denoted by $\mathcal{N}(L)$ and $\mathcal{R}(L)$ respectively, and $\overline{A}$ stands for the closure of set A.

2. Dependence Measure using Hilbert-Schmidt Norm

In this section, we review the theory behind the Hilbert Schimdt (HS) norm and its use in defining dependence measures. Suppose (X, Y) is a random variable on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H}_X = \{f : \mathcal{X} \to \mathbb{R}\}$ be a reproducing kernel Hilbert Space (RKHS) associated with X, feature map $\phi(x) \in \mathcal{H}_X$ $(x \in \mathcal{X})$ and kernel $k_X(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}_X}$ $(x, y \in \mathcal{X})$. The kernel satisfies the property $f(x) = \langle f, k_X(x, .) \rangle_{\mathcal{H}_X}$ for $f \in \mathcal{H}_X$, which is called the reproducing property of the kernel. We can similarly define RKHS $\mathcal{H}_Y$ and kernel k_Y associated with Y. Let us define a class of kernels known as universal kernels, which are critical to this paper.

Definition 1. *A kernel $k_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called universal whenever the associated RKHS $\mathcal{H}_X$ is dense in $C(\mathcal{X})$ — the space of bounded continuous functions over $\mathcal{X}$ — with respect to the L_∞ norm.*

Gaussian and Laplace kernels are two popular kernels which belong to the class of universal kernels.

The cross-covariance operators on these RKHSs capture the dependence of random variables X and Y (Baker, 1973). In order to ensure existence of these operators, we assume that $\mathbb{E}[k_X(X, X)] < \infty$ and $\mathbb{E}[k_Y(Y, Y)] < \infty$. We also assume that all the kernels defined in this paper satisfy the above assumption. The cross-covariance operator (COCO) $\Sigma_{YX} : \mathcal{H}_X \to \mathcal{H}_Y$ is an operator such that

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_Y} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\, E_Y[g(Y)]$$

holds for all $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$. Existence and uniqueness of such an operator can be shown using the Riesz representation theorem (Reed & Simon, 1980).

If X and Y are identical, then the operator Σ_{XX} is called the covariance operator. Both operators above are natural generalizations of the covariance matrix in Euclidean space to Hilbert space. Analogous to correlation matrix in the Euclidean space, we can define operator V_{YX},

$$V_{YX} = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2},$$

where $\mathcal{R}(V_{YX}) \subset \overline{\mathcal{R}(\Sigma_{YY})}$ and $\mathcal{N}(V_{YX})^{\perp} \subset \overline{\mathcal{R}(\Sigma_{XX})}$. This operator is called the normalized cross-covariance operator (NOCCO). Similar to Σ_{YX}, the existence of NOCCO can be proved using Riesz representation theorem. Intuitively, the normalized cross-covariance operator captures the dependence of random variables X and Y discounting the influence of the marginals. We would like to point out that the notation we adopted leads to a few technical difficulties which can easily be addressed (refer (Grünewälder et al., 2012)).

We also define the conditional covariance operators which will be useful to capture conditional independence. Suppose we have another random variable Z on $\mathcal{Z}$ with RHKS $\mathcal{H}_Z$ and kernel k_Z. The normalized conditional cross-covariance operator is defined as:

$$V_{YX|Z} = V_{YX} - V_{YZ}V_{ZX}.$$

Similar to the cross-covariance operator, the conditional cross-covariance operator is a natural extension of conditional covariance matrix to Hilbert space. An interesting aspect of the normalized conditional cross-covariance operator is that it can be expressed in terms of simple products of normalized cross-covariance operators (Fukumizu et al., 2008).

It is not surprising that covariance operators described above can be used for measuring dependence between random variables since they capture the dependence between them. While one can use Σ_{YX} or V_{YX} in defining the dependence measure (Gretton et al., 2005; Fukumizu et al., 2008), we will use the latter in this paper. Let the variable X' denote (X, Z), which will be useful for defining conditional dependence measures. We define the HS norm of a linear operator $L : \mathcal{H}_X \to \mathcal{H}_Y$ as follows.

Definition 2. *(Hilbert-Schmidt Norm) The Hilbert-Schmidt norm of L is defined as*

$$\|L\|_{HS}^2 = \sum_{i,j}\langle v_j, Lu_i\rangle_{\mathcal{H}_Y}^2,$$

where $\{u_i\}$ and $\{v_j\}$ are an orthonormal bases of $\mathcal{H}_X$ and $\mathcal{H}_Y$ respectively, provided the sum converges.

The HS norm is a generalization of the Frobenius norm on matrices and is independent of the choice of the orthonormal bases. An operator is called Hilbert-Schmidt if its HS norm is finite. The covariance operators defined in this paper are assumed to be Hilbert-Schmidt Operators. Fukumizu et al. (2008) define the following dependence measures:

$$D_{HS}(X, Y) = \|V_{YX}\|_{HS}^2,$$
$$D_{HS}(X, Y|Z) = \|V_{YX'|Z}\|_{HS}^2.$$

Note that the measures above are defined for a pair of random variables X and Y. In Section 5 we will generalize this approach and provide dependence and conditional measures with estimators that operate on sets of random variables. Theorem 10 (in the supplementary material) justifies the use of the above dependence measures. The result can be equivalently stated in terms of HS norm of the covariance operators since HS norm of an operator is zero if and only if the operator itself is a null operator. At first it might appear that these measures are strongly linked to the kernel used in constructing the measure but Fukumizu et al. (2008) show a remarkable property that the measures are independent of the kernels, which is captured by the following result. Let $\mathbb{E}_Z[P_{X|Z} \otimes P_{Y|Z}(A \times B)] = \int \mathbb{E}[\mathbb{1}_B(Y)|Z = z]\mathbb{E}[\mathbb{1}_A(X)|Z = z]dP_Z(z)$ for $A \in \mathcal{B}_{\mathcal{X}}$ and $B \in \mathcal{B}_{\mathcal{Y}}$.

Theorem 1. *(Kernel-Free property) Assume that the probabilities P_{XY} and $\mathbb{E}_Z[P_{X|Z} \otimes P_{Y|Z}]$ are absolutely continuous with respect to $\mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$ with probability density functions p_{XY} and $p_{X \perp\!\!\!\perp Y|Z}$, respectively, then we have*

$$D_{HS}(X, Y|Z) = \iint_{\mathcal{X}\times\mathcal{Y}} \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - \frac{p_{X \perp\!\!\!\perp Y|Z}(x,y)}{p_X(x)p_Y(y)}\right)^2 p_X(x)p_Y(y)d\mu_{\mathcal{X}}d\mu_{\mathcal{Y}}.$$

Suppose $\mathcal{Z} = \emptyset$, we have

$$D_{HS}(X, Y) = \iint_{\mathcal{X}\times\mathcal{Y}} \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - 1\right)^2 p_X(x)p_Y(y)d\mu_{\mathcal{X}}d\mu_{\mathcal{Y}}.$$

The above result shows that these measures bear an uncanny resemblance to mutual information and might possibly inherit some of its desirable properties. We show that this intuition is, in fact, true by proving an important consequence of the above result. In particular, the following result holds when the assumptions stated in Theorem 1 are satisfied.

Theorem 2. *(Invariance of dependence measure) Assume that the probabilities P_{XY} and $\mathbb{E}_Z[P_{X|Z} \otimes P_{Y|Z}]$ are absolutely continuous. Let $\mathcal{X} \subset \mathbb{R}^d$, and let*

$\Gamma_X : \mathcal{X} \to \mathbb{R}^d$ be a differentiable injective function. Let Γ_Y and Γ_Z be defined similarly. Under the above assumptions, we have

$$D_{HS}(X, Y|Z) = D_{HS}(\Gamma_X(X), \Gamma_Y(Y)|\Gamma_Z(Z)).$$

As a special case of $\mathcal{Z} = \emptyset$, we have

$$D_{HS}(X, Y) = D_{HS}(\Gamma_X(X), \Gamma_Y(Y)).$$

The proof is in the supplementary material. Though we proved the result for Euclidean spaces, it can be generalized to certain measurable spaces under mild conditions (Hewitt & Stromberg, 1975). We now look at the empirical estimators for the dependence measures defined above. Let $X_{1:m}, Y_{1:m}$ and $Z_{1:m}$ be i.i.d samples from the joint distribution. Let $\widehat{\mu}_X^{(m)} = \frac{1}{m}\sum_{i=1}^m k_X(\cdot, X_i)$ and $\widehat{\mu}_Y^{(m)} = \frac{1}{m}\sum_{i=1}^m k_Y(\cdot, Y_i)$ denote the empirical mean maps respectively. The empirical estimator of Σ_{YX} is

$$\widehat{\Sigma}_{YX}^{(m)} = \frac{1}{m}\sum_{i=1}^m \left(k_Y(\cdot, Y_i) - \widehat{\mu}_Y^{(m)}\right) \langle k_X(\cdot, X_i) - \widehat{\mu}_X^{(m)}, \cdot\rangle_{\mathcal{H}_X}$$

