Supplementary Materials for "K-Nearest-Neighbor Local Sampling Based Conditional Independence Testing"

A Theoretical results

A.1 Proof of Lemma 1

Proof. Recall that Z is a random vector taking values in Euclidean space $(\mathbb{R}^{d_z}, \|\cdot\|_2)$, where d_z is the dimension of Z and $\|\cdot\|_2$ is Euclidean distance. $Z_1, Z_2, \ldots, Z_n$ are i.i.d. random vectors according to $p(z)$. For a fixed $z \in \mathbb{R}^{d_z}$, we denote by $Z_n^{(1)}(z), \ldots, Z_n^{(n)}(z)$ a reordering of $Z_1, Z_2, \ldots, Z_n$ according to the increasing values of $\|Z_i - z\|_2$, that is,

$$\|Z_n^{(1)}(z) - z\|_2 \leq \ldots \leq \|Z_n^{(n)}(z) - z\|_2.$$

Define the set $G = \{z \in \mathbb{R}^{d_z} |\ \forall\delta > 0,\ P(\{\omega : Z(\omega) \in S_z(\delta)\}) > 0\}$, where $S_z(\delta) = \{x \in \mathbb{R}^{d_z} |\ \|x - z\|_2 \leq \delta\}$. For convenience, we omit ω in probability in the following paper. For example, write $P(Z \in S_z(\delta))$ instead of $P(\{\omega : Z(\omega) \in S_z(\delta)\})$. By definition, for $z \in G$, $\forall\delta > 0$, $P(Z \in S_z(\delta)) > 0$. Let G^c be the complement of G. Then, for $z \in G^c$, $\exists r_z > 0$, s.t. $\forall r < r_z$, $P(Z \in S_z(r)) = 0$. Note that $P(\|Z_n^{(k)}(z) - z\|_2 > \delta) = P(Z_n^{(k)}(z) \notin S_z(\delta))$.

In order to prove $\|Z_n^{(k)}(Z) - Z\|_2 \to 0$ a.s., it is sufficient to prove $\forall\delta > 0$, $\lim_{n\to\infty} P(\sup_{m\geq n}\|Z_m^{(k)}(Z) - Z\|_2 > \delta) = 0$. We can obtain

$$\begin{aligned}
P(\sup_{m\geq n}\|Z_m^{(k)}(Z) - Z\|_2 > \delta) &\leq P(\{\sup_{m\geq n}\|Z_m^{(k)}(Z) - Z\|_2 > \delta\} \cap \{Z \in G\}) + P(Z \in G^c) \\
&= \int_G P(\sup_{m\geq n}\|Z_m^{(k)}(z) - z\|_2 > \delta) p(z) dz + P(Z \in G^c) \\
&\leq \int_G \sum\nolimits_{m\geq n} P(Z_m^{(k)}(z) \notin S_z(\delta)) p(z) dz + P(Z \in G^c). \quad (11)
\end{aligned}$$

First, consider the first term of (11). We have

$$\begin{aligned}
P(Z_m^{(k)}(z) \notin S_z(\delta)) = &P(Z_m^{(1)}(z), Z_m^{(2)}(z), \ldots, Z_m^{(m)}(z) \notin S_z(\delta)) \\
&+ P(Z_m^{(1)}(z) \in S_z(\delta), Z_m^{(2)}(z), \ldots, Z_m^{(m)}(z) \notin S_z(\delta)) \\
&+ P(Z_m^{(1)}(z), Z_m^{(2)}(z) \in S_z(\delta), Z_m^{(3)}(z), \ldots, Z_m^{(m)}(z) \notin S_z(\delta)) \\
&+ \ldots + P(Z_m^{(1)}(z), \ldots, Z_m^{(k-1)}(z) \in S_z(\delta), Z_m^{(k)}(z), \ldots, Z_m^{(m)}(z) \notin S_z(\delta)).
\end{aligned}$$

By setting $P(Z \in S_z(\delta)) = \gamma$, we have

$$\begin{aligned}
P(Z_m^{(k)}(z) \notin S_z(\delta)) = &(1-\gamma)^m + C_m^1 \gamma (1-\gamma)^{m-1} + C_m^2 \gamma^2 (1-\gamma)^{m-2} + \ldots \\
&+ C_m^{k-1} \gamma^{k-1} (1-\gamma)^{m-k+1}. \quad (12)
\end{aligned}$$

Consider the j-th term of (12). Let $C_1 := \gamma^j / j!$, $C_2 := C_1 e^j$ and $C_3 := C_2 e^j (1-\gamma)^{-j}$. By using Stirling's approximation, we have

$$\begin{aligned}
\lim_{m\to\infty} C_m^j \gamma^j (1-\gamma)^{m-j} &= \lim_{m\to\infty} \frac{m!}{(m-j)!j!} \gamma^j (1-\gamma)^{m-j} \\
&= \lim_{m\to\infty} C_1 \frac{\sqrt{2\pi m}(\frac{m}{e})^m}{\sqrt{2\pi(m-j)}(\frac{m-j}{e})^{m-j}} (1-\gamma)^{m-j} \\
&= \lim_{m\to\infty} C_2 \sqrt{\frac{m}{(m-j)}} \frac{m^m}{(m-j)^{m-j}} (1-\gamma)^{m-j} \\
&= \lim_{m\to\infty} C_2 e^j (m+j)^j (1-\gamma)^m \\
&= \lim_{m\to\infty} C_2 e^j m^j (1-\gamma)^{m-j} \\
&= \lim_{m\to\infty} C_3 m^j (1-\gamma)^m.
\end{aligned}$$

Thus, there exists $C_4 > 0$ such that $P(Z_m^{(k)}(z) \notin S_z(\delta)) \leq C_4 m^{k-1}(1-\gamma)^m$ when m is large enough. It holds that

$$\frac{C_4 m^{k-1}(1-\gamma)^m}{m^{-2}} = C_4 m^{k+1}(1-\gamma)^m \to 0, \quad \text{as} \quad m \to +\infty.$$

Thus, for $z \in G$ and n large enough, $\forall \delta > 0$, we have $P(Z_m^{(k)}(z) \notin S_z(\delta)) = o(m^{-2})$ and

$$\sum\nolimits_{m \geq n} P(Z_m^{(k)}(z) \notin S_z(\delta)) \leq \sum\nolimits_{m \geq n} \frac{1}{m^2},$$

which shows $\lim_{n\to\infty} \sum_{m \geq n} P(Z_m^{(k)}(z) \notin S_z(\delta)) = 0$ for $z \in G$. So by Lebesgue dominated convergence theorem, we obtain $\lim_{n\to\infty} \int_G \sum_{m \geq n} P(Z_m^{(k)}(z) \notin S_z(\delta)) p(z) dz = 0$.

