Supplement to "High-dimensional Variable Screening via Conditional Martingale Difference Divergence"

Lei Fang

Dr. Bing Zhang Department of Statistics, University of Kentucky,

725 Rose St. Lexington, Kentucky, 40536-0082, E-mail: lfa246@uky.edu

Qingcong Yuan

Biostatistics and Programming, Sanofi,

55 Corporate Drive. Bridgewater, New Jersey, 08807, E-mail: qingcong.yuan@gmail.com

Xiangrong Yin

Dr. Bing Zhang Department of Statistics, University of Kentucky,

725 Rose St. Lexington, Kentucky, 40536-0082, E-mail: yinxiangrong@uky.edu

Chenglong Ye

Dr. Bing Zhang Department of Statistics, University of Kentucky,

725 Rose St. Lexington, Kentucky, 40536-0082, E-mail: chenglong.ye@uky.edu

Supplementary Material

This supplementary file provides additional simulation results, and proofs of the proposition and theorems mentioned in the main article.

S1 Auxiliary Simulation Results

S1.1 Sensitivity Analysis of Using Different Bandwidths in $\mathrm{MC}_{\mathcal{H}}$

We consider the following example from Shao and Zhang (2014). Let $g_1(x) = x, g_2(x) = (2x-1)^2, g_3(x) = sin(2\pi x)/(2 - sin(2\pi x))$, and $g_4(x) = 0.1sin(2\pi x) + 0.2cos(2\pi x) + 0.3sin^2(2\pi x) + 0.4cos^3(2\pi x) + 0.5sin^3(2\pi x)$.

Example A. $Y = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + 1.5g_1(X_5) + 1.5g_2(X_6) + 1.5g_3(X_7) + 1.5g_4(X_8) + 2g_1(X_9) + 2g_2(X_{10}) + 2g_3(X_{11}) + 2g_4(X_{12}) + \sqrt{0.5184}\epsilon$, where the predictors $X_j, j = 1, ..., p$ are $i.i.d.$ from Unif(0,1), and ϵ is independent from the predictors and follows the standard normal distribution. We set $n = 400$ and $p = 1000$, and select $\lfloor n/log(n) \rfloor = 66$ variables.

bandwidth	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$	$\mathcal{P}_6$	$\mathcal{P}_7$	$\mathcal{P}_8$	$\mathcal{P}_9$	$\mathcal{P}_{10}$	$\mathcal{P}_{11}$	$\mathcal{P}_{12}$	$\mathcal{P}_{all}$
0.001	0.21	0.21	0.58	0.76	0.49	0.55	0.97	1.00	0.83	0.88	1.00	1.00	0.00
0.50	0.68	0.46	0.95	0.99	0.96	0.90	1.00	1.000	1.00	1.00	1.00	1.00	0.22
2	0.74	0.24	0.91	0.90	0.97	0.63	1.00	1.00	1.00	0.95	1.00	1.00	0.07
4	0.75	0.11	0.89	0.79	0.97	0.30	1.00	0.99	1.00	0.68	1.00	1.00	0.01
8	0.75	0.07	0.88	0.74	0.97	0.12	1.00	0.98	1.00	0.23	1.00	1.00	0.00
100	0.78	0.05	0.86	0.58	0.98	0.04	0.99	0.88	1.00	0.08	1.00	1.00	0.00

Table 7: Sensitivity on Example A based on 500 replications.

We report the results in Table 7. For X_1 that is linearly related to Y, $\mathcal{P}_1$ increases as the bandwidth increases. But for X_2, X_3 and X_4, the corresponding selection proportions decrease as the bandwidth increases except when the bandwidth is 0.001. Similar pattern can be observed with other predictors.

S1.2 Sensitivity Analysis against the Conditional Set

To demonstrate the numerical robustness of our method against the conditional set, we consider two types of perturbation to the conditional set. The first type is to change the number of variables (selected by $MC_{\mathcal{H}}$) in the conditional set, i.e., d_1. Although we recommend to use $d_1 = \lfloor \sqrt{n/log(n)} \rfloor$ in our paper, using slightly larger d_1 will not significantly worsen the performance of our method, compared to other methods. The second type is to switch a small fixed portion of the variables in the conditional set. Specially, suppose that the conditional set has $a \in \mathbb{Z}$ true active variables and $b \in \mathbb{Z}$ inactive variable. We randomly pick one inactive variable outside the conditional set to randomly replace one of the $b \in \mathbb{Z}$ inactive variable. We carry out the two perturbations on Examples 1, 2, 3. We also carry out the first perturbation to Example 4 under χ^2 error distribution. The results are demonstrated in Figures 1-4 below. For the second type of perturbation, since the curves are pretty flat in Figures 1-3, we calculate the standard deviation of the results over the replications and present them in Table 8. As can be seen, our method is overall the most robust and the most accurate method under both types of perturbation, especially against d_1.

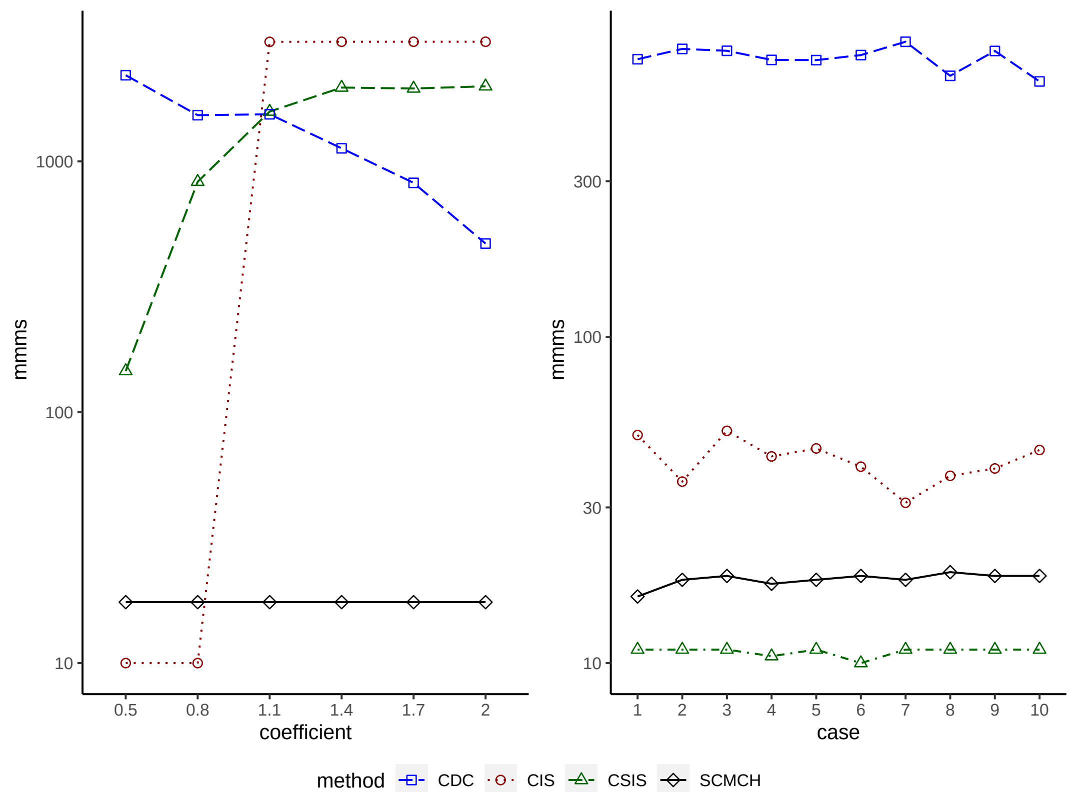

Figure 1: The median minimal model size (mmms) when (left) the conditional set is selected by $MC_{\mathcal{H}}$ where $d_1 = \lfloor \text{coefficient} \cdot \sqrt{n/log(n)} \rfloor$, and (right) the conditional set contains one true active predictor X_1 and another randomly selected nonactive predictor in Example 1 with $\rho = 0.5$.

We also investigate the performance of the methods when we select the conditional set by different methods on Example 2. Table 9 indicates our proposed method has an overall better performance against the choice of the conditional set in this case.

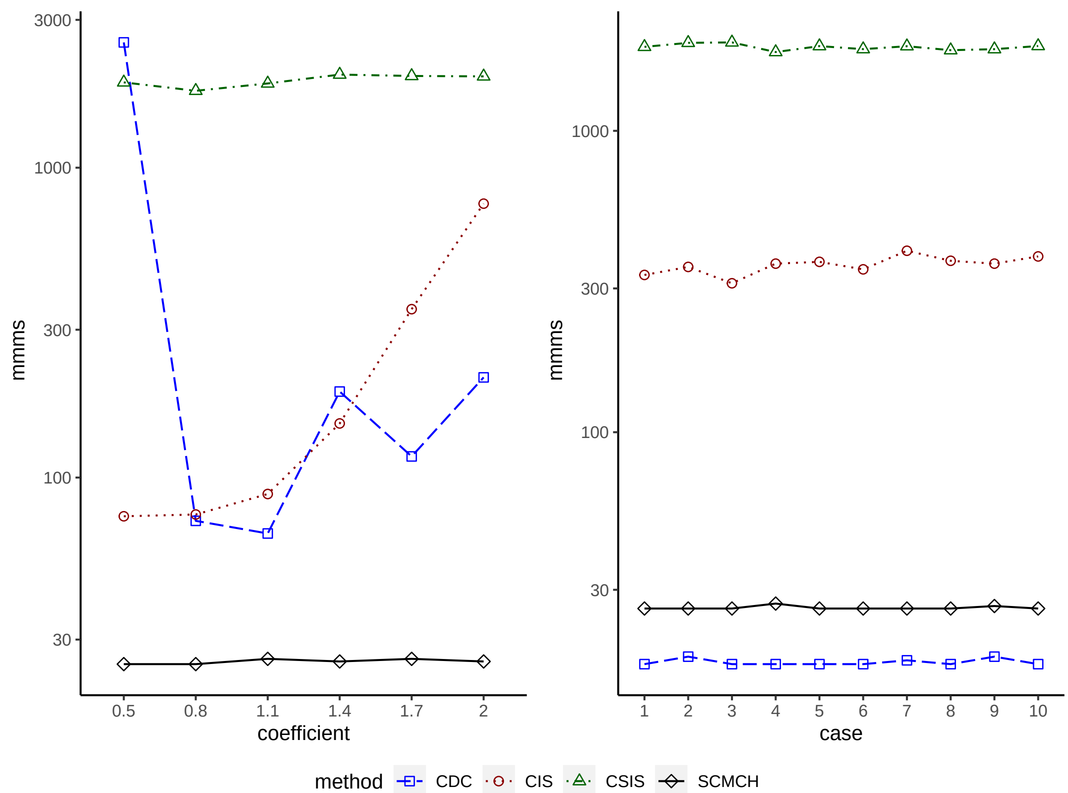

Figure 2: The median minimal model size (mmms) when (left) the conditional set is selected by $MC_{\mathcal{H}}$ where $d_1 = \lfloor \text{coefficient} \cdot \sqrt{n/log(n)} \rfloor$, and (right) the conditional set contains three true active predictor X_1, X_5, X_{10} and another randomly selected nonactive predictor in Example 2 with correlation ρ=0.9.