The empirical covariance operators $\widehat{\Sigma}_{XX}^{(m)}$ and $\widehat{\Sigma}_{YY}^{(m)}$ can be defined in a similar fashion. The empirical normalized cross-covariance operator V_{YX} is

$$\widehat{V}_{YX}^{(m)} = \left(\widehat{\Sigma}_{YY}^{(m)} + \epsilon_m I\right)^{-1/2} \widehat{\Sigma}_{YX}^{(m)} \left(\widehat{\Sigma}_{XX}^{(m)} + \epsilon_m I\right)^{-1/2},$$

where $\epsilon_m > 0$ is the regularization constant (Bach, 2002; Fukumizu et al., 2004). In the later sections, we look at a particular choice of ϵ_m which provides good convergence rates. The empirical conditional cross-covariance operator is

$$\widehat{V}_{YX|Z}^{(m)} = \widehat{V}_{YX}^{(m)} - \widehat{V}_{YZ}^{(m)}\widehat{V}_{ZX}^{(m)}.$$

Let G_X be the centered gram matrix such that $G_{X,ij} = \langle k_X(\cdot, X_i) - \widehat{\mu}_X, k_X(\cdot, X_j) - \widehat{\mu}_X\rangle_{\mathcal{H}_X}$ and $R_X = G_X\left(G_X + m\epsilon_m I\right)^{-1}$. Similarly, we can define G_Y, G_Z and R_Y, R_Z for random variables Y and Z. The empirical dependence measures are then

$$\begin{aligned}\widehat{D}_{HS}(X, Y) &= \|\widehat{V}_{YX}^{(m)}\|_{HS}^2 = Tr\left[R_Y R_X\right], \\ \widehat{D}_{HS}(X, Y|Z) &= \|\widehat{V}_{YX'|Z}^{(m)}\|_{HS}^2 = Tr[R_Y R_{X'} \\ &\quad - 2R_Y R_{X'} R_Z + R_Y R_Z R_{X'} R_Z].\end{aligned}$$

Fukumizu et al. (2008) show the consistency of the above estimators. We now provide an upper bound on the convergence rates of these estimators under certain assumptions.

Theorem 3. *(Consistency of operators) Assume $\Sigma_{YY}^{-3/4}\Sigma_{YX}\Sigma_{XX}^{-3/4}$ is Hilbert-Schmidt. Suppose ϵ_m satisfies $\epsilon_m \to 0$ and $\epsilon_m^3 m \to \infty$. Then, we have convergence in probability in HS norm i.e*

$$\|\widehat{V}_{YX}^{(m)} - V_{YX}\|_{HS} \xrightarrow{P} 0.$$

An upper bound on convergence rate of $\|\widehat{V}_{YX}^{(m)} - V_{YX}\|_{HS}$ is $O_p(\epsilon_m^{-3/2}m^{-1/2} + \epsilon_m^{1/4})$.

Proof Sketch (refer supplementary for complete proof). We can upper bound $\|\widehat{V}_{YX}^{(m)} - V_{YX}\|_{HS}$ by

$$\left\|\widehat{V}_{YX}^{(m)} - \left(\Sigma_{YY} + \epsilon_m I\right)^{-1/2}\Sigma_{YX}\left(\Sigma_{XX} + \epsilon_m I\right)^{-1/2}\right\|_{HS} + \left\|\left(\Sigma_{YY} + \epsilon_m I\right)^{-1/2}\Sigma_{YX}\left(\Sigma_{XX} + \epsilon_m I\right)^{-1/2} - V_{YX}\right\|_{HS}.$$

The first term can be shown to be $O_p\left(\epsilon_m^{-3/2}m^{-1/2}\right)$ using Lemma 3 (in the supplementary material). To prove the second part, consider the complete orthogonal systems $\{\xi_i\}_{i=1}^\infty$ and $\{\psi_i\}_{i=1}^\infty$ for $\mathcal{H}_X$ and $\mathcal{H}_Y$ with an eigenvalue $\lambda_i \geq 0$ and $\gamma_i \geq 0$ respectively. Using the definition of Hilbert-Schimdt norm, we can bound the square of second term by

$$\sum_{i,j=1}^\infty \left(\frac{\epsilon_m\lambda_i + \epsilon_m\gamma_j + \epsilon_m^2}{\lambda_i\gamma_j(\lambda_i + \epsilon_m)(\gamma_j + \epsilon_m)}\right)\langle\psi_j, \Sigma_{YX}\xi_i\rangle^2.$$

Using AM-GM inequality, and assuming $\epsilon_m \ll \lambda_1$, $\epsilon_m \ll \gamma_1$ and that $\Sigma_{YY}^{-3/4}\Sigma_{YX}\Sigma_{XX}^{-3/4}$ is Hilbert-Schmidt, the theorem follows. □

Theorem 4. *(Consistency of estimators) Assume $\Sigma_{YY}^{-3/4}\Sigma_{YX'}\Sigma_{X'X'}^{-3/4}$, $\Sigma_{ZZ}^{-3/4}\Sigma_{ZX'}\Sigma_{X'X'}^{-3/4}$ and $\Sigma_{YZ}^{-3/4}\Sigma_{YZ}\Sigma_{ZZ}^{-3/4}$ are Hilbert-Schmidt. Suppose ϵ_m satisfies $\epsilon_m \to 0$ and $\epsilon_m^3 m \to \infty$. Then, we have*

$$\widehat{D}_{HS}(X, Y) \xrightarrow{P} D_{HS}(X, Y),$$
$$\widehat{D}_{HS}(X, Y|Z) \xrightarrow{P} D_{HS}(X, Y|Z).$$

An upper bound on convergence rate of the estimator $\widehat{D}_{HS}(X, Y|Z)$ is $O_p(\epsilon_m^{-3/2}m^{-1/2} + \epsilon_m^{1/4})$.

The proof is in the supplementary material. The assumptions used in Theorems 3 and 4 depend on the probability distribution and the kernel. A natural question arises if such assumptions are necessary for estimating these measures. The following result answers this question affirmatively. The crux of the result lies in the fact that these dependence measures are intricately connected to functionals of probability

distributions that are typically hard to estimate. In particular, we have

$$D_{HS}(X,Y) = \int\int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{p^2_{XY}(x,y)}{p_X(x)p_Y(y)} - 1\right) d\mu_{\mathcal{X}} d\mu_{\mathcal{Y}}.$$

Since estimation of these functionals is tightly coupled with smoothness assumptions on probability distributions of the random variables, it is reasonable to assume that our assumptions have a similar effect. We prove the result for the special case where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ and $\mathcal{Z} = \emptyset$. Let $\mathbf{P}$ denote the set of all distributions in $[0,1]^d$.

Theorem 5. *Let $\widehat{D}_n$ denote an estimator of D_{HS} on sample size n. For any sequence of estimates $\{\widehat{D}_n\}$ and any sequence $\{a_n\}$ converging to 0, there exists a compact subset $\mathbf{P_0} \subset \mathbf{P}$ for which the uniform rate of convergence is slower than $\{a_n\}$. In other words,*

$$\liminf_n \sup_{\mathbf{P_0}} P\left(|\widehat{D}_n - D| \geq a_n\right) > 0.$$

where $D = D_{HS}(X,Y)$.

The proof is in the supplementary material. An important point to note is that while the dependence measures themselves are invariant to invertible transformations, the estimators do not possess this property. It is often desirable to have this invariance property for the estimators as well since we generally deal with finite sample estimators. This property can be achieved using copula transformation, which is our focus in the next section. Along with the theoretical analysis, we also provide a compelling justification for using copula transformation from a practical point of view in Section 6.

3. Copula Transformation

We review important properties of the copula of multivariate distributions in this section. The use of the transformation will be clear in the later sections. The copula plays an important role in the study of dependence among random variables. They not only capture the dependence between random variables but also help us construct dependence measures which are invariant to any strictly increasing transformation of the marginal variables.