Second, we consider the second term of (11). To prove that $P(Z \in G^c) = 0$, we aim to construct a countable open cover of G^c and show that the probability of the random vector Z falling into each of these open balls is zero. By the property of G^c, for every $z \in G^c$, there exists $r_z > 0$ such that for all $r < r_z$, $P(Z \in S_z(r)) = 0$. Furthermore, using the separability of Euclidean space and the density of the rational number set, we can approximate z using points from $\mathbb{Q}^{d_z}$ with $\mathbb{Q}$ being the rational number set. Therefore, for every $z \in G^c$, there exist $x \in \mathbb{Q}^{d_z} \cap S_z(\frac{r_z}{3})$ and $r \in \mathbb{Q} \cap (\frac{r_z}{3}, \frac{2r_z}{3})$, such that $z \in S_x(r) \subseteq S_z(r_z)$. Because $P(Z \in S_z(r_z)) = 0$, we conclude that $P(Z \in S_x(r)) = 0$. Define

$$\mathcal{F} := \{S_x(r) | \ \exists z \in G^c, \text{ such that } z \in S_x(r) \subseteq S_z(r_z) \text{ with } x \in \mathbb{Q}^{d_z} \text{ and } r \in \mathbb{Q}\}.$$

Note that the elements in set $\mathcal{F}$ are mutually distinct. By the construction of $S_x(r)$, $\mathcal{F}$ forms a countable open cover of G^c, and the probability of Z falling into each open ball in $\mathcal{F}$ is zero. Using the monotonicity and countable additivity properties of probability, we have $P(Z \in G^c) \leq P(Z \in \cup_{S_x(r) \in \mathcal{F}} S_x(r)) \leq \sum_{S_x(r) \in \mathcal{F}} P(Z \in S_x(r)) = 0$. Thus, we conclude that $P(Z \in G^c) = 0$.

We therefore conclude that, $\forall \delta > 0$, $\lim_{n\to\infty} P(\sup_{m \geq n} \|Z_m^{(k)}(Z) - Z\|_2 > \delta) = 0$. This finish the proof.

A.2 Proof of Theorem 2

Proof. By Pinsker's inequality, we have

$$d_{TV}\{p(x|Z), \widehat{p}(x|Z)\} \leq \sqrt{D_{KL}\{p(x|Z), \widehat{p}(x|Z)\}/2}.$$

Note that $I\{\xi = 1\} + \ldots + I\{\xi = k\} = 1$. By the definition of $\widehat{p}(x|Z)$, we obtain

$$\begin{aligned} D_{KL}\{p(x|Z), \widehat{p}(x|Z)\} &= \int p(x|Z) \log \left\{ \frac{p(x|Z)}{p(x|Z_n^{(1)})^{I\{\xi=1\}} \times \ldots \times p(x|Z_n^{(k)})^{I\{\xi=k\}}} \right\} dx \\ &= \int p(x|Z) \log \prod_{l=1}^{k} \frac{p(x|Z)^{I\{\xi=l\}}}{p(x|Z_n^{(l)})^{I\{\xi=l\}}} dx \\ &= \sum_{l=1}^{k} I\{\xi = l\} \int p(x|Z) \log \frac{p(x|Z)}{p(x|Z_n^{(l)})} dx \\ &= \sum_{l=1}^{k} I\{\xi = l\} D_{KL}\{p(x|Z) || p(x|Z_n^{(l)})\}. \end{aligned}$$

Then, by Taylor's expansion, we have

$$\begin{aligned} & D_{KL}\{p(x|Z) || p(x|Z_n^{(l)})\} \\ &= D_{KL}\{p(x|Z) || p(x|Z)\} + \frac{\partial}{\partial z'} D_{KL}\{p(x|Z) || p(x|z')\} \Big|_{z'=Z} (Z_n^{(l)} - Z) \\ &\quad + \frac{1}{2} (Z_n^{(l)} - Z)^T \frac{\partial^2}{\partial z' \partial z'^T} D_{KL}\{p(x|Z) || p(x|z')\} \Big|_{z'=a} (Z_n^{(l)} - Z), \end{aligned}$$

where $a = \lambda Z + (1-\lambda) Z_n^{(l)}$ with $0 \le \lambda \le 1$.

Note that $D_{KL}\{p(x|Z)||p(x|Z)\} = \int p(x|Z) \log \frac{p(x|Z)}{p(x|Z)} dx = 0$. By Lemma 1 and Assumptions 1 and 2, we have

$$\frac{\partial}{\partial z'} D_{KL}\{p(x|Z)||p(x|z')\}\bigg|_{z'=Z} = -\int p(x|Z) \cdot \frac{\partial}{\partial z'} \log p(x|z')\bigg|_{z'=Z} dx$$
$$= -\frac{\partial}{\partial z'} \int p(x|z') dx \bigg|_{z'=Z} = 0$$

and

$$\frac{\partial^2}{\partial z' \partial z'^T} D_{KL}\{p(x|Z)||p(x|z')\}\bigg|_{z'=a} = -\int p(x|Z) \cdot \frac{\partial^2}{\partial z' \partial z'^T} \log p(x|z')\bigg|_{z'=a} dx = I_a(Z).$$

This means

$$D_{KL}\{p(x|Z)||p(x|Z_n^{(l)})\} = \frac{1}{2}(Z_n^{(l)} - Z)^T I_a(Z)(Z_n^{(l)} - Z).$$

Note that $Z_n^{(l)} \to Z$ a.s. implies that $Z_n^{(l)}$ converges to Z in probability. Then $\forall \delta > 0$, for ϵ defined in Assumption 1, we have

$$\begin{aligned} P(D_{KL}\{p(x|Z)||p(x|Z_n^{(l)})\} > \delta) &\le P(\{D_{KL}\{p(x|Z)||p(x|Z_n^{(l)})\} > \delta\} \cap \{\|Z_n^{(l)} - Z\|_2 \le \epsilon\}) \\ &\quad + P(\|Z_n^{(l)} - Z\|_2 > \epsilon) \\ &\le P\left(\frac{1}{2}\beta \|Z_n^{(l)} - Z\|_2^2 > \delta\right) + P(\|Z_n^{(l)} - Z\|_2 > \epsilon) \\ &= o(1). \end{aligned}$$

Thus, $D_{KL}\{p(x|Z)||p(x|Z_n^{(l)})\} = o_p(1)$.