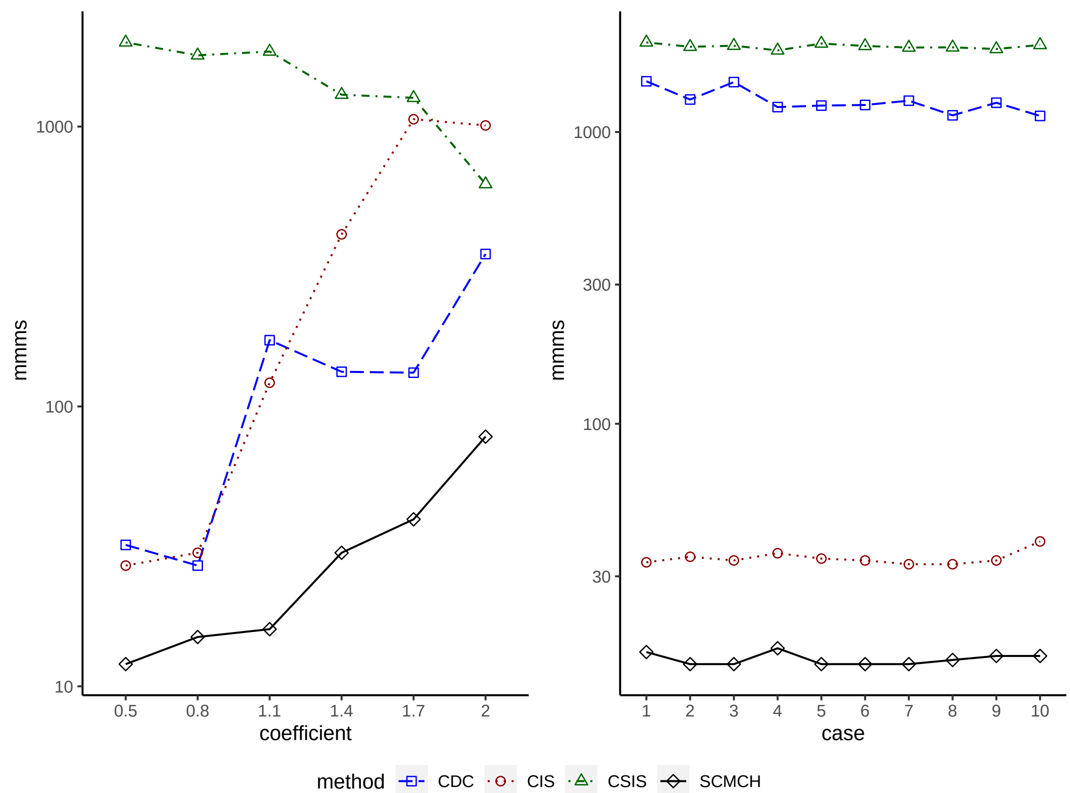

Figure 3: The median minimal model size (mmms) when (left) the conditional set is selected by $\mathrm{MC}_{\mathcal{H}}$ where $d_1 = \lfloor \text{coefficient} \cdot \sqrt{n/log(n)} \rfloor$, and (right) the conditional set contains two true active predictor X_1, X_5 and another randomly selected nonactive predictor in Example 3 with $\tau = 0.5$ and $\rho = 0.9$.

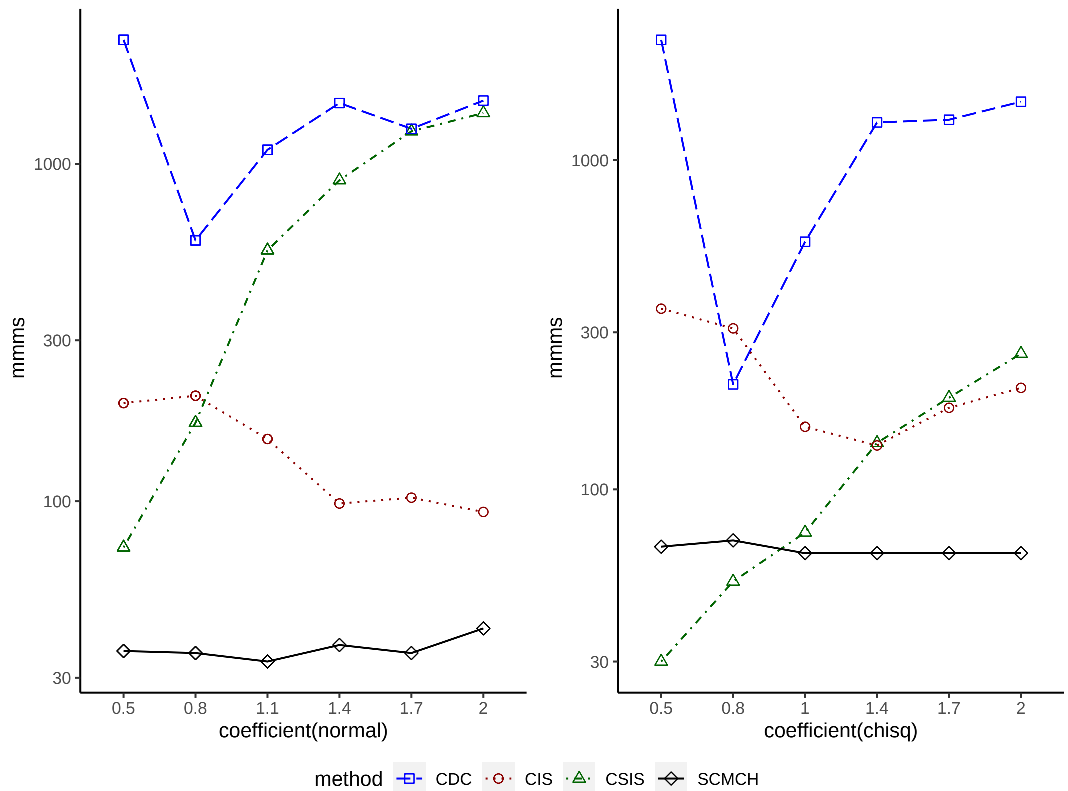

Figure 4: The median minimal model size (mmms) when the conditional set is selected by $\mathrm{MC}_{\mathcal{H}}$ where $d_1 = \lfloor \text{coefficient} \cdot \sqrt{n/log(n)} \rfloor$ in Example 4 with error term generated from (left) normal distribution and (right) χ^2 distribution.

	mmms (sd)		
Method	Example 1	Example 2	Example 3
CSIS	0.34	44.71	37.32
CDC-SIS	59.67	0.42	122.48
CIS	6.36	24.88	1.93
S-CMC$_{\mathcal{H}}$	0.83	0.34	0.74

Table 8: The standard deivation of the mmms Examples 1-3 when conditional set contains active conditional predictors and one inactive predictor.

S1.3 Performance under the Block Correlation Structure

The correlations among active predictors are 0.2 and 0.1 otherwise. In addition, the variance of predictor is set as 1. Table 10 shows the superior performance of S-CMC$_{\mathcal{H}}$ with the block correlation structure among predictors.

S1.4 Additional Simulation Results of Example 2

We evaluate the selection probability for all the interaction terms in Example 2 with a modified dimension $p_1 = 100$. In particular, the input data now has $p_1+p_1(p_1-1)/2$ predictors. The results are presented in Table 11. We can see that if we consider all the two-way interaction terms as predictors, we are able to select all the six terms in the true model with most of the screening methods except RASE$_1$-eBIC. It is worth pointing out that the main effects X_{15}, X_{20}, X_{25} have low selection probability.

	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
Conditonal set selected by Lasso								
CSIS	0.97	0.96	0.92	0.02	0.02	0.06	0.00	2241.0
CDC-SIS	0.67	0.27	0.35	0.02	0.06	0.15	0.00	2178.0
CIS	0.92	0.69	0.70	0.00	0.00	0.00	0.00	2269.0
S-CMC$_{\mathcal{H}}$	0.95	0.94	0.82	0.04	0.02	0.04	0.00	2263.0
Conditonal set selected by SIS								
CSIS	0.97	0.93	0.91	0.03	0.00	0.04	0.00	2179.5
CDC-SIS	0.87	0.88	0.86	0.06	0.16	0.41	0.00	1285.5
CIS	0.99	0.97	0.95	0.02	0.02	0.04	0.00	2368.5
S-CMC$_{\mathcal{H}}$	0.94	0.96	0.94	0.12	0.36	0.75	0.02	1051.0
Conditonal set selected by Forward Regression								
CSIS	0.94	0.91	0.90	0.04	0.03	0.04	0.00	2347.0
CDC-SIS	0.89	0.85	0.87	0.04	0.20	0.48	0.01	1396.0
CIS	0.99	0.96	0.96	0.02	0.01	0.06	0.00	2496.0
S-CMC$_{\mathcal{H}}$	0.95	0.95	0.93	0.15	0.43	0.81	0.07	801.5

Table 9: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 2, with $\rho = 0$ and conditional set selected by different methods.

	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
MDC	0.98	1.00	0.98	0.22	0.27	0.27	0.05	536.0
CSIS(X_{S_1})	1.00	1.00	1.00	0.03	0.05	0.07	0.00	2199.5
CSIS(X_{S_2})	0.96	0.96	0.93	0.08	0.07	0.16	0.00	2574.5
CDC-SIS(X_{S_1})	1.00	1.00	1.00	0.05	0.10	0.22	0.00	1542.0
CDC-SIS(X_{S_2})	1.00	0.99	0.99	0.1	0.07	0.17	0.01	2190.5
DCSIS2	0.59	0.83	0.97	0.29	0.39	0.89	0.06	434.0
CIS(X_{S_1})	1.00	1.00	1.00	0.10	0.21	0.54	0.00	529.0
CIS(X_{S_2})	0.53	0.20	0.27	0.07	0.00	0.13	0.00	1901.5
S-CMC$_{\mathcal{H}}$(X_{S_1})	1.00	1.00	1.00	0.30	0.61	1.00	0.19	121.0
S-CMC$_{\mathcal{H}}$(X_{S_2})	0.98	0.99	0.97	0.25	0.42	0.83	0.10	273.0
RASE$_1$-eBIC	0.84	0.78	0.78	0.01	0.00	0.01	0.00	2245.0

Table 10: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 2, with block design correlation structure.

In this new example, we are indeed only screening the marginally active variables among the $p_1+p_1(p_1-1)/2$ predictors. Thus, if we include all the interaction effects as predictors, we are unable to demonstrate the ability of the conditional variable screening methods on screening conditionally active predictors.

	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{1,15}$	$\mathcal{P}_{5,20}$	$\mathcal{P}_{10,25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
MDC	0.92	0.89	0.92	0.01	0.00	0.04	0.86	1.00	1.00	0.63	23.0
CSIS(X_{S_2})	1.00	1.00	1.00	0.00	0.00	0.02	1.00	1.00	1.00	1.00	7.0
CDC-SIS(X_{S_2})	0.77	0.74	0.76	0.11	0.01	0.00	0.72	0.99	1.00	0.32	176.5
S-CMC$_{\mathcal{H}}$(X_{S_2})	0.99	0.97	0.99	0.13	0.43	0.62	0.91	1.00	1.00	0.86	13.0
RASE$_1$-eBIC	0.49	0.11	0.41	0.07	0.07	0.16	0.36	0.77	0.98	0.01	4238.5
CIS	1	1	0.99	0.02	0.04	0.06	0.61	0.91	0.96	0.51	35.5

Table 11: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 2 with $\rho = 0$ and consider all two-way interaction terms.