Sklar's theorem is central to the theory of copulas. It gives the relationship between a multivariate random variable and its univariate marginals. Suppose $X = \left(X^1, \ldots, X^d\right) \in \mathbb{R}^d$ is a d-dimensional multivariate random variable. Let us denote the marginal cumulative distribution function (cdf) of X^j by $F^j_X : \mathbb{R} \to [0,1]$. In this paper, we assume that the marginals F^j_X are invertible. The copula is defined by Sklar's theorem as follows:

Theorem 6. *(Sklar's theorem). Let $H\left(x_1, \ldots, x_d\right) = Pr\left(X^1 \leq x_1, \ldots, X^d \leq x^d\right)$ be the multivariate cumulative distribution function with continuous marginals $\left\{F^j_X\right\}$. Then there exists a unique copula C such that*

$$H(x_1, \ldots, x_d) = C(F^1_X(x_1), \ldots, F^d_X(x_d)). \quad (1)$$

Conversely, if C is a copula and $\left\{F^j_X\right\}$ are marginal cdfs, then H given in Equation (1) is the joint distribution with marginals $\left\{F^j_X\right\}$.

Let $T_X = \left(T^1_X, \ldots, T^d_X\right)$ denote the transformed variables where $T_X = F_X\left(X\right) = \left(F^1_X(X^1), \ldots, F^d_X(X^d)\right) \in [0,1]^d$. Here, F_X is called copula transformation. The above theorem gives a one-to-one correspondence between the joint distribution of X and T_X. Furthermore, it provides a way to construct dependence measures over the transformed variables since we have information about the copula distribution. An interesting consequence of the copula transformation is that we can get invariance of dependence measures to any strictly increasing transformations of the marginal variables.

4. Hilbert-Schmidt Dependence Measure using Copulas

Consider random variables $X = \left(X^1, \ldots, X^d\right)$, $Y = \left(Y^1, \ldots, Y^d\right)$ and $Z = \left(Z^1, \ldots, Z^d\right)$. We focus on the problem of defining dependence measure $D_C\left(X,Y\right)$ and conditional dependence measure $D_C\left(X,Y|Z\right)$ using copula transformation. The next section generalizes the dependence measure to any set of random variables. Note that we have assumed that the random variables X, Y and Z are all $d-$dimensional for simplicity, but our results hold for random variables with different dimensions.

Let T_X, T_Y and T_Z be the copula transformed variables of X, Y and Z respectively. With slight abuse of notation, we use k_X, k_Y and k_Z to denote the kernels over transformed variables T_X, T_Y and T_Z respectively. In what follows, the kernels are functions of the form $[0,1]^d \times [0,1]^d \to \mathbb{R}$ since they are defined over the transformed random variables. We define the dependence among random variables X and Y as:

$$D_C\left(X,Y\right) = D_{HS}\left(T_X, T_Y\right) = \|V_{T_Y T_X}\|^2_{HS}.$$

The conditional dependence measure is defined as:

$$D_C\left(X,Y|Z\right) = D_{HS}\left(T_X, T_Y|T_Z\right) = \|V_{T_Y T_{X'}}\|^2_{HS},$$

where $T_{X'} = (T_X, T_Z)$. By Theorem 10 (in supplementary) and the fact that the marginal cdfs are invertible, it is easy to see that $D_C(X,Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$ and $D_C(X,Y|Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z$. Our goal is to estimate the dependence measures $D_C(X,Y)$ and $D_C(X,Y|Z)$ using the i.i.d samples $X_{1:m}$, $Y_{1:m}$ and $Z_{1:m}$. Suppose we have the copula transformed variables T_X, T_Y and T_Z, then we can use the estimators

$$\widehat{D}_C(X,Y) = \|\widehat{V}^{(m)}_{T_Y T_X}\|^2_{HS},$$
$$\widehat{D}_C(X,Y|Z) = \|\widehat{V}^{(m)}_{T_Y T_{X'}}\|^2_{HS}$$

for dependence and conditional dependence measures respectively. However, we only have i.i.d samples $X_{1:m}$, $Y_{1:m}$, $Z_{1:m}$, and the marginal cdfs are unknown to us. We have to get the empirical copula transformed variables through these samples by estimating the marginals distribution functions $\left\{F^j_X\right\}$, $\left\{F^j_Y\right\}$ and $\left\{F^j_Z\right\}$. These distribution functions can be estimated efficiently using the rank statistics. For $x \in \mathbb{R}$ and $x^j \in \mathbb{R}$ for $1 \le j \le d$, let

$$\widehat{F}^j_X(x) = \frac{1}{m}\left|\left\{i : 1 \le i \le m, x \le X^j_i\right\}\right|,$$
$$\widehat{F}_X\left(x^1,\ldots,x^d\right) = \left(\widehat{F}^1_X(x^1),\ldots,\widehat{F}^d_X(x^d)\right).$$

$\widehat{F}_X$ is called the empirical copula transformation of X. The samples $\left(\widehat{T}_{X1},\ldots,\widehat{T}_{Xm}\right) = \left(\widehat{F}_X(X_1),\ldots,\widehat{F}_X(X_m)\right)$, called the empirical copula, are estimates of true copula transformation. We can similarly define empirical copula transformations and empirical copula for random variables Y and Z. It should be noted that the samples of empirical copula $\left(\widehat{T}_{X1},\ldots,\widehat{T}_{Xm}\right)$ are not independent even though $X_{1:m}$ are i.i.d samples. We can now use the dependence estimators in (Fukumizu et al., 2008) using empirical copula $\left(\widehat{T}_{X1},\ldots,\widehat{T}_{Xm}\right)$ instead of the samples $(X_1,\ldots,X_m)$. Lemma 2 (in supplementary) shows that the empirical copula is a good approximation of i.i.d. samples $(T_{X1},\ldots,T_{Xm})$.

It is important to note the relationship between measures D_{HS} and D_C. The copula transformation can also be viewed as an invertible transformation and hence, by Theorem 2 we have $D_{HS} = D_C$. Though the measures are identical, their corresponding estimators $\widehat{D}_{HS}$ and $\widehat{D}_C$ are different. At this point, we should also emphasize the difference between our work and Póczos et al. (2012). Although both these works use copula trick to obtain invariance, in contrast to Póczos et al. (2012), we essentially get the same measure even after copula transformation. In other words, the copula transformation in our case does not change the dependence measure and therefore, provides an invariant finite sample estimator to D_{HS}. Thus, we provide a compelling case to use copulas for D_{HS}. Moreover, the invariance property extends naturally to conditional dependence measure in our case.

We now focus on the consistency of the proposed estimators $\widehat{D}_C(X,Y)$ and $\widehat{D}_C(X,Y|Z)$. We assume that the kernel functions k_X, k_Y and k_Z are bounded kernel functions and are Lipschitz continuous on $[0,1]^d$ i.e there exists a $B > 0$ such that

$$|k_X(x_1,x) - k_X(x_2,x)| \le B\|x_1 - x_2\|,$$

for all $x, x_1, x_2 \in [0,1]^d$. The gaussian kernel is one of the popular kernels which is not only universal but also bounded and Lipschitz continuous. In what follows, we assume that conditions required for Theorem 3 and 4 hold for the transformed variables as well. We now show the consistency of the dependence estimators and provide upper bounds on their rates of convergence.

Theorem 7. *(Consistency of copula dependence estimators) Assume kernels k_X, k_Y and k_Z are bounded and Lipschitz continuous. Suppose ϵ_m satisfies $\epsilon_m \to 0$ and $\epsilon^3_m m \to \infty$, then*

(i) $\widehat{D}_C(X,Y) \xrightarrow{P} D_C(X,Y)$.

(ii) $\widehat{D}_C(X,Y|Z) \xrightarrow{P} D_C(X,Y|Z)$.

An upper bound on convergence rate of estimators $\widehat{D}_C(X,Y)$ and $\widehat{D}_C(X,Y|Z)$ is $O_p(\epsilon_m^{-3/2}m^{-1/2} + \epsilon_m^{1/4})$.