Because $I\{\xi = l\} \le 1$ for $l = 1, 2, ..., k$, and k is finite, we obtain

$$D_{KL}\{p(x|Z), \widehat{p}(x|Z)\} = \sum_{l=1}^{k} I\{\xi = l\} D_{KL}\{p(x|Z)||p(x|Z_n^{(l)})\} = o_p(1).$$

Finally, we conclude that $d_{TV}\{p(x|Z), \widehat{p}(x|Z)\} \le \sqrt{D_{KL}\{p(x|Z), \widehat{p}(x|Z)\}/2} = o_p(1)$.

A.3 Proof of Theorem 3

To prove Theorem 3, the following two lemmas are needed.

Lemma 5. *Let $\acute{\boldsymbol{X}}$ be drawn from $\widehat{p}(\cdot|\boldsymbol{Z})$, independently of $\boldsymbol{Y}$. $\acute{\boldsymbol{X}}^{(1)}, \ldots, \acute{\boldsymbol{X}}^{(B)}$ are i.i.d. samples drawn from the k-nearest-neighbor local sampling mechanism based on $(\acute{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{Z})$. For any statistic T, the $B+1$ statistics*

$$T(\acute{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{Z}), T(\acute{\boldsymbol{X}}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z}), \ldots, T(\acute{\boldsymbol{X}}^{(B)}, \boldsymbol{Y}, \boldsymbol{Z})$$

are exchangeable conditionally on $\boldsymbol{Y}$ and $\boldsymbol{Z}$.

Proof. We have that the $B+1$ triples $(\acute{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{Z}), (\acute{\boldsymbol{X}}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z}), \ldots, (\acute{\boldsymbol{X}}^{(B)}, \boldsymbol{Y}, \boldsymbol{Z})$ are i.i.d. samples drawn from the same mechanism after conditionally on $\acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}$ and $\boldsymbol{Z}$, where $\acute{\boldsymbol{X}}_{()}$ is the order statistic of $\acute{\boldsymbol{X}}$. Note that, T is measurable. Thus, $T(\acute{\boldsymbol{X}}, \boldsymbol{Y}, \boldsymbol{Z}), T(\acute{\boldsymbol{X}}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z}), \ldots, T(\acute{\boldsymbol{X}}^{(B)}, \boldsymbol{Y}, \boldsymbol{Z})$ are i.i.d. after conditionally on $\acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}$ and $\boldsymbol{Z}$. Conditionally on $\acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}$ and $\boldsymbol{Z}$, denote their cumulative conditional distribution as $F(\cdot|\acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}, \boldsymbol{Z})$. Denote $\acute{\boldsymbol{X}}$ as $\acute{\boldsymbol{X}}^{(0)}$. Then, for any $t_0, \ldots, t_B \in \mathbb{R}$ and any permutation $\pi = (\pi_{(0)}, \ldots, \pi_{(B)})$ of the indices $\{0, 1, \ldots, B\}$, we have

$$
\begin{aligned}
&P(T(\acute{\boldsymbol{X}}^{(0)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_0, T(\acute{\boldsymbol{X}}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_1, \ldots, T(\acute{\boldsymbol{X}}^{(B)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_B | \boldsymbol{Y}, \boldsymbol{Z}) \\
&= E_{\acute{\boldsymbol{X}}_{()}|\boldsymbol{Y},\boldsymbol{Z}} \{P(T(\acute{\boldsymbol{X}}^{(0)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_0, T(\acute{\boldsymbol{X}}^{(1)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_1, \ldots, T(\acute{\boldsymbol{X}}^{(B)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_B | \acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}, \boldsymbol{Z})\} \\
&= E_{\acute{\boldsymbol{X}}_{()}|\boldsymbol{Y},\boldsymbol{Z}} \{P(T(\acute{\boldsymbol{X}}^{(0)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_0 | \acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}, \boldsymbol{Z}), \ldots, P(T(\acute{\boldsymbol{X}}^{(B)}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_B | \acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}, \boldsymbol{Z})\} \\
&= E_{\acute{\boldsymbol{X}}_{()}|\boldsymbol{Y},\boldsymbol{Z}} \left\{ \prod_{i=0}^{B} F(t_i | \acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}, \boldsymbol{Z}) \right\} \\
&= E_{\acute{\boldsymbol{X}}_{()}|\boldsymbol{Y},\boldsymbol{Z}} \{P(T(\acute{\boldsymbol{X}}^{(\pi_{(0)})}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_0, \ldots, T(\acute{\boldsymbol{X}}^{(\pi_{(B)})}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_B | \acute{\boldsymbol{X}}_{()}, \boldsymbol{Y}, \boldsymbol{Z})\} \\
&= P(T(\acute{\boldsymbol{X}}^{(\pi_{(0)})}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_0, \ldots, T(\acute{\boldsymbol{X}}^{(\pi_{(B)})}, \boldsymbol{Y}, \boldsymbol{Z}) \leq t_B | \boldsymbol{Y}, \boldsymbol{Z}).
\end{aligned}
$$

Thus, the desired result follows.