S1.5 Additional Tables

Method	τ	$\rho = 0$								$\rho = 0.9$							
		$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$	$\mathcal{P}_1$	$\mathcal{P}_5$	$\mathcal{P}_{10}$	$\mathcal{P}_{15}$	$\mathcal{P}_{20}$	$\mathcal{P}_{25}$	$\mathcal{P}_{all}$	$\mathcal{S}_{0.5}$
MDC	0.5	1.00	1.00	0.06	0.03	0.04	0.02	0.00	1910.0	1.00	1.00	0.97	0.49	0.06	0.05	0.49	71.0
	0.75	0.99	1.00	0.24	0.38	0.52	0.55	0.03	986.5	1.00	1.00	1.00	1.00	0.91	0.92	0.86	29.5
CSIS	0.5	1.00	1.00	0.00	0.04	0.24	0.17	0.00	2071.0	1.00	1.00	0.02	0.05	0.31	0.24	0.00	1957.0
	0.75	1.00	1.00	0.00	0.04	0.24	0.17	0.00	2345.5	1.00	1.00	0.02	0.05	0.31	0.24	0.00	2389.5
CDC-SIS	0.5	1.00	1.00	0.86	1.00	0.95	0.92	0.86	8.0	1.00	1.00	0.71	0.98	1.00	1.00	0.70	30.5
	0.75	1.00	1.00	0.86	1.00	0.95	0.92	0.76	18.0	1.00	1.00	0.71	0.98	1.00	1.00	0.70	30.5
QaSIS	0.5	1.00	1.00	0.16	0.16	0.28	0.27	0.03	1085.0	1.00	1.00	1.00	0.90	0.69	0.55	0.90	20.5
	0.75	0.92	0.99	0.14	0.38	0.69	0.72	0.03	600.0	0.99	1.00	1.00	0.97	0.99	0.97	0.92	35.5
DCSIS2	0.5	0.18	0.41	0.04	0.08	0.93	0.90	0.00	1443.5	0.07	0.15	0.06	0.05	1.00	1.00	0.00	918.0
	0.75	0.18	0.41	0.04	0.08	0.93	0.90	0.00	1443.5	0.07	0.15	0.06	0.05	1.00	1.00	0.00	918.0
CIS	0.5	1.00	1.00	0.92	1.00	0.96	1.00	0.92	15.0	1.00	1.00	0.98	1.00	1.00	1.00	0.98	23.0
	0.75	1.00	1.00	0.92	1.00	0.96	1.00	0.88	20.0	1.00	1.00	0.98	1.00	1.00	1.00	0.98	23.0
S-CMC$_{\mathcal{H}}$	0.5	1.00	1.00	0.64	1.00	0.06	0.01	0.64	37.5	1.00	1.00	0.93	0.99	0.29	0.35	0.93	13.0
	0.75	1.00	1.00	0.69	0.96	0.54	0.61	0.27	183.5	1.00	1.00	1.00	1.00	0.99	0.98	0.97	26.0
RaSE$_1$-eBIC	0.5	0.49	0.31	0.00	0.00	0.03	0.01	0.00	1816.0	0.25	0.07	0.00	0.00	0.04	0.03	0.00	2425.5
	0.75	0.49	0.31	0.00	0.00	0.00	0.00	0.00	2427.0	0.25	0.07	0.00	0.00	0.01	0.02	0.00	2427.0

Table 12: The $\mathcal{P}_i$, $\mathcal{P}_{all}$ and $\mathcal{S}_{0.5}$ in Example 3, with $\boldsymbol{X}_S = \{X_1, X_5\}$.

Method	MSE	paired t-test p-value
CSIS	1.381	$< 2.2 \times 10^{-16}$
CDC-SIS	1.446	$< 2.2 \times 10^{-16}$
CIS	1.326	$< 2.2 \times 10^{-16}$
S-CMC$_{\mathcal{H}}$	0.805	–

Table 13: The prediction accuracy in MALT data when conditional set selected by MC$_{\mathcal{H}}$.

S2 Proof of Theorem 1

(a) Let $g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2) = E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle}), g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1) = E(Ve^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1\rangle})$, and $g_{\boldsymbol{U}_2}(\boldsymbol{t}_2) = E(e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2\rangle})$. If we expand the $\text{CMD}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1)$ in the representation of characteristic functions, it is

$$\int |g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2) - g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2 w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2$$
$$= \int (|g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 - g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)\bar{g}_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\bar{g}_{\boldsymbol{U}_2}(\boldsymbol{t}_2)-$$
$$\bar{g}_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)g_{\mathbf{V},\boldsymbol{U}_\mathbf{1}}g_{\boldsymbol{U}_2}(\boldsymbol{t}_2) + |g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2|g_{\boldsymbol{U}_\mathbf{2}}(\boldsymbol{t}_2)|^2)w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2,$$

where

$$|g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}_1^{'}\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}_2^{'}\rangle}],$$
$$g_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)\bar{g}_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\bar{g}_{\boldsymbol{U}_2}(\boldsymbol{t}_2) = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}_1^{'}\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}_2^{''}\rangle}],$$
$$\bar{g}_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1,\boldsymbol{t}_2)g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)g_{\boldsymbol{U}_2}(\boldsymbol{t}_2) = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}_1^{'}\rangle}e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}_2^{''}\rangle}],$$
$$|g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2|g_{\boldsymbol{U}_\mathbf{2}}(\boldsymbol{t}_2)|^2 = E[VV'e^{i\langle \boldsymbol{t}_1,\boldsymbol{U}_1-\boldsymbol{U}_1^{'}\rangle}]\cdot E[e^{i\langle \boldsymbol{t}_2,\boldsymbol{U}_2-\boldsymbol{U}_2^{''}\rangle}].$$

The weight function $w_1(\boldsymbol{t}_1)$ and $w_2(\boldsymbol{t}_2)$ are integrable. With *Bochner's theorem*, for a translation-invariant positive-definite kernel $k(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x}-\boldsymbol{x}')$, we can immediately get

$$\begin{aligned}\text{CMD}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) =& E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}_1^{'})k_2(\boldsymbol{U}_2,\boldsymbol{U}_2^{'}))+ \\ & E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}_1^{'}))\cdot E(k_2(\boldsymbol{U}_2,\boldsymbol{U}_2^{'})) - 2E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}_1^{'})k_2(\boldsymbol{U}_2,\boldsymbol{U}_2^{''}))\end{aligned}$$

(b) To show $\mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \leq 1$, we can rewrite $\mathrm{CMD}^2_{\mathcal{H}}$ as

$$\begin{aligned}\mathrm{CMD}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) =& E[(E(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})) + VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'}) - E_{V,\boldsymbol{U}_1}(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})) - \\ & E_{V',\boldsymbol{U}_1'}(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'}))) \times (E(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) + k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'}) - \\ & E_{\boldsymbol{U}_2}(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) - E_{\boldsymbol{U}_2'}(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'}))].\end{aligned}$$

We have
$$\begin{aligned} v(k_2, \boldsymbol{U}_2) =& E(k_2^2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) + E^2(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) - 2E[k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'}) \cdot k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{''})] \\ =& E(E(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) + k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'}) - E_{\boldsymbol{U}_2}(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})) - E_{\boldsymbol{U}_2^{'}}(k_2(\boldsymbol{U}_2, \boldsymbol{U}_2^{'})))^2,\end{aligned}$$

and

$$\begin{aligned} v(k_{1_V}, \boldsymbol{U}_1) =& E(V^2(V')^2k_1^2(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})) + E^2(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})) - 2E[V^2V'V''k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'}) \cdot k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{''})] \\ =& E(E(VV'k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})) + k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'}) - E_{\boldsymbol{U}_1}(k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})) - E_{\boldsymbol{U}_1^{'}}(k_1(\boldsymbol{U}_1, \boldsymbol{U}_1^{'})))^2,\end{aligned}$$

where $\mathrm{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \leq 1$ follows from an application of the *Cauchy-Schwarz* inequality. Furthermore, it is trivial to see that $\mathrm{CMC}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) \geq 0$.

Recall that $\mathrm{CMC}^2_{\mathcal{H}}$ is a standardized version of $\mathrm{CMD}^2_{\mathcal{H}}$:

$$\mathrm{CMD}_{\mathcal{H}}{}^2(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \iint |E(Ve^{i(\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle + \langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle)}) - E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})|^2 w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2.$$

Thus, $\mathrm{CMC}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = 0$ is equivalent to $E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle}e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})$ for any $\boldsymbol{t}_1 \in \mathbb{R}^p$ and $\boldsymbol{t}_2 \in \mathbb{R}^q$. To prove "Given $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$, $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ a.s. if and only if $\mathrm{CMC}^2_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = 0$", it suffices to prove "Given $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$, $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ a.s. if and only if $E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle}e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})$ for any $\boldsymbol{t}_1 \in \mathbb{R}^p$ and $\boldsymbol{t}_2 \in \mathbb{R}^q$."

We first prove the " only if " direction. By definition, with probability being one,

$$\int V \frac{f(V, \boldsymbol{U}_1, \boldsymbol{U}_2)}{f(\boldsymbol{U}_1, \boldsymbol{U}_2)} dV =: E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1) := \int V \frac{f(V, \boldsymbol{U}_1)}{f(\boldsymbol{U}_1)} dV,$$

where by a slight abuse of notation, we denote $f(\cdot)$ as the corresponding probability density functions or the probability mass functions if the variables are discrete. Plugging in $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$ (i.e., $f(\boldsymbol{U}_1, \boldsymbol{U}_2) = f(\boldsymbol{U}_1)f(\boldsymbol{U}_2)$), we have almost surely

$$\int V \frac{f(V, \boldsymbol{U}_1, \boldsymbol{U}_2)}{f(\boldsymbol{U}_2)} dV = \int V f(V, \boldsymbol{U}_1) dV. \tag{S2.1}$$

Denote $g_1(\boldsymbol{U}_1) := \int V \frac{f(V, \boldsymbol{U}_1, \boldsymbol{U}_2)}{f(\boldsymbol{U}_2)} dV$ and $g_2(\boldsymbol{U}_1) := \int V f(V, \boldsymbol{U}_1) dV$. Recall a function $f(x)$ can be transformed to its Fourier transform $\hat{f}(\xi) := \int_{-\infty}^{\infty} f(x) e^{-i2\pi\xi x} dx$ when the Dirichlet's condition holds, i.e., the integral of $f(x)$ is finite over every finite measure of support of x. Since $g_1(\boldsymbol{U}_1) = g_2(\boldsymbol{U}_1)$ and they are Fourier transformable, taking the Fourier transformation (or discrete Fourier transformation in the case of discrete random variables) on $g_1(\boldsymbol{U}_1)$ yields

$$\hat{g}_1(-\boldsymbol{t}_1/2\pi) = \int_{\mathbb{R}^p} g_1(\boldsymbol{U}_1) e^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1 \rangle} d\boldsymbol{U}_1 = \int_{\mathbb{R}^p} g_2(\boldsymbol{U}_1) e^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1 \rangle} d\boldsymbol{U}_1, \tag{S2.2}$$

which is equivalent to $E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1 \rangle}|\boldsymbol{U}_2) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1 \rangle})$ a.s. for any $\boldsymbol{t}_1 \in \mathbb{R}^p$. Denote $g_3(\boldsymbol{U}_2) := f(\boldsymbol{U}_2) \cdot E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1 \rangle}|\boldsymbol{U}_2)$ and $g_4(\boldsymbol{U}_2) := f(\boldsymbol{U}_2) \cdot E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1 \rangle})$, and we have $g_3(\boldsymbol{U}_2) = g_4(\boldsymbol{U}_2)$. Again, taking the Fourier transformation on $g_3(\boldsymbol{U}_2)$ leads to

$$\hat{g}_3(-\boldsymbol{t}_2/2\pi) = \int_{\mathbb{R}^q} g_3(\boldsymbol{U}_2) e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2 \rangle} d\boldsymbol{U}_2 = \int_{\mathbb{R}^q} g_4(\boldsymbol{U}_2) e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2 \rangle} d\boldsymbol{U}_2, \tag{S2.3}$$

which is equivalent to $E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle} e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})$ for any $\boldsymbol{t}_1 \in \mathbb{R}^p$ and $\boldsymbol{t}_2 \in \mathbb{R}^q$. This completes the "only if" direction.

We next prove the " if" direction. By the reverse Fourier transformation (or known as Fourier inversion theorem): the function $f(x)$ can be recovered by its Fourier transform $\hat{f}(\xi)$. That is, $f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)e^{i2\pi\xi x}dx$. Since $E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle} e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})E(e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle})$ for any $t_1 \in \mathbb{R}^p$ and $t_2 \in \mathbb{R}^q$, we have

$$\int_{\mathbb{R}^q} g_3(\boldsymbol{U}_2)e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}d\boldsymbol{U}_2 = \int_{\mathbb{R}^q} g_4(\boldsymbol{U}_2)e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}d\boldsymbol{U}_2, \tag{S2.4}$$

which implies the equality of their Fourier transform functions, i.e.,

$$\hat{g}_3(-\boldsymbol{t}_2/2\pi) = \hat{g}_4(-\boldsymbol{t}_2/2\pi).$$

Thus,

$$g_3(\boldsymbol{U}_2) = \int_{\mathbb{R}^q} \hat{g}_3(-\boldsymbol{t}_2/2\pi)e^{-i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}d\boldsymbol{U}_2 = \int_{\mathbb{R}^q} \hat{g}_4(-\boldsymbol{t}_2/2\pi)e^{-i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle}d\boldsymbol{U}_2 = g_4(\boldsymbol{U}_2).$$

By the definition of $g_3(\cdot)$ and $g_4(\cdot)$, dividing both sides of $g_3(\boldsymbol{U}_2) = g_4(\boldsymbol{U}_2)$ by $f(\boldsymbol{U}_2)$, we have

$$E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle}|\boldsymbol{U}_2) = E(Ve^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle})$$

a.s. for any $\boldsymbol{t}_1 \in \mathbb{R}^p$, which is equivalent to the equality (S2.2). Similar arguments of applying the reverse Fourier transformation to (S2.2) yields $g_1(\boldsymbol{U}_1) = g_2(\boldsymbol{U}_1)$. Dividing both sides of $g_1(\boldsymbol{U}_1) = g_2(\boldsymbol{U}_1)$ by $f(\boldsymbol{U}_1)$, together with the condition $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$, we have $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$ a.s.. This completes the "if" direction.