Proof Sketch (refer supplementary for complete proof). We first show $\|\widehat{V}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} - V_{T_Y T_X}\|_{HS} \xrightarrow{P} 0$. Consider the following upper bound of $\|\widehat{V}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} - V_{T_Y T_X}\|_{HS}$:

$$\|\widehat{V}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{V}^{(m)}_{T_Y T_X}\|_{HS} + \|\widehat{V}^{(m)}_{T_Y T_X} - V_{T_Y T_X}\|_{HS}. \quad (2)$$

From Theorem 3, it is easy to see that the second term converges with $O_p(\epsilon_m^{-3/2}m^{-1/2} + \epsilon_m^{1/4})$. The first term can be proved to be bounded by $C\epsilon_m^{-3/2}\|\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{\Sigma}^{(m)}_{T_Y T_X}\|_{HS}$ on similar lines as Lemma 3. We then prove that $\|\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{\Sigma}^{(m)}_{T_Y T_X}\|_{HS}$ converges $O_p(m^{-1/2})$ (refer Lemma 1 in the supplementary material), thereby proving the overall rate to be $O_p(\epsilon_m^{-3/2}m^{-1/2})$. The convergence of the dependence estimators follows from the convergence of the operators. □

The above result shows that by using copula transformation, we can ensure that the invariance property holds for finite sample estimators without loss in statistical efficiency. It should be noted that slightly better rates of convergence can be obtained by more restrictive assumptions. It is also noteworthy that under the conditions assumed in this paper, the estimators do not suffer from the curse of dimensionality, and hence, can be used in high dimensions. Moreover, the estimators only use rank statistics rather than the actual values of $X_{1:m}$,$Y_{1:m}$ and $Z_{1:m}$. This provides us with robust estimators since an arbitrarily large outlier sample cannot affect the statistics badly. In addition, this also makes the estimators invariant to monotone increasing transformations.

Let us call the dependence measure defined above as pairwise dependence since it measures dependence between two random variables. Now, the question arises if this approach can be generalized to measure dependence amongst a set of random variables rather than just two random variables. We answer this question affirmatively in the next section.

5. Generalized Dependence Measures

Suppose $\mathbf{S} = \{S_1, \ldots, S_n\}$ is a set of random variables. Similar to our previous assumptions, we assume that random variables $\{S_j\}$ are d-dimensional. We would like to measure the dependence amongst the set of random variables. Recall $S_{i:j}$ represents the random variable $(S_i, \ldots, S_j)$. Note that $S_{i:j}$ is a random variable of $(j-i)d$ dimensions. With slight abuse of notation, the kernel $k_{i:j}$ corresponding to variable $S_{i:j}$ is defined appropriately. We now express the generalized dependence measure as a sum of pairwise dependence measures. For simplicity, we denote $D_C(S_1, \ldots, S_n)$ by $D_C(\mathbf{S})$. The dependence measures are defined as:

$$D_C(\mathbf{S}) = \sum_{j=1}^{n-1} D_C(S_j, S_{j+1:n}), \tag{3}$$

$$D_C(\mathbf{S}|Z) = \sum_{j=1}^{n-1} D_C(S_j, S_{j+1:n}|Z). \tag{4}$$

Note that the set dependence measure is a sum of $n-1$ pairwise dependence measures. The following result justifies the use of these dependence measures.

Theorem 8. *(Generalized dependence measure) If the product kernels $k_j k_{j+1:n}$ for $j = \{1, \ldots, n-1\}$ are universal, we have (i) $D_C(\mathbf{S}) = \mathbf{0} \Leftrightarrow (S_1, \ldots, S_n)$ are independent. (ii) $D_C(\mathbf{S}|Z) = 0 \Leftrightarrow (S_1, \ldots, S_n)$ are independent given Z.*

The proof is in the supplementary material. We can now use the pairwise dependence estimators for estimating set dependence. The following estimators are used for measuring dependence

$$\widehat{D}_C(\mathbf{S}) = \sum_{j=1}^{n-1} \widehat{D}_C(S_j, S_{j+1:n}), \tag{5}$$

$$\widehat{D}_C(\mathbf{S}|Z) = \sum_{j=1}^{n-1} \widehat{D}_C(S_j, S_{j+1:n}|Z). \tag{6}$$

Let us assume that the conditions required for Theorem 7 are satisfied. The following theorem states the consistency of the dependence estimators proposed above, and provides an upper bound on their rates of convergence.

Theorem 9. *(Consistency of generalized estimators) Suppose the kernels defined above are bounded and Lipschitz continuous, then*

(i) $\widehat{D}_C(\mathbf{S}) \xrightarrow{P} D_C(\mathbf{S})$.

(ii) $\widehat{D}_C(\mathbf{S}|Z) \xrightarrow{P} D_C(\mathbf{S}|Z)$.

An upper bound on convergence rate of $\widehat{D}_C(\mathbf{S})$ and $\widehat{D}_C(\mathbf{S}|Z)$ is $O_p(n\epsilon_m^{-3/2} m^{-1/2} + n\epsilon_m^{1/4})$.

Proof. The theorem follows easily from Theorem 7 and Equations (3), (4), (5) and (6). □

6. Experimental Results

In this section, we empirically illustrate the theoretical contributions of the paper. We compare the performance of $D_{HS}(X,Y)$ and $D_{HS}(X,Y|Z)$ (referred to as NHS) with $D_C(X,Y)$ and $D_C(X,Y|Z)$ (referred to as CHS), that is with and without copula respectively. In the following experiments, we choose gaussian kernels and choose σ by median heuristic. We fix $\epsilon_m = 10^{-6}$ for our experiments.

6.1. Synthetic Dataset

In the first simulation, we constructed the following random variables: $X^1 \sim U[0, 4\pi]$, $X^2 = (500\ V_1, 1000\tanh(V_2), 500\sinh(V_3))$, where $V_1,\ V_2,\ V_3 \sim U[0,1]$, and $Y = 1000\tanh(X^1)$. 200 sample points are generated using the distributions specified above. The task in this experiment was to choose a feature between X^1 and X^2 that contains the most information about Y. Note that Y is a deterministic function of X^1 and independent of X^2. Therefore, we expect the dependence for (X^1, Y) to be high and that of (X^2, Y) to be low. Figure 1 shows

dependence measure (DM) comparison of NHS and CHS. It can be seen that while CHS chooses the correct feature X^1, NHS chooses the incorrect feature X^2.

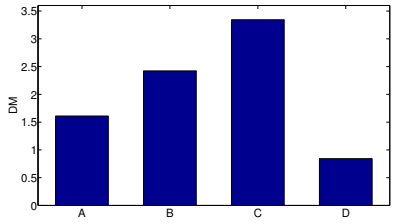

Figure 1. Columns (A) and (B) represent the dependence of (X^1, Y) and (X^2, Y) as measured by NHS while Columns (C) and (D) represent the dependence of (X^1, Y) and (X^2, Y) as measured by CHS.

The next simulation is designed to prove the significance of invariance property on finite samples. As mentioned earlier, though NHS is invariant to invertible transformation, its estimators do not retain this property. Thanks to copula transformation, CHS does not suffer from this issue. Let $X \sim U[0, 4\pi]$, $Y = 50\ X$ and $Z = 10 \tanh(Y/100)$. Note that Z is an invertible transformation of Y. The dependence of (X, Z) under CHS is not reported here as CHS is invariant to any monotone increasing transformations. We now demonstrate the asymptotic nature of the invariance property of NHS on the same data. Figure 2 clearly shows that while both methods have almost the same dependence measure for (X, Y), this dependence measured by NHS is reduced significantly when Y is transformed to Z. We can clearly see that NHS requires a large sample before it exhibits the invariance property. This problem furthur amplifies as we move to higher dimensions, making it undesirable for higher dimensional tasks.

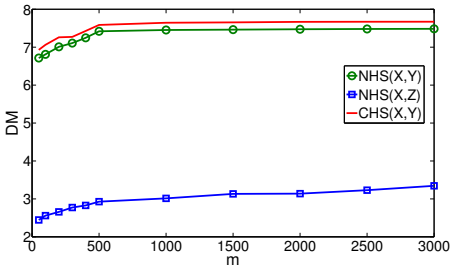

Figure 2. DM with varying number of samples.

We demonstrate the performance of conditional dependence measures in the following experiment. Let $X,\ V \sim U[0, 4\pi]$, $Y = 50 \sin(V)$ and $Z = \log(X + Y)$. Observe that though X and Y are independent, they become dependent when conditioned on Z. The results in Figure 3 clearly indicate that NHS fails to detect the dependence between X and Y when conditioned on Z while CHS successfully captures this dependence.

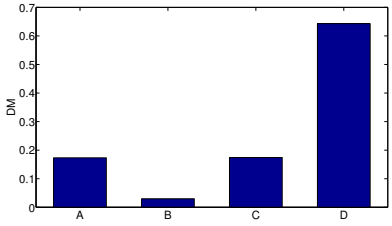

Figure 3. Columns (A) and (B) represent the dependence of (X, Y) and $(X, Y)|Z$ as measured by NHS while Columns (C) and (D) represent the dependence of (X, Y) and $(X, Y)|Z$ as measured by CHS.