Let $\stackrel{d}{=}$ denotes equality in distribution. We present the following Lemma:

Lemma 6. *For any two bi-tuples $(\boldsymbol{U}, \boldsymbol{V})$ and $(\boldsymbol{U}', \boldsymbol{V}')$, if $\forall \boldsymbol{u},\ (\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \stackrel{\mathrm{d}}{=} (\boldsymbol{V}'|\boldsymbol{U}' = \boldsymbol{u})$, we have $d_{TV}\{(\boldsymbol{U}, \boldsymbol{V}), (\boldsymbol{U}', \boldsymbol{V}')\} = d_{TV}(\boldsymbol{U}, \boldsymbol{U}')$.*

Proof. Denote the joint density functions of $(\boldsymbol{U}, \boldsymbol{V})$ and $(\boldsymbol{U}', \boldsymbol{V}')$ by $p_{\boldsymbol{U},\boldsymbol{V}}(\boldsymbol{u}, \boldsymbol{v})$ and $p_{\boldsymbol{U}',\boldsymbol{V}'}(\boldsymbol{u}', \boldsymbol{v}')$, respectively. According to the equivalent definition of the TV distance, we obtain

$$
\begin{aligned}
d_{TV}\{(\boldsymbol{U}, \boldsymbol{V}), (\boldsymbol{U}', \boldsymbol{V}')\} &= \frac{1}{2} \iint \left| p_{\boldsymbol{U},\boldsymbol{V}}(\boldsymbol{u}, \boldsymbol{v}) - p_{\boldsymbol{U}',\boldsymbol{V}'}(\boldsymbol{u}, \boldsymbol{v}) \right| d\boldsymbol{u} d\boldsymbol{v} \\
&= \frac{1}{2} \iint \left| p_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{u}) p_{\boldsymbol{U}}(\boldsymbol{u}) - p_{\boldsymbol{V}'|\boldsymbol{U}'}(\boldsymbol{v}|\boldsymbol{u}) p_{\boldsymbol{U}'}(\boldsymbol{u}) \right| d\boldsymbol{u} d\boldsymbol{v} \\
&= \frac{1}{2} \iint p_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{u}) \left| p_{\boldsymbol{U}}(\boldsymbol{u}) - p_{\boldsymbol{U}'}(\boldsymbol{u}) \right| d\boldsymbol{u} d\boldsymbol{v} \\
&= \frac{1}{2} \int \left[\int p_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{u}) d\boldsymbol{v} \right] \left| p_{\boldsymbol{U}}(\boldsymbol{u}) - p_{\boldsymbol{U}'}(\boldsymbol{u}) \right| d\boldsymbol{u} \\
&= \frac{1}{2} \int \left| p_{\boldsymbol{U}}(\boldsymbol{u}) - p_{\boldsymbol{U}'}(\boldsymbol{u}) \right| d\boldsymbol{u} \\
&= d_{TV}(\boldsymbol{U}, \boldsymbol{U}').
\end{aligned}
$$

Now we present the proof of Theorem 3:

Proof. Let $\widetilde{\boldsymbol{X}}^{(1)}, \ldots, \widetilde{\boldsymbol{X}}^{(B)}$ be i.i.d. drawn from the k-nearest-neighbor local sampling mechanism, see Algorithm 3. Now let $\acute{\boldsymbol{X}}$ be an additional sample drawn from $\widehat{p}(\cdot|\boldsymbol{Z})$ independently of $\boldsymbol{Y}$. Let $\acute{\boldsymbol{X}}^{(1)}, \ldots, \acute{\boldsymbol{X}}^{(B)}$ be i.i.d. drawn from the k-nearest-neighbor local sampling mechanism after we observe $\acute{\boldsymbol{X}}$ instead of $\boldsymbol{X}$. Because $(\acute{\boldsymbol{X}}^{(1)}, \ldots, \acute{\boldsymbol{X}}^{(B)})$, conditionally on $\acute{\boldsymbol{X}}, \boldsymbol{Y}$ and $\boldsymbol{Z}$, is generated from the same mechanism as $(\widetilde{\boldsymbol{X}}^{(1)}, \ldots, \widetilde{\boldsymbol{X}}^{(B)})$, conditionally on $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$, for all $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$
((\widetilde{\boldsymbol{X}}^{(1)}, \ldots, \widetilde{\boldsymbol{X}}^{(B)}) | \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{Z}) \stackrel{\mathrm{d}}{=} ((\acute{\boldsymbol{X}}^{(1)}, \ldots, \acute{\boldsymbol{X}}^{(B)}) | \acute{\boldsymbol{X}} = \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{Z}).
$$

Then, by applying Lemma 6, we obtain

$$
\begin{aligned}
&d_{TV}\{(\boldsymbol{X}, \widetilde{\boldsymbol{X}}^{(1)}, \ldots, \widetilde{\boldsymbol{X}}^{(B)} | \boldsymbol{Y}, \boldsymbol{Z}), (\acute{\boldsymbol{X}}, \acute{\boldsymbol{X}}^{(1)}, \ldots, \acute{\boldsymbol{X}}^{(B)} | \boldsymbol{Y}, \boldsymbol{Z})\} \\
&= d_{TV}\{(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}), (\acute{\boldsymbol{X}}|\boldsymbol{Y}, \boldsymbol{Z})\} = d_{TV}\{p(\cdot|\boldsymbol{Z}), \widehat{p}(\cdot|\boldsymbol{Z})\}.
\end{aligned}
$$

Define $\chi_\alpha^B := \left\{ (\boldsymbol{x}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(B)}) \middle| \ \left[1 + \sum_{b=1}^{B} 1\{T(\boldsymbol{x}^{(b)}, \boldsymbol{Y}, \boldsymbol{Z}) \geq T(\boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{Z})\} \right] / (1 + B) \leq \alpha \right\}$, where $1(\cdot)$ is the indicator function. Note that in our case, the statistic T is selected to be $\widehat{\mathrm{CMI}}$. Then, it follows that