(c) The proof is straightforward by the definition of $\text{CMC}_{\mathcal{H}}(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$ and is omitted here.

(d) The proof is straightforward by the definition of $\text{MC}_{\mathcal{H}}(V|\boldsymbol{U})$ and is omitted here.□

S3 Derivation of the U-statistic of $\text{CMC}_{\mathcal{H}}$ in Definition 3

Recall the following notations: $a_{ij} = V_i V_j k_1(\boldsymbol{U}_{1i}, \boldsymbol{U}_{1j})$ and $b_{ij} = k_2(\boldsymbol{U}_{2i}, \boldsymbol{U}_{2j})$, $a^*_{ij} = a_{ij} - \dfrac{1}{n-2}\sum_{j=1}^n a_{ij} - \frac{1}{n-2}\sum_{i=1}^n a_{ij} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^n a_{ij}$, and $b^*_{ij} = b_{ij} - \frac{1}{n-2}\sum_{j=1}^n b_{ij} - \frac{1}{n-2}\sum_{i=1}^n b_{ij} + \frac{1}{(n-1)(n-2)}\sum_{i,j=1}^n b_{ij}$. In addition, we let $a_{i.} = \sum_{j=1}^n a_{ij}$, $a_{.j} = \sum_{i=1}^n a_{ij}$, $a_{..} = \sum_{i,j=1}^n a_{ij}$, $b_{i.} = \sum_{j=1}^n b_{ij}$, $b_{.j} = \sum_{i=1}^n b_{ij}$, $b_{..} = \sum_{i,j=1}^n b_{ij}$, $\bar{a}_{i.} = \frac{1}{n-2}a_{i.}$, $\bar{a}_{.j} = \frac{1}{n-2}a_{.j}$, $\bar{a}_{..} = \frac{1}{(n-1)(n-2)}a_{..}$. $\bar{b}_{i.} = \frac{1}{n-2}b_{i.}$, $\bar{b}_{.j} = \frac{1}{n-2}b_{.j}$, and $\bar{b}_{..} = \frac{1}{(n-1)(n-2)}b_{..}$.

Then,

$$\begin{aligned}\sum_{i\neq j} A_{i,j}B_{i,j} =& \sum_{i\neq j}(a_{ij}b_{ij} - a_{ij}\bar{b}_{i.} - a_{ij}\bar{b}_{.j} + a_{ij}\bar{b}_{..} - \bar{a}_{i.}b_{ij} + \bar{a}_{i.}\bar{b}_{i.} + \bar{a}_{i.}\bar{b}_{.j} - \bar{a}_{i.}\bar{b}_{..}\\ & - \bar{a}_{.j}b_{ij} + \bar{a}_{.j}\bar{b}_{i.} + \bar{a}_{.j}\bar{b}_{.j} - \bar{a}_{.j}\bar{b}_{..} + \bar{a}_{..}b_{ij} - \bar{a}_{..}\bar{b}_{i.} - \bar{a}_{..}\bar{b}_{.j} + \bar{a}_{..}\bar{b}_{..})\\ =& \sum_{i\neq j} a_{ij}b_{ij} - \sum_i a_{i.}\bar{b}_{i.} - \sum_j a_{.j}\bar{b}_{.j} + a_{..}\bar{b}_{..}\\ & - \sum_i \bar{a}_{i.}b_{i.} + (n-1)\sum_i \bar{a}_{i.}\bar{b}_{i.} + \sum_{i\neq j}\bar{a}_{i.}\bar{b}_{.j} - (n-1)\sum_j \bar{a}_{.j}\bar{b}_{..}\\ & - \sum_j \bar{a}_{.j}b_{.j} + \sum_{i\neq j}\bar{a}_{i.}\bar{b}_{.j} + (n-1)\sum_i \bar{a}_{i.}\bar{b}_{i.} - (n-1)\sum_j \bar{a}_{.j}\bar{b}_{..}\\ & + \bar{a}_{..}b_{..} - (n-1)\sum_i \bar{a}_{..}\bar{b}_{i.} - (n-1)\sum_j \bar{a}_{..}\bar{b}_{.j} + n(n-1)\bar{a}_{..}\bar{b}_{..}.\end{aligned}$$

Let $T_1 = \sum_{i\neq j} a_{ij}b_{ij}$, $T_2 = a_{..}b_{..}$, $T_3 = \sum_i a_{i.}b_{i.}$. Then,

$$\begin{aligned}\sum_{i\neq j} A_{i,j}B_{i,j} =& T_1 - \frac{1}{n-2}T_3 - \frac{1}{n-2}T_3 + -\frac{1}{(n-1)(n-2)}T_2\\ & - \frac{1}{n-2}T_3 + \frac{n-1}{(n-2)^2}T_3 + \frac{1}{(n-2)^2}(T_2 - T_3) - \frac{1}{(n-2)^2}T_2\\ & - \frac{1}{n-2}T_3 + \frac{1}{(n-2)^2}(T_2 - T_3) + \frac{n-1}{(n-2)^2}T_3 - \frac{1}{(n-2)^2}T_2\\ & + \frac{1}{(n-1)(n-2)}T_2 - \frac{1}{(n-2)^2}T_2 - \frac{1}{(n-2)^2}T_2 + \frac{n}{(n-1)(n-2)^2}T_2\\ =& T_1 - \frac{2}{n-2}T_3 + \frac{1}{(n-1)(n-2)}T_2.\end{aligned}$$

Let $(n)_k = n!/(n-k)!$ and I_k^n be the collections of k-tuples of indices (chosen from 1,2, ...,n) such that each index occurs exactly once. Then corresponding U-statistics estimator is

$$E(VV'k_1(\boldsymbol{U}_{1i},\boldsymbol{U}^{'}_{1j})k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) = (n)_2^{-1}E(\sum_{(i,j)\in I_2^n}^{n} V_iV_jk_1(\boldsymbol{U}_{1i},\boldsymbol{U}^{'}_{1j})k_2(\boldsymbol{U}_{1i},\boldsymbol{U}^{'}_{1j}))$$
$$= (n)_2^{-1}E(T_1),$$

$$E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1))\cdot E(k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) = (n)_4^{-1}E(\sum_{(i,j,q,r)\in I_2^n}^{n} V_iV_jk_1(\boldsymbol{U}_{1i},\boldsymbol{U}^{'}_{1j})k_2(\boldsymbol{U}_{2q},\boldsymbol{U}^{'}_{2r}))$$
$$= (n)_4^{-1}E(T_2 - 4T_3 + 2T_1),$$

and

$$E(VV'k_1(\boldsymbol{U}_1,\boldsymbol{U}^{'}_1)k_2(\boldsymbol{U}_2,\boldsymbol{U}^{'}_2)) = (n)_3^{-1}E(\sum_{(i,j,r)\in I_2^n}^{n} V_iV_jk_1(\boldsymbol{U}_{1i},\boldsymbol{U}^{'}_{1j})k_2(\boldsymbol{U}_{2i},\boldsymbol{U}^{'}_{2r})) = (n)_3^{-1}E(T_2 - T_1).$$

Combine the expectations, we get $\text{CMD}_{\mathcal{H}} = E(T_1 - \frac{2}{n-2}T_3 + \frac{1}{(n-1)(n-2)}T_2)$. Then the unbiased estimator is $T_1 - \frac{2}{n-2}T_3 + \frac{1}{(n-1)(n-2)}T_2$, which can be simplified to the format in Definition 3. □

S4 Proof of Theorem 2

If $\mathbb{E}(V^2) < \infty$, we need to prove almsot surely convergence for the following two expressions.

$$lim_{n\to\infty}\widehat{\text{CMD}}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) = \text{CMD}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1), \tag{S4.5}$$

and

$$lim_{n\to\infty}\widehat{\text{CMC}}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) = \text{CMC}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1). \tag{S4.6}$$

By the *Strong law of large numbers* for U-statistics, we can immediately get the result. □

S5 Proof of Theorem 3

In this proof, we require the assumptions that $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$, and $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$.

(a). We define the process

$$\Gamma_n(\boldsymbol{t}_1, \boldsymbol{t}_2) := \sqrt{n}(g^n_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) - g^n_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1) g^n_{\boldsymbol{U}_2}(\boldsymbol{t}_2)),$$

where $g^n_{V,\boldsymbol{U}_1,\boldsymbol{U}_2} = \frac{1}{n}\sum_{j=1}^n V_i e^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_{1j}\rangle} e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_{2j}\rangle}$, $g^n_{V,\boldsymbol{U}_1} = \frac{1}{n}\sum_{j=1}^n V_i e^{i\langle \boldsymbol{t}_1, \boldsymbol{U}_{1j}\rangle}$ and $g^n_{\boldsymbol{U}_2} = \frac{1}{n}\sum_{j=1}^n e^{i\langle \boldsymbol{t}_2, \boldsymbol{U}_{2j}\rangle}$. Denote $\boldsymbol{s} = (\boldsymbol{t}_1, \boldsymbol{t}_2)$. Define the norm $\|\gamma(\boldsymbol{s})\|^2_{\mathcal{H}} = \iint |\gamma(\boldsymbol{t}_1, \boldsymbol{t}_2)|^2 w_1(\boldsymbol{t}_1) w_2(\boldsymbol{t}_2) d\boldsymbol{t}_1 d\boldsymbol{t}_2$ for complex function $\gamma(\cdot)$ defined on $\mathbb{R}^p \times \mathbb{R}^q$. Since $n\widehat{\mathrm{CMD}}_{\mathcal{H}}^2(V, \boldsymbol{U}_2|\boldsymbol{U}_1) = \|\Gamma_n(\boldsymbol{s})\|^2_{\mathcal{H}}$, our goal is to show

$$\|\Gamma_n(\boldsymbol{s})\|^2_{\mathcal{H}} \xrightarrow[n\to\infty]{D} \|\Gamma(\boldsymbol{s})\|^2_{\mathcal{H}}.$$

First, by some direct calculations and the aforementioned independent assumptions, we have $E(\Gamma_n(\boldsymbol{t}_1, \boldsymbol{t}_2)) = 0$ and

$$\begin{aligned}
&E[\Gamma_n(\boldsymbol{t}_1,\boldsymbol{t}_2)\overline{\Gamma_n(\boldsymbol{t}'_1, \boldsymbol{t}'_2)}]\\
=&E(n(g^n_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}_1, \boldsymbol{t}_2) - g^n_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)g^n_{\boldsymbol{U}_2}(\boldsymbol{t}_2))(\overline{g^n_{V,\boldsymbol{U}_1,\boldsymbol{U}_2}(\boldsymbol{t}'_1, \boldsymbol{t}'_2)} - \overline{g^n_{V,\boldsymbol{U}_1}(\boldsymbol{t}'_1)g^n_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)}))\\
=&\frac{(n-1)^2}{n^2}F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, \boldsymbol{t}_2 - \boldsymbol{t}'_2) + \frac{n-1}{n}g_{\boldsymbol{U}_2}(\boldsymbol{t}_2 - \boldsymbol{t}'_2)[\frac{1}{n}F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, 0) -\\
&g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\overline{g_{V,\boldsymbol{U}_1}(\boldsymbol{t}'_1)}] + \frac{n-1}{n}[g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)\overline{g_{V,\boldsymbol{U}_1}(\boldsymbol{t}'_1)} +\\
&\frac{n-2}{n}F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, 0)]g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)} - \frac{(n-1)^2}{n^2}(F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, \boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)} +\\
&g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)F(\boldsymbol{t}_1 - \boldsymbol{t}'_1, -\boldsymbol{t}'_2)),
\end{aligned}$$

where $F(\boldsymbol{t}_1, \boldsymbol{t}_2) = E(V^2 \exp(i\langle \boldsymbol{t}_1, \boldsymbol{U}_1\rangle)\exp(i\langle \boldsymbol{t}_2, \boldsymbol{U}_2\rangle))$. In particular,

$$\begin{aligned}
E|\Gamma_n(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 = &\frac{n-1}{n}E(V^2)(1 + \frac{n-2}{n}|g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) - \frac{n-1}{n}|g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2(1 - |g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) -\\
&\frac{(n-1)^2}{n^2}[F(0, \boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}'_2)} + g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)F(0, -\boldsymbol{t}'_2)].
\end{aligned}$$

Let $D(\delta) = \{\boldsymbol{s} : \delta \le |\boldsymbol{s}|_{p+q} \le 1/\delta\}$ for $\delta > 0$. We then define

$$Q_n(\delta) = \int_{D(\delta)} |\Gamma_n(\boldsymbol{s})|^2 dw \text{ and } Q(\delta) = \int |\Gamma(\boldsymbol{s})|^2 dw,$$

where $dw = w(\boldsymbol{s})d\boldsymbol{s}$. Note that in this paper, $w(\boldsymbol{s}) = w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)$. Also, for a given δ, $w(\boldsymbol{s})$ is bounded on $D(\delta)$.