6.2. Housing Dataset

We evaluate the performance of the dependence measures on the Housing dataset from the UCI repository. The importance of scale invariance on real-world data is demonstrated through this experiment. This dataset consists of 506 sample points each of which has 12 real valued and 1 integer valued attributes. We will only consider the real value attributes for this experiment and discard the integer attribute. Our aim is to predict the median value of owner-occupied homes based on other attributes like per capital crime, percentage of lower status of the population etc. This dataset is particularly interesting since it contains features of different nature and scale. In this experiment, we would like to predict the single most relevant feature for predicting the median value. Features 13 and 6 achieve the least prediction errors amongst all features using linear regressors (see supplementary material for more details). While CHS predicts these two features as the most relevant features, NHS performs poorly by selecting features 1 and 6 as the most relevant features.

7. Conclusion

In this paper we developed new dependence and condition dependence measures and estimators which are invariant to any monotone increasing transformations of the variables. We showed that under certain conditions the convergence rates of the estimators are polynomial, but when the conditions are less restrictive, then similarly to other existing estimators the convergence can be arbitrarily slow. We generalized these measures and estimators to sets of variables as well, and illustrated the applicability of our method with a few numerical experiments.

References

Bach, Francis R. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.

Baker, Charles R. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:pp. 273–289, 1973.

Borgwardt, K., Gretton, A., Rasch, M., Kriegel, H., Schölkopf, B., and Smola, A. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Fortet, R. and Mourier, E. Convergence de laréparation empirique vers la réparation théorique. *Ann. Scient. École Norm*, 70:266–285, 1953.

Fukumizu, K., Gretton, A., Sun, X., and Schoelkopf, B. Kernel measures of conditional dependence. In *NIPS 20*, pp. 489–496, Cambridge, MA, 2008. MIT Press.

Fukumizu, Kenji, Bach, Francis R., and Jordan, Michael I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Fukumizu, Kenji, Bach, Francis R., and Gretton, Arthur. Statistical convergence of kernel cca. In *Advances in Neural Information Processing Systems 18*, 2005.

Gretton, A., Herbrich, R., and Smola, A. The kernel mutual information. In *Proc. ICASSP*, 2003.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pp. 63–77, 2005.

Grünewälder, S, Lever, G, Baldassarre, L, Patterson, S, Gretton, A, and Pontil, M. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, volume 2, pp. 1823–1830, 2012.

Hewitt, E. and Stromberg, K. *Real and Abstract Analysis: A Modern Treatment of the Theory of Functions of a Real Variable*. Graduate Texts in Mathematics. Springer, 1975. ISBN 9780387901381.

Poczos, B. and Schneider, J. Nonparametric estimation of conditional information and divergences. In *International Conference on AI and Statistics (AISTATS)*, JMLR Workshop and Conference Proceedings, 2012.

Póczos, Barnabás, Ghahramani, Zoubin, and Schneider, Jeff G. Copula-based kernel dependency measures. In *ICML*, 2012.

Reed, M. and Simon, B. *Functional Analysis*. Academic Press, 1980. ISBN 9780080570488.

Rényi, A. On measures of dependence. *Acta. Math. Acad. Sci. Hungar*, 10:441–451, 1959.

Rényi, A. On measure of entropy and information. In *4th Berkeley Symposium on Math., Stat., and Prob.*, pp. 547–561, 1961.

Ritov, Y. and Bickel, P. J. Achieving information bounds in non and semiparametric models. *The Annals of Statistics*, 18(2):pp. 925–938, 1990.

Schweizer, B. and Wolff, E. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9, 1981.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794, 2007.

Tsallis, C. Possible generalization of boltzmann-gibbs statistics. *J. Statist. Phys.*, 52(1-2):479–487, 1988.

Supplementary Material

A. Proof of Theorem 2

Proof. We will prove the general case of conditional dependence measure since the other case follows trivially as a special case when $\mathcal{Z} = \emptyset$. The kernel-free property of the dependence measures is used to prove the result. The proof essentially uses *change of variables* formulas for transformation of random variables. From Theorem 1, we have

$$D_{HS}(X, Y|Z) = \int \int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - \frac{p_{X \perp\!\!\!\perp Y|Z}(x,y)}{p_X(x)p_Y(y)} \right)^2 p_X(x)p_Y(y)d\mu_{\mathcal{X}}d\mu_{\mathcal{Y}}.$$

Let $U = \Gamma_X(X)$, $V = \Gamma_Y(Y)$ and $W = \Gamma_Z(Z)$. Let $J_X = |det(\frac{d\Gamma_X^{-1}(y)}{dy})|$. We can similarly define J_Y and J_Z. We first observe that

$$\begin{aligned} p_{UV}(u,v) &= p_{XY}(\Gamma_X^{-1}(u), \Gamma_Y^{-1}(v))J_X J_Y, \\ p_U(u) &= p_X(\Gamma_X^{-1}(u))J_X. \end{aligned}$$

We can similarly calculate the joint probability and marginal distributions of other variables. Furthermore,

$$\begin{aligned} p_{U \perp\!\!\!\perp V|W}(u,v) &= \int_{\mathcal{W}} p_{U|W}(u|w)p_{V|W}(v|w)p_W(w)d\mu_{\mathcal{W}} \\ &= \int_{\mathcal{Z}} p_{X|Z}(\Gamma_X^{-1}(u)|\Gamma_Z^{-1}(w))\ J_X\ p_{Y|Z}(\Gamma_Y^{-1}(v)|\Gamma_Z^{-1}(w))J_Y\ p_Z(\Gamma^{-1}(w))\ J_Z\ \frac{d\mu_{\mathcal{Z}}}{J_Z} \\ &= \int_{\mathcal{Z}} p_{X|Z}(\Gamma_X^{-1}(u)|\Gamma_Z^{-1}(w))\ p_{Y|Z}(\Gamma_Y^{-1}(v)|\Gamma_Z^{-1}(w))p_Z(\Gamma^{-1}(w))d\mu_{\mathcal{Z}}\ J_X\ J_Y \\ &= p_{X \perp\!\!\!\perp Y|Z}(\Gamma^{-1}(u), \Gamma^{-1}(v))\ J_X\ J_Y. \end{aligned}$$

Using the above relations, we have

$$\begin{aligned} D_{HS}(U, V|W) &= \int \int_{\mathcal{U}\times\mathcal{V}} \left(\frac{p_{XY}(\Gamma_X^{-1}(u), \Gamma_Y^{-1}(v))\ J_X\ J_Y}{p_X(\Gamma_X^{-1}(u))p_Y(\Gamma_Y^{-1}(v))\ J_X\ J_Y} - \frac{p_{X \perp\!\!\!\perp Y|Z}(\Gamma_X^{-1}(u), \Gamma_Y^{-1}(v))\ J_X\ J_Y}{p_X(\Gamma_X^{-1}(u))p_Y(\Gamma_Y^{-1}(v))\ J_X\ J_Y} \right)^2 p_X(\Gamma_X^{-1}(u)) \\ &\qquad \times p_Y(\Gamma_Y^{-1}(v))\ J_X\ J_Y \frac{d\mu_{\mathcal{X}}}{J_X} \frac{d\mu_{\mathcal{Y}}}{J_Y} \\ &= \int \int_{\mathcal{X}\times\mathcal{Y}} \left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} - \frac{p_{X \perp\!\!\!\perp Y|Z}(x,y)}{p_X(x)p_Y(y)} \right)^2 p_X(x)p_Y(y)d\mu_{\mathcal{X}}d\mu_{\mathcal{Y}} \\ &= D_{HS}(X, Y|Z). \end{aligned}$$

□

B. Proof of Theorem 3

Proof. We will first prove the convergence of $\|\widehat{V}_{YX}^{(m)} - V_{YX}\|_{HS}$. It is easy to see that

$$\begin{aligned} \left\| \widehat{V}_{YX}^{(m)} - V_{YX} \right\|_{HS} \leq & \left\| \widehat{V}_{YX}^{(m)} - (\Sigma_{YY} + \epsilon_m I)^{-1/2} \Sigma_{YX} (\Sigma_{XX} + \epsilon_m I)^{-1/2} \right\|_{HS} \\ & + \left\| (\Sigma_{YY} + \epsilon_m I)^{-1/2} \Sigma_{YX} (\Sigma_{XX} + \epsilon_m I)^{-1/2} - V_{YX} \right\|_{HS}. \end{aligned}$$

From Lemma 3 (in the supplementary material), we know

$$\left\| \widehat{V}_{YX}^{(m)} - (\Sigma_{YY} + \epsilon_m I)^{-1/2} \Sigma_{YX} (\Sigma_{XX} + \epsilon_m I)^{-1/2} \right\|_{HS} = O_p\left(\epsilon_m^{-3/2} m^{-1/2} \right).$$