$$\begin{aligned}
P(p \leq \alpha|\boldsymbol{Y},\boldsymbol{Z}) &= P((\boldsymbol{X},\widetilde{\boldsymbol{X}}^{(1)},\ldots,\widetilde{\boldsymbol{X}}^{(B)}) \in \chi_\alpha^B|\boldsymbol{Y},\boldsymbol{Z}) \\
&= P((\acute{\boldsymbol{X}},\acute{\boldsymbol{X}}^{(1)},\ldots,\acute{\boldsymbol{X}}^{(B)}) \in \chi_\alpha^B|\boldsymbol{Y},\boldsymbol{Z}) + P((\boldsymbol{X},\widetilde{\boldsymbol{X}}^{(1)},\ldots,\widetilde{\boldsymbol{X}}^{(B)}) \in \chi_\alpha^B|\boldsymbol{Y},\boldsymbol{Z}) \\
&\quad - P((\acute{\boldsymbol{X}},\acute{\boldsymbol{X}}^{(1)},\ldots,\acute{\boldsymbol{X}}^{(B)}) \in \chi_\alpha^B|\boldsymbol{Y},\boldsymbol{Z}) \\
&\leq P((\acute{\boldsymbol{X}},\acute{\boldsymbol{X}}^{(1)},\ldots,\acute{\boldsymbol{X}}^{(B)}) \in \chi_\alpha^B|\boldsymbol{Y},\boldsymbol{Z}) \\
&\quad + d_{TV}\{(\boldsymbol{X},\widetilde{\boldsymbol{X}}^{(1)},\ldots,\widetilde{\boldsymbol{X}}^{(B)}|\boldsymbol{Y},\boldsymbol{Z}),(\acute{\boldsymbol{X}},\acute{\boldsymbol{X}}^{(1)},\ldots,\acute{\boldsymbol{X}}^{(B)}|\boldsymbol{Y},\boldsymbol{Z})\} \\
&= P((\acute{\boldsymbol{X}},\acute{\boldsymbol{X}}^{(1)},\ldots,\acute{\boldsymbol{X}}^{(B)}) \in \chi_\alpha^B|\boldsymbol{Y},\boldsymbol{Z}) + d_{TV}\{p(\cdot|\boldsymbol{Z}),\widehat{p}(\cdot|\boldsymbol{Z})\}.
\end{aligned}$$

Applying Lemma 5 and the property of rank test, we obtain $P((\acute{\boldsymbol{X}},\acute{\boldsymbol{X}}^{(1)},\ldots,\acute{\boldsymbol{X}}^{(B)}) \in \chi_\alpha^B|\boldsymbol{Y},\boldsymbol{Z}) \leq \alpha$. Finally, we have $P(p \leq \alpha|\boldsymbol{Y},\boldsymbol{Z}) \leq \alpha + d_{TV}\{p(\cdot|\boldsymbol{Z}),\widehat{p}(\cdot|\boldsymbol{Z})\}$.

Because the TV distance is bounded by 1, marginalizing the above inequality over $\boldsymbol{Y}$ and $\boldsymbol{Z}$ and applying Theorem 2 and Lebesgue dominated convergence theorem lead to

$$P(p \leq \alpha|H_0) \leq \alpha + E(d_{TV}\{p(\cdot|\boldsymbol{Z}),\widehat{p}(\cdot|\boldsymbol{Z})\}) = \alpha + o(1).$$

A.4 Proof of Theorem 4

Here we present the assumptions given in [6] that ensure the consistency of CMI estimator. We denote the point (x,y,z) as $\omega \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_z}$. Let $f(\omega) = p(x,y,z)$ be the joint density function of (X,Y,Z), $g(\omega) = p(x,z)p(y|z)$ be the joint density function of (X,Y,Z) under H_0, and $\phi(\omega) = \phi(x,y,z)$ be the joint density of (X,Y',Z) produced by Algorithm 1. The classifier in Algorithm 2 is trained using the label probability $\gamma_\theta(\omega) := P_\theta(l=1|\omega)$ with parameter θ. According to Algorithm 2, $P(l=1) = P(l=0) = 1/2$. Define the population binary-cross entropy loss over the joint distribution of data and label as

$$\text{BCE}(\gamma_\theta) = -\{E_{Wl}(l \log \gamma_\theta(W) + (1-l)\log(1-\gamma_\theta(W))\}.$$

Let $\gamma'(\omega) := \gamma_{\theta'}(\omega)$ be the point-wise minimizer of binary-cross entropy loss based on $f(\omega)$ and $\phi(\omega)$, and $\gamma''(\omega) := \gamma_{\theta''}(\omega)$ be the point-wise minimizer of binary-cross entropy loss based on $f(\omega)$ and $g(\omega)$.

Assumption (A1): $f(\cdot)$ and $\phi(\cdot)$ admit densities in a compact subset $\mathcal{W} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_z}$.

Assumption (A2): For some constant $\alpha, \zeta > 0$, $\alpha \leq f(\omega), \phi(\omega) \leq \zeta$, $\forall\omega$.

Assumption (A3): $\gamma'(\omega), \gamma''(\omega) \in [\tau, 1-\tau]$ and clip predictions such that $\gamma_\theta(\omega) \in [\tau, 1-\tau]$ $\forall\omega,\theta$, with $0 < \tau \leq \alpha/(\alpha+\zeta)$.

Assumption (A4): The classifier class $\mathcal{C}_\theta$ is parametrized by θ within a compact domain $\Theta \subset \mathbb{R}^h$. There exists a constant K such that $||\theta||_2 \leq K$, and the classifier's output is L-Lipschitz with respect to θ.

Assumption (A5): $\int p(z)^{1-1/d}dz \leq C_5$, $\forall d \geq 2$, where C_5 is a constant.

We denote the CMI estimator $\widehat{\text{CMI}}$ based on Algorithm 2 as $\widehat{D}_{KL}^{(n)}(f||\phi)$. The true CMI of (X,Y,Z) is $\text{CMI} := I(X;Y|Z) = D_{KL}(f||g)$. Then we have the following Lemma:

Lemma 7. *Under Assumptions 1 and 2 and (A1)-(A5), we have* $\widehat{D}_{KL}^{(n)}(f||\phi) \xrightarrow{P} D_{KL}(f||g)$.

Proof. By the definition of convergence in probability, it is sufficient to prove $\forall\delta > 0$, $\forall\eta > 0$, $\exists N$, when $n > N$, $P(|\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||g)| > \delta) < \eta$.

Note that

$$\begin{aligned}
&P(|\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||g)| > \delta) \\
&= P(|\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||\phi) + D_{KL}(f||\phi) - D_{KL}(f||g)| > \delta).
\end{aligned}$$

Applying Theorem 1 in [6], we have $\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||\phi) \xrightarrow{P} 0$, which means for $\delta/2 > 0$ and $\eta > 0$, $\exists N_1$, when $n > N_1$, $P(|\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||\phi)| > \delta/2) < \eta$.