The **outline of the proof** is as follows: if we can show

(i) $Q_n(\delta) \xrightarrow{D} Q(\delta)$ for each $\delta > 0$, ;

(ii) $\limsup_{n\to\infty} E|Q_n(\delta) \xrightarrow{D} \|\Gamma_n(\boldsymbol{s})\|^2_{\mathcal{H}}| \to 0$ as $\delta \to 0$;

(iii) $E|Q(\delta) - \|\Gamma(\boldsymbol{s})\|_{\mathcal{H}}| \to 0$ as $\delta \to 0$, then $\|\Gamma_n(\boldsymbol{s})\|^2 \xrightarrow[n\to\infty]{D} \|\Gamma(\boldsymbol{s})\|^2$ follows from Theorem 8.6.2 of Resnick (2019).

Given $\epsilon = 1/p > 0$, $p \in \mathbb{N}$, choose a partition $\{D_k\}_{k=1}^N$ of $D(\delta)$ into $N = N(\epsilon)$ measurable sets with diameter at most ϵ. Then we can write $Q_n(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma_n(\boldsymbol{s})|^2 dw$ and $Q(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma(\boldsymbol{s})|^2 dw$. Define $Q_n^p(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma_n(\boldsymbol{s}_k)|^2 dw$ and $Q^p(\delta) = \sum_{k=1}^N \int_{D_k} |\Gamma(\boldsymbol{s}_k)|^2 dw$, where $\{\boldsymbol{s}_k\}_{k=1}^N$ is a set of distinct points such that $\boldsymbol{s}_k \in D_k$. Note that the choice of $\{\boldsymbol{s}_k\}_{k=1}^N$ depends on the partition $\{D_k\}_{k=1}^N$ of $D(\delta)$.

Note that, for any fixed $\boldsymbol{s}$, $\xi_n(\boldsymbol{s})$ is a complex-valued U-statistics for estimating $E\xi_n(\boldsymbol{s}) = 0$. As known in the U-statistics theory, we can apply the Central Limit Theorem to obtain the asymptotic distribution of a U-statistic. Specifically, assume $U_n = \binom{n}{m}^{-1} \sum h(X_{i_1}, \cdots, X_{i_m})$ is a U-statistics of a parameter $\theta := Eh(X_{i_1}, \cdots, X_{i_m})$ with kernel h and order $m \leq n$. Define $\hat{U}_n = \sum_{i=1}^n E[U_n|X_i] - (n-1)\theta$. Then $\sqrt{n}(U_n - \theta)$ and $\sqrt{n}(\hat{U}_n - \theta)$ have the same limiting distribution since $\sqrt{n}(U_n - \hat{U}_n) \xrightarrow{p} 0$. Denote $h_1(t) = E[h(t, X_2, \cdots, X_m)]$. We can write

$$\sqrt{n}(\hat{U}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (mh_1(X_i) - \theta m).$$

Applying CLT to the RHS, we have $\sqrt{n}(\hat{U}_n - \theta) \xrightarrow{D} N(0, m^2 var(h_1(X_1)))$. In our case, by the assumptions $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$ and $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$, we have $\theta = 0$. So we can find some complex function h_s such $\frac{1}{\sqrt{n}} \sum_{i=1}^n (h_s(X_i) - 0)$ and $\sqrt{n}(\xi_n(\boldsymbol{s}) - 0)$ has the same limiting joint distribution of the real part and the imaginary part. With some calculations, it not hard to show that $\sqrt{n}(\xi_n(\boldsymbol{s}) - 0) \xrightarrow{D} \Gamma(\boldsymbol{s})$ for any given $\boldsymbol{s}$, where $\xrightarrow{D}$ means convergence in the joint distribution of the real part and the imaginary part. Now we consider a N-dimensional random variable $\mathbf{Z}_i = (h_{s_1}(X_i), ..., h_{s_N}(X_i))$. Denote $\boldsymbol{\xi}_n = (\xi_n(\boldsymbol{s}_1), ..., \xi_n(\boldsymbol{s}_N))$. We have that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i$ and $\sqrt{n}(\boldsymbol{\xi}_n - \mathbf{0})$ have the same limiting distribution. Applying multivariate central limit theorem to $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i$ and applying continuous mapping theorem of the mapping $f : x \to |x|^2$ for complex number x, we have

$$\begin{pmatrix} |\Gamma_n(\boldsymbol{s}_1)|^2 \\ |\Gamma_n(\boldsymbol{s}_2))|^2 \\ \vdots \\ |\Gamma_n(\boldsymbol{s}_N)|^2 \end{pmatrix} \xrightarrow{D} \begin{pmatrix} |\Gamma(\boldsymbol{s}_1)|^2 \\ |\Gamma(\boldsymbol{s}_2)|^2 \\ \vdots \\ |\Gamma(\boldsymbol{s}_N)|^2 \end{pmatrix}$$

Therefore, we have $Q_n^p(\delta) \to Q^p(\delta)$. Following Theorem 8.6.2 of Resnick (2019), (i) holds if we show

$$\limsup_{p\to\infty} E|Q^p(\delta) - Q(\delta)| = 0 \tag{S5.7}$$

and

$$\limsup_{p\to\infty} \limsup_{n\to\infty} E|Q_n^p(\delta) - Q_n(\delta)| = 0. \tag{S5.8}$$

Let $\beta_n(\epsilon) = \sup_{\boldsymbol{s},\boldsymbol{s}'} E\,||\Gamma_n(\boldsymbol{s})|^2 - |\Gamma_n(\boldsymbol{s}')|^2|$ and $\beta(\epsilon) = \sup_{\boldsymbol{s},\boldsymbol{s}'} E\,||\Gamma(\boldsymbol{s})|^2 - |\Gamma(\boldsymbol{s}')|^2|$, where the supremum is taken over all $\boldsymbol{s} = (\boldsymbol{t}_1, \boldsymbol{t}_2)$ and $\boldsymbol{s}' = (\boldsymbol{t}_1', \boldsymbol{t}_2')$, when $\delta <$

$|\boldsymbol{s}|_{p+q}, |\boldsymbol{s}'|_{p+q} < 1/\delta$ and $|\boldsymbol{s}-\boldsymbol{s}'|_{p+q} < \epsilon$. By checking the form of $\text{Cov}_\Gamma(\boldsymbol{s},\boldsymbol{s}')$ and by applying the Cauchy-Swartz inequality, we derive that

$$\begin{aligned}
\beta(\epsilon) &= \sup_{\boldsymbol{s},\boldsymbol{s}'} E|(\Gamma(\boldsymbol{s})-\Gamma(\boldsymbol{s}'))\overline{\Gamma(\boldsymbol{s})} + \Gamma(\boldsymbol{s}')(\overline{\Gamma(\boldsymbol{s})}-\overline{\Gamma(\boldsymbol{s}')})| \\
&\leq \sup_{\boldsymbol{s},\boldsymbol{s}'} E^{1/2}|\Gamma(\boldsymbol{s})-\Gamma(\boldsymbol{s}')|^2(E^{1/2}|\Gamma(\boldsymbol{s})|^2 + E^{1/2}|\Gamma(\boldsymbol{s}')|^2) \\
&\leq C\sup_{\boldsymbol{s},\boldsymbol{s}'} E^{1/2}|\Gamma(\boldsymbol{s})-\Gamma(\boldsymbol{s}')|^2 \\
&\leq \sup_{\boldsymbol{s},\boldsymbol{s}'} C|\text{Cov}_\Gamma(\boldsymbol{s},\boldsymbol{s}) - \text{Cov}_\Gamma(\boldsymbol{s},\boldsymbol{s}') - \text{Cov}_\Gamma(\boldsymbol{s}',\boldsymbol{s}) + \text{Cov}_\Gamma(\boldsymbol{s}',\boldsymbol{s}')|^{1/2},
\end{aligned}$$

where the second inequality holds for some constant C because $E|\Gamma(\boldsymbol{s})|^2$ is upper bounded by definition. Since the integrand functions inside Cov_Γ are uniformly continuous in $\boldsymbol{s} \in \mathcal{R}^{p+q}$, it can be easily shown that $\beta(\epsilon) \to 0$ as $\epsilon \to 0$. To show (S5.7), we notice that

$$\begin{aligned}
E|Q^p(\delta) - Q(\delta)| &= E\left|\int_{D(\delta)} |\Gamma(\boldsymbol{s})|^2 dw - \sum_{k=1}^{N}\int_{D_k} |\Gamma(\boldsymbol{s}_k)|^2 dw\right| \\
&= E\left|\sum_{k=1}^{N}\int_{D_k} (|\Gamma(\boldsymbol{s})|^2 - |\Gamma(\boldsymbol{s}_k)|^2) dw\right| \\
&\leq \beta(1/p)\int_{D(\delta)} dw \to 0, \text{as } p \to \infty
\end{aligned}$$

Using exactly the same argument, we can show (S5.8) holds, so (i) holds.

Proof of (ii). We first calculate

$$E\left|\int_{D(\delta)} |\Gamma_n(\boldsymbol{s})|^2 dw\right| - \int_{\mathcal{R}^{p+q}} |\Gamma_n(\boldsymbol{s})|^2 dw\Big| = \int_{|s|_{p+q}<\delta} E|\Gamma_n(\boldsymbol{s})|^2 dw + \int_{|s|_{p+q}>1/\delta} E|\Gamma_n(\boldsymbol{s})|^2 dw.$$

From the calculation of Theorem 3(a),

$$E|\Gamma_n(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 = \frac{n-1}{n}E(V^2)(1+\frac{n-2}{n}|g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) - \frac{n-1}{n}|g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2(1-|g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) - \frac{(n-1)^2}{n^2}[F(0,\boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}_2')} + g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)F(0,-\boldsymbol{t}_2')].$$

Thus,

$$\int E|\Gamma_n(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 dw = \frac{n-1}{n}E(V^2)[\int 1dw + \frac{n-2}{n}E(k_2(\boldsymbol{U}_2-\boldsymbol{U}_2')] - \frac{n-1}{n}E(VV'k_1(\boldsymbol{U}_1-\boldsymbol{U}_1'))(1-E(k_2(\boldsymbol{U}_2-\boldsymbol{U}_2')) - \frac{(n-1)^2}{n^2}2E(V^2k_2(\boldsymbol{U}_2-\boldsymbol{U}_2')).$$

Since kernel k_1 and k_2 are bounded, and by assumption, $E(V^2)$ is also bounded, therefore, the above $\int E|\Gamma_n(\boldsymbol{t}_1,\boldsymbol{t}_2)|^2 dw$ is bounded. By our assumptions on the weight functions (Remark 3), i.e., $\iint w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2 = 1$, $\int_{|(\boldsymbol{t}_1,\boldsymbol{t}_2)|_{p+q}<\delta} w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2 \to 0$ as $\delta \to 0$, and $\int_{|(\boldsymbol{t}_1,\boldsymbol{t}_2)|_{p+q}>1/\delta} w_1(\boldsymbol{t}_1)w_2(\boldsymbol{t}_2)d\boldsymbol{t}_1 d\boldsymbol{t}_2 \to 0$ as $\delta \to 0$, we have that for any small ϵ, there exist δ_0, n_0, such that when $n \geq n_0$ and $\delta \leq \delta_0$, $\int_{|\boldsymbol{s}|_{p+q}<\delta} E|\Gamma_n(\boldsymbol{s})|^2 dw < \epsilon$ and $\int_{|\boldsymbol{s}|_{p+q}>1/\delta} E|\Gamma_n(\boldsymbol{s})|^2 dw < \epsilon$. Thus we have shown (ii).

