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

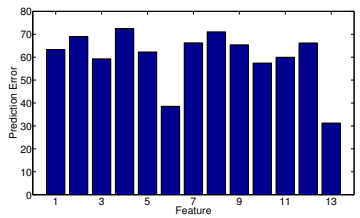

Figure 4. Prediction Error with linear regressors for all features

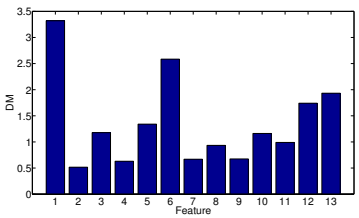

Figure 5. Dependence measure of features using NHS

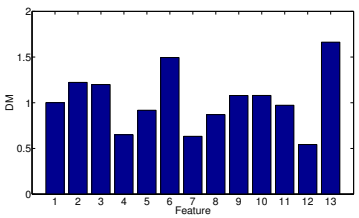

Figure 6. Dependence measure of features using CHS

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.