Now consider the term $D_{KL}(f||\phi) - D_{KL}(f||g)$. Applying Lemma 3 in [6], we have $\gamma'(\omega)/\{1 - \gamma'(\omega)\} = f(\omega)/\phi(\omega)$ and $\gamma''(\omega)/\{1 - \gamma''(\omega)\} = f(\omega)/g(\omega)$. Then, by the definition of KL divergence, it follows that

$$\begin{aligned}
|D_{KL}(f||\phi) - D_{KL}(f||g)| &= \left| E_{f(\omega)} \log \frac{f(\omega)}{\phi(\omega)} - E_{f(\omega)} \log \frac{f(\omega)}{g(\omega)} \right| \\
&= \left| E_{f(\omega)} \left(\log \frac{\gamma'(\omega)}{1-\gamma'(\omega)} - \log \frac{\gamma''(\omega)}{1-\gamma''(\omega)} \right) \right| \\
&\leq E_{f(\omega)} \left| \log \frac{1-\gamma'(\omega)}{\gamma'(\omega)} - \log \frac{1-\gamma''(\omega)}{\gamma''(\omega)} \right| \\
&\leq \frac{1-\tau}{\tau} E_{f(\omega)} \left| \frac{1-\gamma'(\omega)}{\gamma'(\omega)} - \frac{1-\gamma''(\omega)}{\gamma''(\omega)} \right| \\
&= \frac{1-\tau}{\tau} E_{f(\omega)} \left| \frac{\phi(\omega) - g(\omega)}{f(\omega)} \right| \\
&= \frac{1-\tau}{\tau} \iiint |\phi(x,y,z) - g(x,y,z)| dxdydz \\
&= \frac{2(1-\tau)}{\tau} d_{TV}(\phi, g).
\end{aligned}$$

The second inequality follows from Lagrange's mean value theorem and Assumption (A3).

Applying Theorem 1 in [9], $\forall \epsilon_1 \leq \epsilon$ with ϵ being defined in Assumption 1, we have $d_{TV}(\phi, g) \leq b(n)$, where

$$b(n) = \frac{1}{2} \sqrt{\frac{\beta}{4} \frac{C_5 2^{1/d_z} \Gamma(1/d_z)}{(n\gamma_{d_z})^{1/d_z} d_z} + \frac{\beta \epsilon_1 G(2c_{d_z}\epsilon_1^2)}{4}} + \exp\left(-\frac{1}{2} n \gamma_{d_z} c_{d_z} \epsilon_1^{d_z+2} \right) + G(2c_{d_z}\epsilon_1^2).$$

Here, β is defined in Assumption 1, C_5 is defined in Assumption (A5), d_z is the dimension of Z, $\Gamma(\cdot)$ is the gamma function, γ_{d_z} is the volume of the unit radius l_2 ball in $\mathbb{R}^{d_z}$, c_{d_z} is defined in Assumption 2, and $\forall \delta > 0$, $G(\delta) = P(p(Z) \leq \delta)$.

Because ϵ_1 can be arbitrary small, we conclude that $\lim_{n\to\infty} b(n) = 0$. So we arrive at $\lim_{n\to\infty} |D_{KL}(f||\phi) - D_{KL}(f||g)| = 0$, which means for $\delta/2 > 0$, $\exists N_2$, when $n > N_2$, $|D_{KL}(f||\phi) - D_{KL}(f||g)| < \delta/2$.

Then for $\delta > 0$ and $\eta > 0$, take $N = \max(N_1, N_2)$, when $n > N$,

$$\begin{aligned}
&P(|\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||\phi) + D_{KL}(f||\phi) - D_{KL}(f||g)| > \delta) \\
&\leq P(|\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||\phi)| + |D_{KL}(f||\phi) - D_{KL}(f||g)| > \delta) \\
&\leq P(|\widehat{D}_{KL}^{(n)}(f||\phi) - D_{KL}(f||\phi)| > \delta/2) < \eta
\end{aligned}$$

holds. This finish the proof.

We therefore conclude that $\widehat{\text{CMI}}$ is a consistent estimator of CMI. When considering $\widehat{\text{CMI}}^{(b)}$ based on the sample $(\widetilde{\boldsymbol{X}}^{(b)}, \boldsymbol{Y}, \boldsymbol{Z})$ $(b = 1, \ldots, B)$ drawn from the k-nearest-neighbor local sampling mechanism as depicted in Algorithm 3, we can state: Under Assumptions 1, 2, (A1)-(A3) with $f(\omega)$ and $\phi(\omega)$ replaced by densities of the distribution of $(\widetilde{X}, Y, Z)$ and the corresponding 1-NN distribution, respectively, and Assumptions (A4)-(A5), $\forall b = 1, \ldots, B$, $\widehat{\text{CMI}}^{(b)}$ is a consistent estimator of $\text{CMI}^{(b)}$, where $\text{CMI}^{(b)} = I(\widetilde{X}^{(b)}; Y|Z)$. Now let's present the proof of Theorem 4.

Proof. Write $P(\cdot|H_1)$ as $P_{H_1}(\cdot)$. By Markov inequality, it follows that

$$
\begin{aligned}
P_{H_1}(p > \alpha) &= P_{H_1}\left(\frac{1 + \Sigma_{b=1}^{B_n} 1(\widehat{\mathrm{CMI}}^{(b)} \geq \widehat{\mathrm{CMI}})}{1 + B_n} > \alpha\right) \\
&\leq \frac{1}{\alpha(1+B_n)} E_{H_1}\left(1 + \sum_{b=1}^{B_n} 1(\widehat{\mathrm{CMI}}^{(b)} \geq \widehat{\mathrm{CMI}})\right) \\
&= \frac{1}{\alpha(1+B_n)} + \frac{B_n}{\alpha(1+B_n)} P_{H_1}\big(\widehat{\mathrm{CMI}}^{(1)} \geq \widehat{\mathrm{CMI}}\big) \\
&\leq \frac{1}{\alpha(1+B_n)} + \frac{1}{\alpha} P_{H_1}\big(\widehat{\mathrm{CMI}}^{(1)} \geq \widehat{\mathrm{CMI}}\big).
\end{aligned}
$$