To prove the second part, consider the complete orthogonal systems $\{\xi_i\}_{i=1}^{\infty}$ and $\{\psi_i\}_{i=1}^{\infty}$ for $\mathcal{H}_X$ and $\mathcal{H}_Y$ such that $\Sigma_{XX}\xi_i = \lambda_i\xi_i$ with an eigenvalue $\lambda_i \geq 0$ and $\Sigma_{YY}\psi_i = \gamma_i\psi_i$ with an eigenvalue $\gamma_i \geq 0$ respectively. Now consider the second term,

$$\begin{aligned}
&\left\| (\Sigma_{YY} + \epsilon_m I)^{-1/2} \Sigma_{YX} (\Sigma_{XX} + \epsilon_m I)^{-1/2} - V_{YX} \right\|_{HS}^2 \\
&= \left\| (\Sigma_{YY} + \epsilon_m I)^{-1/2} \Sigma_{YX} (\Sigma_{XX} + \epsilon_m I)^{-1/2} - \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2} \right\|_{HS}^2 \\
&= \sum_{i,j=1}^{\infty} \left\langle \psi_j, \left((\Sigma_{YY} + \epsilon_m I)^{-1/2} \Sigma_{YX} (\Sigma_{XX} + \epsilon_m I)^{-1/2} - \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1/2} \right)\xi_i \right\rangle^2 \\
&= \sum_{i,j=1}^{\infty} \left\langle \psi_j, \frac{1}{(\lambda_i + \epsilon_m)^{1/2}(\gamma_j + \epsilon_m)^{1/2}}\Sigma_{YX}\xi_i - \frac{1}{\lambda_i^{1/2}\gamma_j^{1/2}}\Sigma_{YX}\xi_i \right\rangle^2 \\
&\leq \sum_{i,j=1}^{\infty} \left(\frac{\epsilon_m\lambda_i + \epsilon_m\gamma_j + \epsilon_m^2}{\lambda_i\gamma_j(\lambda_i + \epsilon_m)(\gamma_j + \epsilon_m)} \right) \langle \psi_j, \Sigma_{YX}\xi_i \rangle^2 .
\end{aligned}$$

The first transition follows from the definition of HS norm. Using arithmetic-geometric-harmonic mean inequality, we get

$$\frac{\epsilon_m\lambda_i + \epsilon_m\gamma_j + \epsilon_m^2}{(\lambda_i + \epsilon_m)(\gamma_j + \epsilon_m)} \leq \frac{1}{2}\left(\frac{\epsilon_m\lambda_i + \epsilon_m\gamma_j + \epsilon_m^2}{\lambda_i\gamma_j} \right)^{1/2}.$$

Assuming $\epsilon_m \ll \lambda_1$ and $\epsilon_m \ll \gamma_1$, we have

$$\frac{\epsilon_m\lambda_i + \epsilon_m\gamma_j + \epsilon_m^2}{\lambda_i\gamma_j} \leq \frac{2\epsilon_m(\lambda_1 + \gamma_1)}{\lambda_i\gamma_j}.$$

Using the above inequality, it is easy to see that,

$$\begin{aligned}
\sum_{i,j=1}^{\infty} \left(\frac{\epsilon_m\lambda_i + \epsilon_m\gamma_j + \epsilon_m^2}{\lambda_i\gamma_j(\lambda_i + \epsilon_m)(\gamma_j + \epsilon_m)} \right) \langle \psi_j, \Sigma_{YX}\xi_i \rangle^2 &\leq \frac{1}{\sqrt{2}} \sum_{i,j=1}^{\infty} \frac{\epsilon_m^{1/2}(\lambda_1 + \gamma_1)^{1/2}}{\lambda_i^{3/2}\gamma_j^{3/2}} \langle \psi_j, \Sigma_{YX}\xi_i \rangle^2 \\
&= O_p(\epsilon_m^{1/2}).
\end{aligned}$$

The last step is obtained by finiteness of

$$\frac{1}{\sqrt{2}} \sum_{i,j=1}^{\infty} \frac{1}{\lambda_i^{3/2}\gamma_j^{3/2}} \langle \psi_j, \Sigma_{YX}\xi_i \rangle^2 ,$$

which follows from our assumption that $\Sigma_{YY}^{-3/4}\Sigma_{YX}\Sigma_{XX}^{-3/4}$ is Hilbert-Schmidt. Therefore,

$$\|\widehat{V}_{YX}^{(m)} - V_{YX}\|_{HS} = O_p(\epsilon_m^{-3/2}m^{-1/2} + \epsilon_m^{1/4}).$$

The convergence rate of $\widehat{D}_{HS}(X, Y)$ follows from the above result by using triangle inequality and the fact that $\|\widehat{V}_{YX}^{(m)}\|_{HS} \leq 1$. □

Proof of Theorem 4

Proof. We have,

$$\begin{aligned}
&\|\widehat{V}_{YX|Z}^{(m)} - V_{YX|Z}\|_{HS} \\
&\leq \|\widehat{V}_{YX'}^{(m)} - V_{YX'}\|_{HS} + \|\widehat{V}_{YZ}^{(m)}\widehat{V}_{ZX'}^{(m)} - V_{YZ}V_{ZX'}\|_{HS}.
\end{aligned}$$

The first term can be bounded using Theorem 3. The second term can be upper bounded by

$$\left\| \left(\widehat{V}_{YZ}^{(m)} - V_{YZ} \right) V_{ZX'} \right\|_{HS} + \left\| \widehat{V}_{YZ}^{(m)} \left(\widehat{V}_{ZX'}^{(m)} - V_{ZX'} \right) \right\|_{HS}.$$

Using the fact that $\|\widehat{V}_{YZ}^{(m)}\|_{HS} \leq 1$ and Theorem 3, we have

$$\|\widehat{V}_{YX|Z}^{(m)} - V_{YX|Z}\|_{HS} = O_p(\epsilon_m^{-3/2} m^{-1/2} + \epsilon_m^{1/4}).$$

The convergence rate of $\widehat{D}_{HS}(X, Y|Z)$ follows from the above result by using triangle inequality, the fact that $\|\widehat{V}_{YX}^{(m)}\|_{HS}$, $\|\widehat{V}_{YZ}^{(m)}\|_{HS}$ and $\|\widehat{V}_{ZX}^{(m)}\|_{HS}$ are bounded and the operators are Hilbert-Schmidt. □

Proof of Theorem 5

Proof. For simplicity we will sketch the proof only for the 2-dimensional case ($P = P_{X,Y}(x, y)$). The higher dimensional case can be treated similarly. When the marginal distributions P_X and P_Y are uniform, then $D_{HS}(X, Y)$ has a very simple form:

$$\begin{aligned} D_{HS}(X, Y) &= \iint_{\mathcal{X}\times\mathcal{Y}} (p(x, y) - 1)^2 \, dxdy \\ &= \iint_{\mathcal{X}\times\mathcal{Y}} p^2(x, y) \, dxdy - 1. \end{aligned}$$

Ritov & Bickel (1990) proved that for 1-dimensional distributions there is a subset of distributions such that the uniform convergence rate for estimating $\int p^2(x)dx$ can be arbitrarily slow (Theorem 11).

All we have to show is that this theorem can be extended to the set of 2-dimensional continuous distributions that have uniform marginal distributions. For simplicity, let us denote $\iint_{\mathcal{X}\times\mathcal{Y}} p^2(x, y)dxdy$ by $\int p^2$. For one dimensional case, this is $\int_{\mathcal{X}} p^2(x)dx$.

The main idea in the proof of Ritov & Bickel (1990) is to reduce the $\int p^2$ estimation problem to a Bayesian two class classification problem. First, for each sample size n they construct a finite set of random densities $\mathbf{P}_{0n}$ in a specific way. The distribution of the random density $p \in \mathbf{P}_{0n}$ is denoted by $\pi_n(p)$.