Proof of (iii). We can also show (iii) using similar arguments.

Thus, the desired result follows.

(b). According to the first assertion, we have

$$E||\Gamma||_{\mathcal{H}}^2 = \int Cov_\Gamma((\boldsymbol{t}_1, \boldsymbol{t}_2), (\boldsymbol{t}_1, \boldsymbol{t}_2))dw$$
$$= \int \{[E(V^2) - |g_{V,\boldsymbol{U}_1}(\boldsymbol{t}_1)|^2](1 - |g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2) + 2E(V^2)|g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)|^2 - F(0, \boldsymbol{t}_2)\overline{g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)} -$$
$$g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)F(0, -\boldsymbol{t}_2)\}dw.$$

Under the assumption $E(V^2|\boldsymbol{U}_2) = E(V^2)$, which implies $F(0, \boldsymbol{t}_2) = E(V^2)g_{\boldsymbol{U}_2}(\boldsymbol{t}_2)$, we have $E\|\Gamma\|_{\mathcal{H}}^2 = E(V^2) - E(V^2)E(k_2(\boldsymbol{U}_2 - \boldsymbol{U}_2')) - E(VV'k_1(\boldsymbol{U}_1 - \boldsymbol{U}_1')) + E(VV'k_1(\boldsymbol{U}_1 - \boldsymbol{U}_1')k_2(\boldsymbol{U}_2 - \boldsymbol{U}_2'))$. Let $S_n = (\frac{1}{n}\sum_i V_i^2 - \frac{1}{n(n-1)}\sum_{i\neq j} a_{ij})(1 - \frac{1}{n(n-1)}\sum_{i\neq j} b_{ij})$, where $a_{ij} = V_iV_jk_1(\boldsymbol{U}_{1i}, \boldsymbol{U}_{1j})$ and $b_{ij} = k_2(\boldsymbol{U}_{2i}, \boldsymbol{U}_{2j})$. Then, by the SLLN for U-statistics, $S_n \xrightarrow[n\to\infty]{a.s} E\|\Gamma\|_{\mathcal{H}}^2$. The below proof follows a similar argument in the proof of Corollary 2 of Szekely er al.(2007), we present the whole steps for the completeness. The defined Γ_n and Γ are zero mean process, according to Kuo (2006), Chapter 1, Section 2, the squared norm $\|\Gamma\|_{\mathcal{H}}^2$ of the zero-mean Gaussian process Γ has the representation

$$\|\Gamma\|_{\mathcal{H}}^2 \overset{D}{=} \sum_{j=1}^{\infty} \lambda_j Z_j \tag{S5.9}$$

where Z_j are independent standard normal random variables, and the nonnegative constants $\{\lambda_j\}$ depend on the distribution of $(\boldsymbol{U}_1, \boldsymbol{U}_2, V)$. Also, under $\boldsymbol{U}_1 \perp \boldsymbol{U}_2$ and $E(V|\boldsymbol{U}_1, \boldsymbol{U}_2) = E(V|\boldsymbol{U}_1)$, $n\widehat{\mathrm{CMD}}_{\mathcal{H}}^2(V, \boldsymbol{U}_2|\boldsymbol{U}_1)$ converges in distribution to $\|\Gamma\|_{\mathcal{H}}^2$.

Therefore $n\widehat{\mathrm{CMD}}_{\mathcal{H}}^2(V, \boldsymbol{U}_2|\boldsymbol{U}_1)/S_n \xrightarrow[n\to\infty]{D} Q$, where E(Q)=1 and Q is a nonnegative quadratic form of centered Gaussian random variable.

(c). Suppose that $\mathrm{CMD}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) > 0$, since $\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) \xrightarrow[n\to\infty]{a.s.} \mathrm{CMD}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) > 0$, therefore $n\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1) \xrightarrow[n\to\infty]{a.s.} \infty$. By the SLLN, S_n converges to a constant, and therefore $n\widehat{\mathrm{CMD}}_{\mathcal{H}}(V,\boldsymbol{U}_2|\boldsymbol{U}_1)/S_n \xrightarrow[n\to\infty]{a.s.} \infty$. □

S6 Proof of Theorem 4

To avoid the abuse of subscripts and ease the notation, we use Z to denote $\boldsymbol{X}_S$ in this section. Let $S_1^j = E[YY'k_1(Z,Z')k_2(X_j,X_j')]$, $S_2^j = E[YY'k_1(Z,Z')]E[k_2(X_j,X_j')]$ and $S_3^j = E[YY'k_1(Z,Z')k_2(X_j,X_j'')]$, where (X_j',Y',Z') and (X_j'',Y'',Z'') are *i.i.d.* copies of (X_j,Y,Z). Denote $(n)_k = n!/(n-k)!$ and let I_k^n be the collections of k-tuples of indices (chosen from $\{1,2,...,n\}$) such that each index occurs exactly once. Correspondingly, their unbiased sample counterparts are

$$S_{1n}^j = (n)_2^{-1} \sum_{(k,l)\in I_2^n} Y_kY_lk_1(Z_k,Z_l)k_2(X_{jk},X_{jl}),$$

$$S_{2n}^j = (n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n} Y_kY_lk_1(Z_k,Z_l)k_2(X_{jh},X_{jq}),$$

$$S_{3n}^j = (n)_3^{-1} \sum_{(k,l,h)\in I_3^n} Y_kY_lk_1(Z_k,Z_l)k_2(X_{jk},X_{jh}).$$

Since $(\mathrm{CMD}^j_{\mathcal{H}})^2 = S_1^j + S_2^j - 2S_3^j$ and $(\widehat{\mathrm{CMD}}^j_{\mathcal{H}})^2 = S_{1n}^j + S_{2n}^j - 2S_{3n}^j$. We will establish the consistency of S_{1n}^j, S_{2n}^j and S_{3n}^j separately.

Consistency of S_{1n}^j. Since S_{1n}^j is a U-statistic with the corresponding kernel function

$$h_1(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_l) := Y_k Y_l k_1(Z_k, Z_l) k_2(X_{jk}, X_{jl}).$$

Rewrite

$$S_{1n}^j = \frac{1}{n(n-1)} \sum_{k \neq l} h_1 I\{|h_1| \leq M\} + \frac{1}{n(n-1)} \sum_{k \neq l} h_1 I\{|h_1| > M\} := S_{1n,1}^j + S_{1n,2}^j.$$

Correspondingly, its population counterpart can also be decomposed as

$$S_1^j = E[h_1 I\{|h_1| \leq M\}] + E[h_1 I\{|h_1| > M\}] := S_{1,1}^j + S_{1,2}^j.$$

Note that $S_{1n,1}^j$ and $S_{1n,2}^j$ are unbiased estimators of $S_{1,1}^j$ and $S_{1,2}^j$, respectively.

(I) To show the consistency of $S_{1n,1}^j$, we note that all the U-statistics can be expressed as an average of averages of i.i.d. random variables (see Serfling (2009), Section 5.1.6). Denote $m = \lfloor n/2 \rfloor$, and define $\Omega(X_{j1}, Y_1, Z_1; ...; X_{jn}, Y_n, Z_n) = \frac{1}{m} \sum_{r=0}^{m-1} h_1^{(r)} I\{|h_1^{(r)}| \leq M\}$, where $h_1^{(r)} = h_1(X_{j\ 1+2r}, Y_{1+2r}, Z_{1+2r}; X_{j\ 2+2r}, Y_{2+2r}, Z_{2+2r})$. Then, we have

$$S_{1n,1}^j = (n!)^{-1} \sum_{n!} \Omega(X_{ji_1}, Y_{i_1}, Z_{i_1}; ...; X_{ji_n}, Y_{i_n}, Z_{i_n}),$$

where the summation is over $n!$ possible permutations $(i_1, ..., i_n)$ of $(1, ..., n)$. By

Jensen's inequality, for $t > 0$, we have

$$E[\exp(tS^j_{1n,1})] = E[\exp\{t(n!)^{-1}\sum_{n!}\Omega(X_{ji_1}, Y_{i_1}, Z_{i_1}; ...; X_{ji_n}, Y_{i_n}, Z_{i_n})\}$$
$$\le (n!)^{-1}\sum_{n!} E[\exp(t\sum_{r=0}^{m-1} h_1^{(r)} I\{|h_1^{(r)}| \le M\}/m)]$$
$$= E^m[\exp(th_1^{(r)} I\{|h_1^{(r)}| \le M\}/m)].$$

It implies

$$P(S^j_{1n,1} - S^j_{1,1} \ge \epsilon) \le \exp(-t\epsilon)\exp(-tS^j_{1,1})E[\exp(t\tilde{S}^j_{1n,1})]$$
$$\le \exp(-t\epsilon) \cdot E^m\{\exp[t(h_1^{(r)} I\{|h_1^{(r)}| \le M\} - S^j_{1,1})/m]\}$$
$$\le \exp(-t\epsilon) \cdot \exp\{t^2M^2/(2m)\},$$

where the first and third inequalities are due to *Markov's inequality* and *Hoeffding's inequality* (see Lemma 1 of Li et al. (2012)), respectively. Set $t = \epsilon m/M^2$. By the symmetry of U-statistics, we then obtain $P(|S^j_{1n,1} - S^j_{1,1}| \ge \epsilon) \le 2\exp\{-\epsilon^2 m/(2M^2)\}$.