Because $\widetilde{X}^{(1)} \perp\!\!\!\perp Y|Z$, $\mathrm{CMI}^{(1)} = I(\widetilde{X}^{(1)}; Y|Z) = 0$. Then, $\forall \delta > 0$,

$$
\begin{aligned}
P_{H_1}\big(\widehat{\mathrm{CMI}}^{(1)} \geq \widehat{\mathrm{CMI}}\big) &\leq P_{H_1}\big(\{\widehat{\mathrm{CMI}} \leq \widehat{\mathrm{CMI}}^{(1)}\} \cap \{|\widehat{\mathrm{CMI}}^{(1)} - \mathrm{CMI}^{(1)}| \leq \delta\}\big) \\
&\quad + P_{H_1}\big(|\widehat{\mathrm{CMI}}^{(1)} - \mathrm{CMI}^{(1)}| > \delta\big) \\
&\leq P_{H_1}(\widehat{\mathrm{CMI}} \leq \delta) + P_{H_1}\big(|\widehat{\mathrm{CMI}}^{(1)} - \mathrm{CMI}^{(1)}| > \delta\big).
\end{aligned}
$$

Next, we have

$$
\begin{aligned}
P_{H_1}(\widehat{\mathrm{CMI}} \leq \delta) &\leq P_{H_1}(\{\widehat{\mathrm{CMI}} \leq \delta\} \cap \{|\widehat{\mathrm{CMI}} - \mathrm{CMI}| \leq \delta\}) + P_{H_1}(|\widehat{\mathrm{CMI}} - \mathrm{CMI}| > \delta) \\
&\leq P_{H_1}(\mathrm{CMI} - \delta \leq \widehat{\mathrm{CMI}} \leq \delta) + P_{H_1}(|\widehat{\mathrm{CMI}} - \mathrm{CMI}| > \delta).
\end{aligned}
$$

Thus, we conclude that

$$
\begin{aligned}
P_{H_1}(p \leq \alpha) \geq& 1 - \frac{1}{\alpha(1+B_n)} - \frac{1}{\alpha}\big[P_{H_1}(\mathrm{CMI} - \delta \leq \widehat{\mathrm{CMI}} \leq \delta) \\
&+ P_{H_1}(|\widehat{\mathrm{CMI}} - \mathrm{CMI}| > \delta) + P_{H_1}\big(|\widehat{\mathrm{CMI}}^{(1)} - \mathrm{CMI}^{(1)}| > \delta\big)\big].
\end{aligned}
$$

Under H_1, $\mathrm{CMI} > 0$. Take $\delta = \mathrm{CMI}/4 > 0$, we obtain

$$
\lim_{n\to\infty} P_{H_1}(p \leq \alpha) \to 1.
$$

B Additional Empirical Results

B.1 The choice of the neighbor order k

To investigate the impact of the parameter k on our proposed approach, we employ a linear uniform model. To accomplish this, we generate synthetic data in the following manner:

$$
\begin{aligned}
&H_0 : X = \epsilon_x, \;\; Y = \epsilon_y, \;\; \text{and} \;\; Z \sim \mathrm{Uniform}(-1, 1), \\
&H_1 : X = \epsilon_x, \;\; Y = \alpha X + 0.5\epsilon_y, \;\; \text{and} \;\; Z \sim \mathrm{Uniform}(-1, 1),
\end{aligned} \tag{13}
$$

where ϵ_x and ϵ_y are generated independently from the uniform distribution over the interval $[-1, 1]$. The parameter α is randomly generated within the range of $[0, 2]$. As is shown in Figure 3, our method achieves effective control of type I error and exhibits the highest power under H_1 across all dimensions when $k = 7$. Therefore, we consistently set $k = 7$ in all experiments.

B.2 Empirical results for Scenario (13)

We demonstrate the effectiveness of our approach and compare it with alternative methods in Scenario (13). The results are shown in Figure 4, which pertains to high-dimensional Z, and Figure 5, which focuses on low-dimensional Z. The results consistently demonstrate that our test achieves favorable performance in terms of the type I error and power under H_1. Although LPCIT, CMIknn, and NNSCIT effectively control the type I error, they exhibit noticeably lower power compared to our method, when the dimension exceeds 60, often by a substantial margin. Furthermore, KCIT, GCIT, and CCIT all yield high power under H_1, but they either always or sometimes suffer from inflated type I errors.

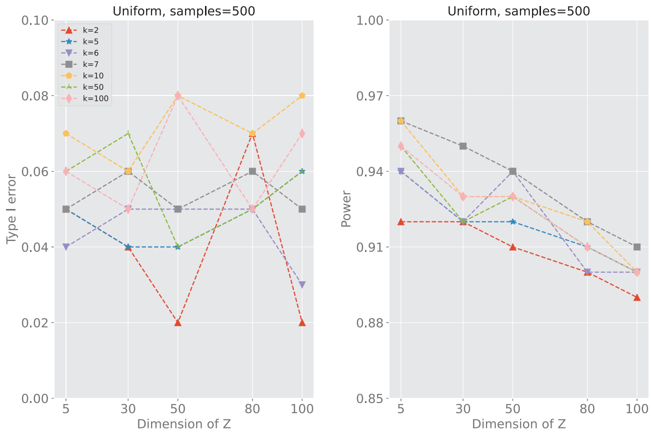

Figure 3: Comparison of the type I error (lower is better) and power under H_1 (higher is better) for our test in Scenario (13) across different values of k.

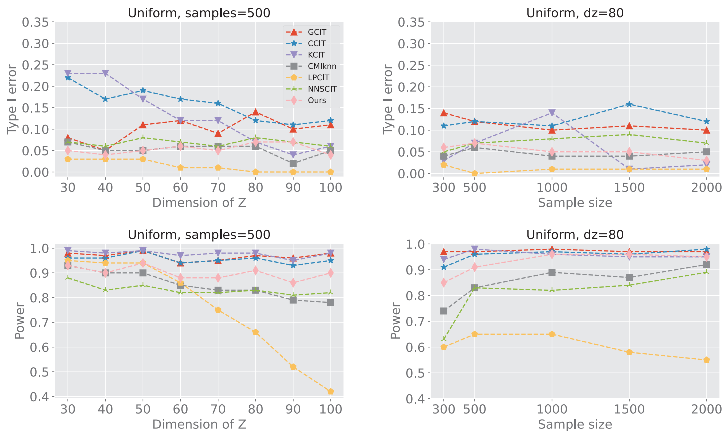

Figure 4: Comparison of the type I error (lower is better) and power under H_1 (higher is better) of our test with six SOTA tests in Scenario (13). **Left:** The results when varying the dimension of Z. **Right:** The results when varying the sample size.