The first class consists of densities $p \in \mathbf{P}_{0n}$ such that $\int p^2 = 1 + \frac{9}{12}a_n$. In the second class we have distributions $p \in \mathbf{P}_{0n}$ such that $\int p^2 = 1 + 3a_n$. The densities in $\mathbf{P}_{0n}$ are constructed such a way such that for the posterior probabilities we will have

$$\begin{aligned} \pi_n\left(\int p^2 = 1 + \frac{9}{12}a_n | X_1, \ldots, X_n \right) &= 1/2 + o_{\pi_n}(1), \\ \pi_n\left(\int p^2 = 1 + 3a_n | X_1, \ldots, X_n \right) &= 1/2 + o_{\pi_n}(1) \end{aligned}$$

This implies that even after having n samples, the probability to predict whether $\int p^2 = 1 + \frac{9}{12}a_n$ or $\int p^2 = 1 + 1 + 3a_n$ is close to $1/2$. From this it follows that

$$\inf_{\theta_n} P(|\theta_n - \int p^2| > a_n | X_1, \ldots, X_n) \to_{\pi_n} \frac{1}{2},$$

and thus $\int P[|\theta_n - \int p^2| > a_n]\pi_n dP \to 1/2$, which will prove that

$$\liminf_n \sup_{\mathbf{P_0}} P\left(|\theta_n - \int p^2| > a_n \right) \geq 1/2 > 0.$$

□

In Ritov & Bickel (1990), the main idea of the construction of the random densities is to split the $[0, 1]$ support uniformly to $m = n^3$ disjunct parts that is $[\frac{i-1}{m}, \frac{i}{m}]$ $(i = 1, \ldots, m)$, and define the random densities in each of these parts independently from each other such that for the density p either $\int p^2 = 1 + \frac{9}{12}a_n$ or $\int p^2 = 1 + 3a_n$ holds, and when there is only one observation in the $[(i - 1)/m, i/m]$ interval, then it will not provide any information about whether the random density p belongs to the first or the second class. It is easy to see that this construction can be generalized to two (and even higher dimensions) such a way that the marginal distributions can be kept uniform.

C. Proof of Theorem 7

Proof. In order to prove the consistency of the $\widehat{D}_C(X,Y)$, we need to show $\left\|\widehat{V}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} - V_{T_YT_X}\right\|_{HS} \xrightarrow{P} 0$. Consider the decomposition,

$$\left\|\widehat{V}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{V}^{(m)}_{T_YT_X}\right\|_{HS} + \left\|\widehat{V}^{(m)}_{T_YT_X} - V_{T_YT_X}\right\|_{HS}. \tag{7}$$

From Theorem 3, it is easy to see that the second term

$$\left\|\widehat{V}^{(m)}_{T_YT_X} - V_{T_YT_X}\right\|_{HS} \xrightarrow{P} 0$$

and it converges with $O_p(\epsilon_m^{-3/2}m^{-1/2} + \epsilon_m^{1/4})$. Now consider the first term,

$$\left\|\widehat{V}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{V}^{(m)}_{T_YT_X}\right\|_{HS} =$$
$$\left\| \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{-1/2} \widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_X\widehat{T}_X} + \epsilon_m I\right)^{-1/2} - \left(\widehat{\Sigma}^{(m)}_{T_YT_Y} + \epsilon_m I\right)^{-1/2} \widehat{\Sigma}^{(m)}_{T_YT_X} \left(\widehat{\Sigma}^{(m)}_{T_XT_X} + \epsilon_m I\right)^{-1/2} \right\|_{HS}.$$

This can be upper bounded by the following:

$$\left\| \left\{ \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{-1/2} - \left(\widehat{\Sigma}^{(m)}_{T_YT_Y} + \epsilon_m I\right) \right\} \widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_X\widehat{T}_X} + \epsilon_m I\right)^{-1/2} \right\|_{HS} \tag{8}$$
$$+ \left\| \left(\widehat{\Sigma}^{(m)}_{T_YT_Y} + \epsilon_m I\right)^{-1/2} \left\{ \widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{\Sigma}^{(m)}_{T_YT_X} \right\} \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_X\widehat{T}_X} + \epsilon_m I\right)^{-1/2} \right\|_{HS} \tag{9}$$
$$+ \left\| \left(\widehat{\Sigma}^{(m)}_{T_YT_Y} + \epsilon_m I\right)^{-1/2} \widehat{\Sigma}^{(m)}_{T_YT_X} \left\{ \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_X\widehat{T}_X} + \epsilon_m I\right)^{-1/2} - \left(\Sigma^{(m)}_{T_XT_X} + \epsilon_m I\right)^{-1/2} \right\} \right\|_{HS}. \tag{10}$$

The first term (8) in the above expression can be rewritten as

$$\left\| \left\{ \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{-1/2} \left\{ \left(\widehat{\Sigma}^{(m)}_{T_YT_Y} + \epsilon_m I\right)^{3/2} - \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{3/2} \right\} + \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} - \widehat{\Sigma}^{(m)}_{T_YT_Y}\right) \right\}$$
$$\times \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{-3/2} \widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_X\widehat{T}_X} + \epsilon_m I\right)^{-1/2} \Big\|_{HS}.$$

Using the facts $\left\| \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{-1/2} \right\|_{HS} \leq \frac{1}{\sqrt{\epsilon_m}}$,

$$\left\| \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{-1/2} \widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_X\widehat{T}_X} + \epsilon_m I\right)^{-1/2} \right\|_{HS} \leq 1$$

and Lemma 4, the above term can be upper bounded by

$$\frac{1}{\epsilon_m}\left\{ \frac{3}{\sqrt{\epsilon_m}} \max\left\{ \|\widehat{\Sigma}^{(m)}_{T_YT_Y} + \epsilon_m I\|^{1/2}_{HS}, \|\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\|^{1/2}_{HS} \right\} + 1 \right\} \|\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} - \widehat{\Sigma}^{(m)}_{T_YT_Y}\|_{HS}.$$

We similarly bound the third term (10). Again using the fact $\left\| \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_Y} + \epsilon_m I\right)^{-1/2} \right\|_{HS} \leq \frac{1}{\sqrt{\epsilon_m}}$ and $\left\| \left(\widehat{\Sigma}^{(m)}_{\widehat{T}_X\widehat{T}_X} + \epsilon_m I\right)^{-1/2} \right\|_{HS} \leq \frac{1}{\sqrt{\epsilon_m}}$, we can easily see that the second term is bounded by $\frac{1}{\epsilon_m}\|\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{\Sigma}^{(m)}_{T_YT_X}\|_{HS}$. Let us prove the following lemma which will be useful in completing the proof.

Lemma 1. *Suppose kernels k_X, k_Y and k_Z are bounded and Lipschitz continuous, then $\|\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{\Sigma}^{(m)}_{T_YT_X}\|_{HS} \xrightarrow{P} 0$ and its convergence rate is $O_p(m^{-1/2})$.*

Proof. We have

$$\|\widehat{\Sigma}^{(m)}_{\widehat{T}_Y\widehat{T}_X} - \widehat{\Sigma}^{(m)}_{T_YT_X}\|_{HS} =$$
$$\left\| \frac{1}{m}\sum_{i=1}^{m} \left\{ \left(k_Y(\cdot,\widehat{T}_{Yi}) - \widehat{\mu}^{(m)}_{\widehat{T}_Y}\right) \left\langle k_X(\cdot,\widehat{T}_{Xi}) - \widehat{\mu}^{(m)}_{\widehat{T}_X}, \cdot \right\rangle_{\mathcal{H}_X} - \left(k_Y(\cdot,T_{Yi}) - \widehat{\mu}^{(m)}_{T_Y}\right) \left\langle k_X(\cdot,T_{Xi}) - \widehat{\mu}^{(m)}_{T_X}, \cdot \right\rangle_{\mathcal{H}_X} \right\} \right\|_{HS}.$$

This can be upper bounded by using the following:

$$\Big\|\frac{1}{m}\sum_{i=1}^{m}\Big\{\Big(k_Y(\cdot,\widehat{T}_{Yi})-\widehat{\mu}^{(m)}_{\widehat{T}_Y}\Big)-\Big(k_Y(\cdot,T_{Yi})-\widehat{\mu}^{(m)}_{T_Y}\Big)\Big\}\Big\langle k_X(\cdot,\widehat{T}_{Xi})-\widehat{\mu}^{(m)}_{\widehat{T}_X},\cdot\Big\rangle_{\mathcal{H}_X}\Big\|_{HS}$$
$$+\Big\|\frac{1}{m}\sum_{i=1}^{m}\Big(k_Y(\cdot,\widehat{T}_{Yi})-\widehat{\mu}^{(m)}_{\widehat{T}_Y}\Big)\Big\{\Big\langle k_X(\cdot,\widehat{T}_{Xi})-\widehat{\mu}^{(m)}_{\widehat{T}_X},\cdot\Big\rangle_{\mathcal{H}_X}-\Big\langle k_X(\cdot,T_{Xi})-\widehat{\mu}^{(m)}_{T_X},\cdot\Big\rangle_{\mathcal{H}_X}\Big\}\Big\|_{HS}.$$