(II) Next we show the consistency of $S^j_{1n,2}$. By *Cauchy-Schwarz inequality* and *Markov's inequality*, we have

$$(S^j_{1,2})^2 = (E[h_1 I\{|h_1| > M\}])^2 \le E[h_1^2] \cdot P(|h_1| > M)$$
$$\le E[h_1^2]E[|h_1|^q] \cdot M^{-q}$$

for any $q \in \mathbb{N}$. Since $|ab| \le (a^2 + b^2)/2$ for any $a, b \in \mathbb{R}$, we get

$$|h_1(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_l)| \le \frac{1}{2}(Y_k^2 + Y_l^2)k_1(Z_k, Z_l)k_2(X_{jk}, X_{jl}) \le K^2Y_k^2$$

as the kernel k_1 and k_2 are upper bounded by some constant $K > 0$. Hence $E[|h_1|^q]$ is bounded based on Assumption (A1). Thus, if we let $M = n^\gamma$ for $0 < \gamma < 1/2 - \kappa$, then $S^j_{1,2} \leq \epsilon/2$ for sufficiently large n (in the sense that we take $\epsilon = cn^{-\kappa}$ and q can be any integer greater than $2\kappa/\gamma$). Hence, $P(|S^j_{1n,2} - S^j_{1,2}| \geq \epsilon) \leq P(|S^j_{1n,2}| \geq \epsilon/2)$. Since the event $\{|S^j_{1n,2}| \geq \epsilon/2\}$ implies $\{Y_k^2 \geq M/K^2,\ \text{for some}\, 1 \leq k \leq n\}$, we have

$$\begin{aligned} P\{|S^j_{1n,2}| \geq \epsilon/2\} &\leq P(\cup_{k=1}^n \{Y_k^2 \geq M/K^2\}) \\ &\leq \sum_{k=1}^n P(\{Y_k^2 \geq M/K^2\}) \\ &\leq nP(\{Y_k^2 \geq M/K^2\}). \end{aligned}$$

By Assumption (A1) and *Markov's inequality*, there exists a constant C such that $P(\{Y_k^2 \geq M/K^2\} \leq C\exp(-sM/K^2)$ for any k and $s \in (0, 2s_0]$. Consequently, for sufficiently large n, $\max_{1\leq j\leq p} P(|S^j_{1n,2} - S^j_{1,2}| \geq \epsilon) \leq \max_{1\leq j\leq p} P(|S^j_{1n,2}| \geq \epsilon/2) \leq \max_{1\leq p\leq n} nP(\{Y_k^2 \geq M/K^2\}) \leq nC\exp(-sM/K^2)$. In combination with the convergence result of $S^j_{1n,1}$, for large enough n, we have

$$\begin{aligned} P(|S^j_{1n} - S^j_1| \geq 2\epsilon) &\leq P(|S^j_{1n,1} - S^j_{1,1}| \geq \epsilon) + P(|S^j_{1n,2} - S^j_{1,2}| \geq \epsilon) \\ &\leq 2\exp(-\epsilon^2 n^{1-2\gamma}/4) + Cn\exp(-sn^\gamma/K^2). \end{aligned}$$

Consistency of S^j_{2n}. We rewrite S^j_{2n} as follows:

$$\begin{aligned}S^j_{2n} =&(n)_4{}^{-1} \sum_{k<l<h<q} 4 \cdot [Y_kY_l k_1(Z_k, Z_l)k_2(X_{jh}, X_{jq})\\ &\qquad + Y_kY_h k_1(Z_k, Z_h)k_2(X_{jl}, X_{jq})\\ &\qquad + Y_kY_q k_1(Z_k, Z_q)k_2(X_{jl}, X_{jh})\\ &\qquad + Y_lY_h k_1(Z_l, Z_h)k_2(X_{jq}, X_{jk})\\ &\qquad + Y_lY_q k_1(Z_l, Z_q)k(X_{jh}, X_{jk})\\ &\qquad +Y_hY_q k_1(Z_h, Z_q)k_2(X_{jl}, X_{jk})]\\ :=&24(n)_4^{-1} \sum_{k<l<h<q} h_2(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_l; X_{jh}, Y_h, Z_h; X_{jq}, Y_q, Z_q),\end{aligned}$$

where $h_2(X_{jk}, Y_k, Z_{jk}; X_{jl}, Y_l, Z_{jl}; X_{jh}, Y_h, Z_{jh}; X_{jq}, Y_q, Z_{jq})$ is the kernel function. Following the same argument in (I), we write S^j_{2n} as $S^j_{2n} = 24(n)_4^{-1} \sum_{k<l<h<q} h_2 I(|h_2| \leq M) + 24(n)_4^{-1} \sum_{k<l<h<q} h_2 I(|h_2| \geq M) = S^j_{2n,1} + S^j_{2n,2}$ and their population versions $S^j_2 = E[h_2 I\{|h_2| \leq M\}] + E[h_2 I\{|h_2| \geq M\}] = S^j_{2,1} + S^j_{2,2}$. Using the same argument as for $S^j_{1n,1}$, we can show that

$$P(|S^j_{2n,1} - S^j_{2,1}| \geq \epsilon) \leq 2\exp\{-\epsilon^2 m'/(2M^2)\},$$

where $m' = \lfloor n/4 \rfloor$ since S^j_{2n} is a fourth-order U-statistics.

Now it remains to establish the uniform convergence of the other part $S^j_{2n,2}$. Note that $|h_2(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_l; X_{jh}, Y_h, Z_h; X_{jq}, Y_q, Z_q)| \leq [\frac{1}{4}(Y_k^2+Y_l^2+Y_h^2+Y_q^2)]*K^2$, so the event $\{|S^j_{2n}| \geq \epsilon/2\}$ implies the event $\{Y_k^2 \geq M/K^2\}$ for some $1 \leq k \leq n$.

Therefore, following the similar argument as in the part of $S^j_{1n,1}$, we have $P(|S^j_{2n,2} - S^j_{2,2}| \geq \epsilon) \leq P(|S^j_{2n,2}| \geq \epsilon/2) \leq P(\cup_{k=1}^n [Y_k^2 \geq M/K^2]) \leq Cn\exp(-sM/K^2)$ for any k and $s \in (0, 2s_0]$. Combining the two convergence results for $S^j_{2n,1}$ and $S^j_{2n,2}$ with $M = n^\gamma$ for some $0 < \gamma < 1/2 - \kappa$, it follows that $P(|S^j_{2n} - S^j_2| \geq 2\epsilon) \leq 2\exp(-\epsilon^2 n^{1-2\gamma}/8) + Cn\exp(-sn^\gamma/K^2)$.

Consistency of S^j_{3n}. We rewrite S^j_{3n} as follows:

$$
\begin{aligned}
S^j_{3n} =& (n)_3^{-1} \sum_{k<l<h} [Y_k Y_l k_1(Z_k, Z_l) k_2(X_{jk}, X_{jh}) \\
& + Y_k Y_h k_1(Z_k, Z_h) k_2(X_{jk}, X_{jl}) \\
& + Y_l Y_k k_1(Z_l, Z_k) k_2(X_{jl}, X_{jh}) \\
& + Y_l Y_h k_1(Z_l, Z_h) k_2(X_{jl}, X_{jk}) \\
& + Y_h Y_k k_1(Z_h, Z_k) k_2(X_{jh}, X_{jl}) \\
& + Y_h Y_l k_1(Z_h, Z_l) k_2(X_{jh}, X_{jk})] \\
:=& 6(n)_3^{-1} \sum_{k<l<h} h_3(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_k; X_{jh}, Y_h, Z_h),
\end{aligned}
$$

where $h_3(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_k; X_{jh}, Y_h, Z_h)$ is the kernel function. Again, we write S^j_{3n} as $S^j_{3n} = 6\{n(n-1)(n-2)\}^{-1} \sum_{k<l<h} h_3 I(|h_3| \leq M) + 6\{n(n-1)(n-2)\}^{-1} \sum_{k<l<h} h_3 I(|h_3| \geq M) := S^j_{3n,1} + S^j_{3n,2}$ and its population counterpart as $S^j_3 = E[h_3 I\{|h_3| \leq M\}] + E[h_3 I\{|h_3| \geq M\}] := S^j_{3,1} + S^j_{3,2}$. Using the same ar-

gument as in the part of $S^j_{1n,1}$, we can show that

$$P(|S^j_{3n,1} - S^j_{3,1}| \geq \epsilon) \leq 2\exp\{-\epsilon^2 m'/(2M^2)\},$$

where $m' = \lfloor n/3 \rfloor$ since S^j_{3n} is a third-order U-statistic. It remains to establish the uniform convergence of $S^j_{3n,2}$. Note that $|h_3(X_{jk}, Y_k, Z_k; X_{jl}, Y_l, Z_k; X_{jh}, Y_h, Z_h)| \leq [\frac{1}{3}(Y_k^2 + Y_l^2 + Y_h^2)] \cdot K^2$, so the event $\{|S^j_{3n}| \geq \epsilon/2\}$ implies $\{Y_k^2 > M/K^2\}$ for some $1 \leq k \leq n$. Therefore, following a similar argument as in the part of S^j_{3n}, we have

$$\begin{aligned}P(|S^j_{3n,1} - S^j_{3,1}| \geq \epsilon) &\leq P(|S^j_{3n,2}| \geq \epsilon/2)\\ &\leq P(\cup_{k=1}^n [Y_k^2 \geq M/K^2])\\ &\leq Cn\exp(-sM/K^2),\end{aligned}$$

for any k and $s \in (0, 2s_0]$. Combining the two convergence results for $S^j_{3n,1}$ and $S^j_{3n,2}$ with $M = n^\gamma$ for some $0 < \gamma < 1/2 - \kappa$, it follows that

$$P(|S^j_{3n} - S^j_3| \geq 2\epsilon) \leq 2\exp(-\epsilon^2 n^{1-2\gamma}/6) + Cn\exp(-sn^\gamma/K^2)).$$

Combining all the consistency results, we have

$$\begin{aligned}&P\{|(2S^j_{3n} - S^j_{1n} - S^j_{2n}) - (2S^j_3 - S^j_1 - S^j_2)| \geq \epsilon\}\\ &\leq P(|S^j_{3n} - S^j_3| \geq \frac{\epsilon}{4}) + P(|S^j_{2n} - S^j_2| \geq \frac{\epsilon}{4}) + P(|S^j_{1n} - S^j_1| \geq \frac{\epsilon}{4})\\ &= O\{\exp(-c_1\epsilon^2 n^{1-2\gamma}) + n\exp(-c_2 n^\gamma)\}\end{aligned}$$

for some positive constants c_1 and c_2 and the bound is uniform with respect to $j = 1, ..., p$. Analyzing the denominator of $\widehat{\omega}_j$ would have the same form of convergence rate, so we omit the details here. Let $\epsilon = cn^{-\kappa}$, where κ satisfies $0 < \kappa + \gamma < 1/2$, we then have

$$P\{\max_{1\leq j\leq p-d_1}|\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa}\} \leq (p-d_1)\max_{1\leq j\leq p-d_1} P\{|\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa}\}$$
$$\leq O((p-d_1)[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]).$$

If $\mathcal{D}_S \nsubseteq \widehat{\mathcal{D}}_S$, then there exists some $j \in \mathcal{D}_S$ such that $\widehat{\omega}_j < cn^{-\kappa}$. According to Assumption (A2), for this particular j , we have $|\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa}$, which implies that $A = \{\mathcal{D}_S \nsubseteq \widehat{\mathcal{D}}_S\} \subseteq \{|\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa}, \text{for some } j \in \mathcal{D}_S\} = B$ and hence $B^c \subseteq A^c$. Finally,

$$P(A^c) \geq P(B^c) = 1 - P(B) = 1 - P(|\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa}, \text{for some } j \in \mathcal{D}_S)$$
$$\geq 1 - s_n \max_{(j\in\mathcal{D}_S)} P(|\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa})$$
$$\geq 1 - O(s_n[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^{\gamma})]).$$

where the first inequality above is due to *Bonferroni's inequality.*

Ranking Consistence. Recall the assumption that $\delta = \min_{j\in\mathcal{D}_S}\omega_j - \max_{j\in\mathcal{D}_S}\omega_j$,

$$P\left\{\min_{j\in\mathcal{D}_S}\widehat{\omega}_j \leq \max_{j\in\mathcal{I}_S}\widehat{\omega}_j\right\}$$
$$= P\left\{\min_{j\in\mathcal{D}_S}\widehat{\omega}_j - \min_{j\in\mathcal{D}_S}\omega_j + \delta \leq \max_{j\in\mathcal{I}_S}\widehat{\omega}_j - \max_{j\in\mathcal{I}_S}\omega_j\right\}$$
$$\leq P\left\{\max_{j\in\mathcal{D}_S}|\widehat{\omega}_j - \omega_j| \geq \delta/2\right\} + P\left\{\max_{j\in\mathcal{I}_S}|\widehat{\omega}_j - \omega_j| \geq \delta/2\right\}.$$

Hence,

$$P\left\{\min_{j\in\mathcal{I}_S}\widehat{\omega}_j < \max_{j\in\mathcal{D}_S}\widehat{\omega}_j\right\}$$
$$\geq 1 - 2O((p-d_1)[\exp(-c_1'\delta^2 n^{1-2\gamma}) + n\exp(-c_2 n^{\gamma})]).$$

□

S7 Proof of Proposition 1

Following the proof of Proposition 2 in Shao and Zhang (2014), we rewrite $|y_{l_\tau} - \hat{y}_{l_\tau}|$ as $|I(y_l \leq q_\tau) - I(y_l \leq \widehat{q}_\tau)|$, which is $I(\widehat{q}_\tau < y_l \leq q_\tau) + I(q_\tau < y_l \leq \widehat{q}_\tau)$. Then $P(\frac{1}{n}\sum_{l=1}^n |\hat{y}_{l_\tau} - y_{l_\tau}| > \epsilon) \leq P(\frac{1}{n}\sum_{l=1}^n |\hat{y}_{l_\tau} - y_{l_\tau}| > \epsilon, |\widehat{q}_\tau - q_\tau| \leq \delta) + P(|\widehat{q}_\tau - q_\tau| \geq \delta| =: P_1 + P_2$. For P_2, we apply Serfling (2009) Theorem 2.3.2 and get $P_2 = P(|\widehat{q}_\tau - q_\tau| \geq \delta) \leq 2\exp(-2nL(\delta)^2)$, where $L(\delta) = \min\{F_Y(q_\tau+\delta) - \tau, \tau - F_Y(q_\tau+\delta)\}$. Under the Assumption (B1), we have $G_1(\delta_0)\delta \leq L(\delta) \leq G_2(\delta_0)\delta$. Let $\delta = \min(\epsilon/\{4G_2(\delta_0)\}, \delta_0)$ if $\epsilon < \epsilon_0 = 4G_2(\delta_0)\delta_0$. Then for $\epsilon \in (0, \epsilon_0)$, we have

$$P_2 \leq 2\exp(-2nG_1(\delta_0)^2\delta^2) \leq 2\exp(-2n\frac{G_1(\delta_0)^2}{16G_2(\delta_0)^2}\epsilon^2).$$

Setting $P_{q_\tau} = P(|y_l - q_\tau| \leq \delta)$, we can find a bound for P_1:

$$P_1 \leq P(\frac{1}{n}\sum_{l=1}^n I(|y_l - q_\tau| \leq \delta) > \epsilon) = P(\frac{1}{n}\sum_{l=1}^n I(|y_l - q_\tau| \leq \delta) - P_{q_\tau} > \epsilon - P_{q_\tau}).$$

By Hoeffding's inequality, $P_1 \leq \exp(-2(\epsilon - P_{q_\tau})^2 n)$. Since $P_{q_\tau} \leq 2\delta G_2(\delta_0) \leq \epsilon/2$ when $\epsilon \in (0, \epsilon_0)$, then $I_1 \leq \exp(-2n\epsilon^2/4)$. Together with the bound for P_2, we have

$P\left(\frac{1}{n}\sum_{l=1}^{n}|\widehat{y_{l_\tau}} - y_{l_\tau}| > \epsilon\right) \leq 3\exp(-2nc_1\epsilon^2).$ □

S8 Proof of Theorem 5

We shall show the uniform consistency of $\widehat{\omega}_j(\hat{Y}_\tau) = \text{CMC}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$ under Assumptions (B1) and (B2). Due to the similarity of its numerator and denominator, we only present the numerator part, which is the consistency of $\widehat{\text{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S))$. First we show the consistency of $\widehat{\text{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)$. Then we show that $\widehat{\text{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$ is a consistent estimator of $\widehat{\text{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)$. Following the similar procedures in the proof of Section S6, for any $\gamma \in (0, 1/2-\kappa)$, there exist positive constants c_1 and c_2 such that

$$P(|\text{CMD}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S) - \widehat{\text{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)| \geq \epsilon) \leq C[\exp\{-c_1\epsilon^2 n^{1-2\gamma}\} + n\exp(-c_2 n^\gamma)] \tag{S8.10}$$

for a sufficiently small ϵ (i.e. $\epsilon = cn^{-\kappa}$, which will be defined later). Next we focus on the difference between $\widehat{\text{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)$ and $\widehat{\text{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$. Denote $\widehat{T}^j_{1n} = (n)_2^{-1}\sum_{(k,l)\in I_2^n} \hat{y}_{k_\tau}\hat{y}_{l_\tau}k_1(Z_k, Z_l)k_2(X_{jk}, X_{jl})$, $\widehat{T}^j_{2n} = (n)_4^{-1}\sum_{(k,l,h,q)\in I_4^n} \hat{y}_{k_\tau}\hat{y}_{l_\tau}k_1(Z_k, Z_l)k_2(X_{jh}, X_{jq})$, and $\widehat{T}^j_{3n} = (n)_3^{-1}\sum_{(k,l,h)\in I_3^n} \hat{y}_{k_\tau}\hat{y}_{h_\tau}k_1(Z_k, Z_h)k_2(X_{jk}, X_{jl})$. Similarly, T^j_{1n}, T^j_{2n} and T^j_{3n} are defined as $\{\hat{y}_{k_\tau}\}_{k=1}^n$ replaced with $\{W_k\}_{k=1}^n$. Let $C_0 = \tau + 1$. By using the *triangle* *inequality* and the boundedness of y_{k_τ} and $\hat{y}_{k_\tau}$, we can derive that

$$
\begin{aligned}
&|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S) - \widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)| \leq |\widehat{T}^j_{1n} - T^j_{1n}| + |\widehat{T}^j_{2n} - T^j_{2n}| + 2|\widehat{T}^j_{3n} - T^j_{3n}| \\
&= |(n)_2^{-1} \sum_{(k,l)\in I_2^n} [\hat{y}_{k_\tau}\hat{y}_{l_\tau} - y_{k_\tau}y_{l_\tau}] k_1(Z_k, Z_l) k_2(X_{jk}, X_{jl})| + \\
&|(n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n}^{n} [\hat{y}_{k_\tau}\hat{y}_{l_\tau} - y_{k_\tau}y_{l_\tau}] k_1(Z_k, Z_l) k_2(X_{jh}, X_{jq})| + \\
&2|(n)_3^{-1} \sum_{(k,l,h)\in I_3^n} [\hat{y}_{k_\tau}\hat{y}_{h_\tau} - y_{k_\tau}y_{h_\tau}] k_1(Z_k, Z_h) k_2(X_{jk}, X_{jl})| \\
&\leq (n)_2^{-1} \sum_{(k,l)\in I_2^n} [|\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k, Z_l) k_2(X_{jk}, X_{jl}) + \\
&(n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n} [|\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k, Z_l) k_2(X_{jh}, X_{jq}) + \\
&2(n)_3^{-1} \sum_{(k,l,h)\in I_3^n} [|\hat{y}_{k_\tau}(\hat{y}_{h_\tau} - y_{h_\tau})| + |y_{h_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k, Z_l) k_2(X_{jk}, X_{jl}).
\end{aligned}
$$

For the first part:

$$
\begin{aligned}
&(n)_2^{-1} \sum_{(k,l)\in I_2^n} [|\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k, Z_l) k_2(X_{jk}, X_{jl}) \\
&\leq K\{(n)_2^{-1} \sum_{(k,l)\in I_2^n} |\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + (n)_2^{-1} \sum_{(k,l)\in I_2^n} |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|\} \\
&\leq K\{C_0(n)_2^{-1} \sum_{(k,l)\in I_2^n} |(\hat{y}_{l_\tau} - y_{l_\tau})| + C_0(n)_2^{-1} \sum_{(k,l)\in I_2^n} |(\hat{y}_{k_\tau} - y_{k_\tau})|\} \\
&= 2KC_0(n)_2^{-1} \sum_{(k,l)\in I_2^n} |(\hat{y}_{l_\tau} - y_{l_\tau})|.
\end{aligned}
$$

For the second part:

$$
\begin{aligned}
&(n)_4^{-1} \sum_{(k,l,h,q)\in I_4^n} [|\hat{y}_{k_\tau}(\hat{y}_{l_\tau} - y_{l_\tau})| + |y_{l_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k, Z_l) k_2(X_{jh}, X_{jq}) \\
&\leq 2KC_0 (n)_4^{-1} \sum_{(k,h,q)\in I_3^n} |\hat{y}_{k_\tau} - y_{k_\tau}|.
\end{aligned}
$$

For the third part:

$$
\begin{aligned}
&2(n)_3^{-1} \sum_{(k,l,h)\in I_3^n} [|\hat{y}_{k_\tau}(\hat{y}_{h_\tau} - y_{h_\tau})| + |y_{h_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|] k_1(Z_k, Z_l) k_2(X_{jk}, X_{jl}) \\
&\leq 2K\{(n)_3^{-1} \sum_{(k,l,h)\in I_3^n} [|\hat{y}_{k_\tau}(\hat{y}_{h_\tau} - y_{h_\tau})| + 2(n)_3^{-1} \sum_{(k,l,h)\in I_3^n} [|\hat{y}_{h_\tau}(\hat{y}_{k_\tau} - y_{k_\tau})|\} \\
&= 2KC_0 (n)_3^{-1} \Big(\sum_{(k,h,l)\in I_3^n} |\hat{y}_{h_\tau} - y_{h_\tau}| + \sum_{(k,l)\in I_2^n}^{n} |(\hat{y}_{k_\tau} - y_{k_\tau})| \Big).
\end{aligned}
$$

Combining the above three parts together, we get

$$
\begin{aligned}
&|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S) - \widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)| \\
&\leq 4KC_0 (n)_2^{-1} \sum_{(k,l)\in I_2^n} |(\hat{y}_{l_\tau} - y_{l_\tau})| + 4KC_0 (n)_3^{-1} \sum_{(k,h,l)\in I_3^n}^{n} |\hat{y}_{h_\tau} - y_{h_\tau}| \\
&= \frac{8KC_0}{n} \sum_{l=1}^{n} |(\hat{y}_{l_\tau} - y_{l_\tau})|.
\end{aligned}
$$

By Proposition 1, we have:

$$
P\Big(\frac{8C_0}{n} \sum_{k=1}^{n} |(\hat{y}_{l_\tau} - y_{l_\tau})| * Z \geq \epsilon\Big) = P\Big(\frac{1}{n} \sum_{k=1}^{n} |(\hat{y}_{l_\tau} - y_{l_\tau})| \geq \frac{\epsilon}{8KC_0}\Big) \leq 3\exp(-2nc_1\epsilon^2).
$$

Consequently, in view of (S8.10), we have that

$$\begin{aligned}
&P(|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S) - \mathrm{CMD}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)| \geq 2\epsilon) \\
\leq &P(|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S) - \mathrm{CMD}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)| \geq \epsilon)+ \\
&P(|\widehat{\mathrm{CMD}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S) - \widehat{\mathrm{CMD}}^2_{\mathcal{H}}(Y_\tau, X_j|\boldsymbol{X}_S)| \geq \epsilon) \\
\leq &C[\exp\{-c_1\epsilon^2 n^{1-2\gamma}\} + n\exp(-c_2 n^\gamma)],
\end{aligned}$$

for a sufficiently small $\epsilon > 0$ and some positive constant c_1 and c_2. The analysis of the denominator of $\widehat{\mathrm{CMC}}^2_{\mathcal{H}}(\hat{Y}_\tau, X_j|\boldsymbol{X}_S)$ will generate a similar form of the convergence rate. Therefore, if we set $\epsilon = cn^{-\kappa}$, where κ satisfies $0 < \kappa + \gamma < 1/2$, we have

$$\begin{aligned}
P\{\max_{1\leq j\leq p-d_1} |\widehat{\omega}_j(\hat{Y}_\tau) - \omega_j(\hat{Y}_\tau)| \geq cn^{-\kappa}\} &\leq p \max_{1\leq j\leq p-d_1} P\{|\widehat{\omega}_j(\hat{Y}_\tau) - \omega_j(\hat{Y}_\tau)| \geq cn^{-\kappa}\} \\
&\leq O((p-d_1)[\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n\exp(-c_2 n^\gamma)]).
\end{aligned}$$

□

The proofs of Theorems 6-10 in the appendix follow exactly the similar lines of argument as in their corresponding $\mathrm{CMD}_{\mathcal{H}}$ or $\mathrm{CMC}_{\mathcal{H}}$ versions, we thus omit the details here.

Bibliography

Kuo, H.-H. (2006). Gaussian measures in banach spaces. *Gaussian Measures in Banach Spaces*, pages 1–109.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Resnick, S. (2019). *A probability path*. Springer.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.