Using triangle inequality, the first term of the above expression upper bounded by the following decomposition

$$\frac{1}{m}\sum_{i=1}^{m}\Big\|\Big\{\Big(k_Y(\cdot,\widehat{T}_{Yi})-\widehat{\mu}^{(m)}_{\widehat{T}_Y}\Big)-\Big(k_Y(\cdot,T_{Yi})-\widehat{\mu}^{(m)}_{T_Y}\Big)\Big\}\Big\langle k_X(\cdot,\widehat{T}_{Xi})-\widehat{\mu}^{(m)}_{\widehat{T}_X},\cdot\Big\rangle_{\mathcal{H}_X}\Big\|_{HS}. \tag{11}$$

Observe that each $i=\{1,\dots,m\}$, we have

$$\Big\|\Big\{\Big(k_Y(\cdot,\widehat{T}_{Yi})-\widehat{\mu}^{(m)}_{\widehat{T}_Y}\Big)-\Big(k_Y(\cdot,T_{Yi})-\widehat{\mu}^{(m)}_{T_Y}\Big)\Big\}\Big\langle k_X(\cdot,\widehat{T}_{Xi})-\widehat{\mu}^{(m)}_{\widehat{T}_X},\cdot\Big\rangle_{\mathcal{H}_X}\Big\|^2_{HS}$$
$$\leq\Big\|\Big(k_Y(\cdot,\widehat{T}_{Yi})-\widehat{\mu}^{(m)}_{\widehat{T}_Y}\Big)-\Big(k_Y(\cdot,T_{Yi})-\widehat{\mu}^{(m)}_{T_Y}\Big)\Big\|^2_{\mathcal{H}_Y}\Big\|k_X(\cdot,\widehat{T}_{Xi})-\widehat{\mu}^{(m)}_{\widehat{T}_X}\Big\|^2_{\mathcal{H}_X}.$$

The previous step is obtained by using the definition of the HS norm. Since the kernel K_Y is Lipschitz Continuous, we know

$$\Big\|K_Y(.,T_{Yi})-K_Y(.,\widehat{T}_{Yi})\Big\|_{\mathcal{H}_Y}\leq B\Big\|\widehat{T}_{Yi}-T_{Yi}\Big\|$$

for some constant B. Moreover, the term

$$\Big\|k_X(\cdot,\widehat{T}_{Xi})-\widehat{\mu}^{(m)}_{\widehat{T}_X}\Big\|^2_{\mathcal{H}_X}$$

is bounded since the kernel K_X is bounded. Using the Lipschitz Continuity and bounded properties of kernel, it is easy to see that the expression (11) can bounded by

$$c\frac{1}{m}\sum_{i=1}^{m}\Big\|\widehat{T}_{Yi}-T_{Yi}\Big\|,$$

where c is some constant. Thanks to Lemma 2, it is easy to see that the above term is $O_p(m^{-1/2})$. By using a similar analysis, we can show that the second term is $O_p(m^{-1/2})$. □

Using the above lemma, it is easy to see that both the terms of 7 are $O_p(\epsilon_m^{-3/2}m^{-1/2})$. Hence the overall convergence rate is $O_p(\epsilon_m^{-3/2}m^{-1/2}+\epsilon_m^{1/4})$. Therefore,

$$\|\widehat{T}_Y-V_{T_YT_X}\|_{HS}=O_p(\epsilon_m^{-3/2}m^{-1/2}).$$

To prove the consistency and convergence rate of the dependence measures, we follow similar procedure as in Theorem 4 by using triangle inequality, and the facts that the operators are Hilbert-Schmidt and the HS norm of the estimators is bounded by 1. □

D. Proof of Theorem 8

Proof. Suppose $S_1,\dots,S_n$ are independent then it is easy to see that $D_C(S_1,\dots,S_n)=0$. Now, consider the case when $D_C(S_1,\dots,S_n)=0$. Then each term

$$D_C(S_j,S_{j+1:n})=0,\ \text{for } j=\{1,\dots,n-1\},$$

since they are non-negative. By product rule of probability

$$P(S_1,\dots,S_n)=\prod_{j=1}^{n-1}P(S_j|S_{j+1},\dots,S_n).$$

Since $D(S_j, S_{j+1:n})$ is 0, $P(S_j|S_{j+1}, \ldots, S_n) = P(S_j)$ for $j = \{1, \ldots, n-1\}$. Therefore,

$$P(S_1, \ldots, S_n) = \prod_{j=1}^{n-1} P(S_j).$$

Hence $(S_1, \ldots, S_n)$ are independent. The conditional dependence case can be proven similarly. □

E. Theorems & Lemmas used in this paper

In order to prove the results in our paper, we need the following theorems and lemmas (refer (Fukumizu et al., 2005; 2008; Ritov & Bickel, 1990) for details on these results).

Theorem 10. *(i) If the product $k_X k_Y$ is a universal kernel on $\mathcal{X} \times \mathcal{Y}$, then we have*

$$V_{YX} = O \Leftrightarrow X \perp\!\!\!\perp Y.$$

(ii) If the product $k_{X'} k_Y$ is a universal kernel on $(\mathcal{X} \times \mathcal{Z}) \times \mathcal{Y}$ and $k_{\mathcal{Z}}$ is universal, then

$$V_{YX'} = O \Leftrightarrow X \perp\!\!\!\perp Y|Z.$$

Theorem 11. *Let $a_n \in \mathbb{R}$ be a sequence converging to 0. Let $\theta_n = \theta_n(X_1, \ldots, X_n)$ a sequence of estimators for $D = \int p^2$, where $\{X_i\}$ is an i.i.d. series of random variables. Then there exists $\mathbf{P_0} \subset \mathbf{P}$, a compact subset of continuous distributions on $[0, 1]$ such that the uniform rate of convergence of θ_n is slower than a_n:*

$$\liminf_n \sup_{\mathbf{P_0}} P\left(|\widehat{D}_n - D| \geq a_n\right) > 0.$$

Lemma 2. *Let $X_{1:m}$ be an i.i.d sample from a probability distribution over $\mathbb{R}^d$ with marginal cdfs $\left\{F_X^j\right\}$. Let F_X and $\widehat{F}_X$ be e copula and empirical copula as defined above. Then, for any $\epsilon \geq 0$,*

$$Pr\left[\sup_{x \in \mathbb{R}^d} \|F_X(x) - \widehat{F}_X(x)\|_2\right] \leq 2d \exp\left(-\frac{2m\epsilon^2}{d}\right).$$

Lemma 3. *Suppose V_{YX} is Hilbert-Schmidt and $\epsilon_m \to 0$ as $m \to \infty$. Then we have*

$$\|\widehat{V}_{YX}^{(m)} - (\Sigma_{YY} + \epsilon_m I)^{-1/2} \Sigma_{YX} (\Sigma_{XX} + \epsilon_m I)^{-1/2} \| = O_p(\epsilon_m^{-3/2} m^{-1/2}).$$

Lemma 4. *Suppose A and B are positive, self-adjoint, Hilbert-Schmidt operators on a Hilbert space. Then,*

$$\|A^{3/2} - B^{3/2}\|_{HS} \leq 3 \left(\max\{\|A\|, \|B\|\}\right)^{1/2} \|A - B\|_{HS}.$$

F. Experiment Details

Housing Dataset

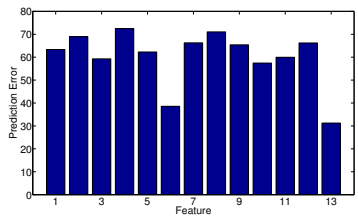

Figure 4. Prediction Error with linear regressors for all features

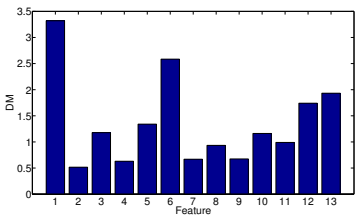

Figure 5. Dependence measure of features using NHS

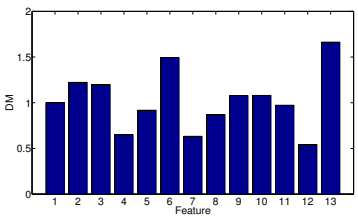

Figure 6. Dependence measure of features using CHS

We evaluate the performance of the dependence measures on the Housing dataset from the UCI repository. The importance of scale invariance on real-world data is demonstrated through this experiment. As already mentioned, our goal in the experiment was to predict the median value of owner-occupied homes based on other attributes. We used 300 instances for training and the rest of the data for testing. We trained linear regressors on each features in order to determine their explanatory strength. The prediction errors on the test are shown in Figure 4. The dependence measure estimates of NHS and CHS for all features are reported in Figures 5 and 6 respectively. It can be seen in these illustrations that CHS gives high dependence measures for most relevant features while NHS does not prefer the most relevant features.