B.3 Additional empirical results for Scenario I

In Figure 6, we present the type I error and power under H_1 in low dimensions of Z ranging from 5 to 30 for Scenario I (Eq. (9)) with Gaussian or Laplace noises. It can be observed that our test and LPCIT consistently achieve good and stable performance in terms of type I error and power under H_1, when the dimensionality of Z is lower than 30. On the other hand, GCIT, CCIT, and KCIT exhibit high power under H_1 but fail to control the type I error. NNSCIT and CMIknn demonstrate relatively good control of type I errors but lack sufficient power under H_1.

B.4 Computational efficiency analysis

Figure 7 shows the timing performance of all methods for a single test under Scenario I with Laplace noises. Our test is found to be highly computationally efficient even when dealing with large sample sizes and high-dimensional conditioning sets. In contrast, CMIknn and CCIT for sample

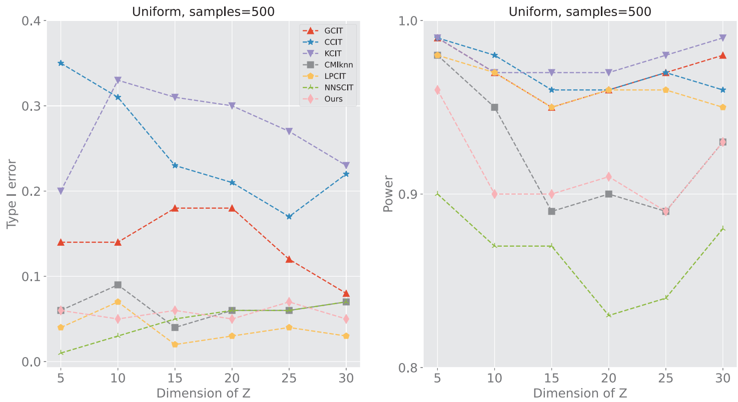

Figure 5: Comparison of the type I error (lower is better) and power under H_1 (higher is better) of our test with six SOTA tests in Scenario (13).

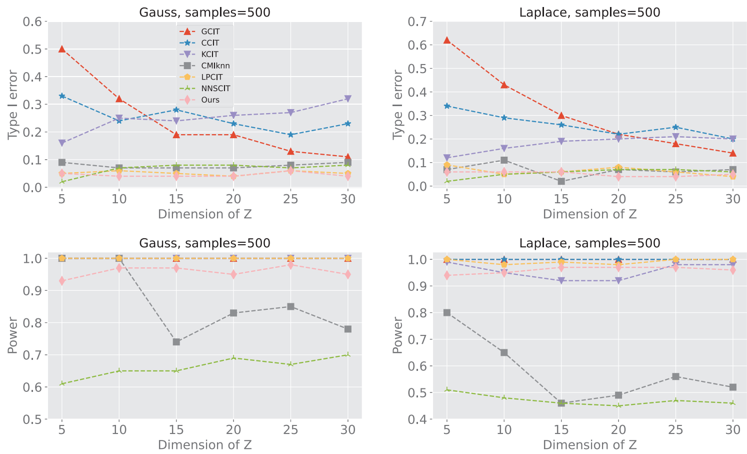

Figure 6: Comparison of the type I error (lower is better) and power under H_1 (higher is better) of our method with six SOTA methods on the post-nonlinear model under Gaussian or Laplace distributions in Scenario I. **Left:** The results under Gaussian distribution. **Right:** The results under Laplace distribution.

sizes exceeding 1000, and LPCIT for dimension of Z higher than 50 are impractical due to their prohibitively long running time.

B.5 The detailed experimental setup for Scenario III

For the chain structure $Y \to Z \to X$, we generate synthetic data as follows:

$$H_0: \ Y \sim N(1,1), \ Z = Ya + \epsilon_1, \ X = Z^T b + \epsilon_2,$$
$$H_1: \ Y \sim N(1,1), \ Z = Ya + \epsilon_1, \ X = Z^T b + Y + \epsilon_2,$$

where a and b are both d_z-dimensional, the entries of a and b are both randomly and uniformly sampled from $[0, 0.3]$, ϵ_1 is generated from a d_z-dimensional standard multivariate Gaussian distribution, and ϵ_2 is sampled from a standard Gaussian distribution.

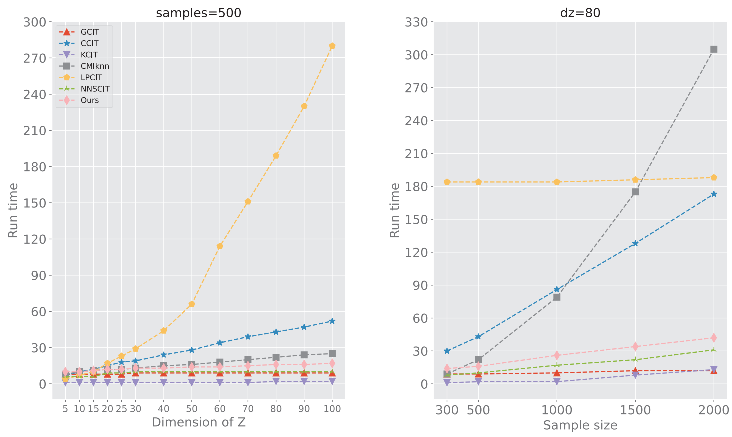

Figure 7: Running times in seconds as a function of sample size or dimension of Z on the post-nonlinear model under Laplace distribution in Scenario I. **Left:** The results when varying the dimension of Z. **Right:** The results when varying the sample size.

C Real Data Analysis

In order to showcase the superior performance of our test, we conduct a comparative evaluation against other state-of-the-art (SOTA) CI tests using real datasets. We assess the effectiveness of our method along with six SOTA approaches on two specific datasets: the ABALONE dataset [1] and the Flow-Cytometry dataset [8].

C.1 Real ABALONE dataset

The ABALONE dataset [1] comprises measurements obtained from a study conducted to predict the age of abalones based on their physical characteristics. The dataset is publicly available at the UCI Machine Learning Repository and can be downloaded from https://archive.ics.uci.edu/ml/datasets/abalone. In our evaluation, we consider the graph structure recovered by [4] as the ground truth, as depicted in Figure 4 of their paper. This graph represents the causal relationships among the 8 variables in the dataset. We specifically select 35 CI relations and 35 non-CI relations from this graph. The philosophy used is that a node X is independent of all other nodes Y in the graph when conditioned on its parents, children, and parents of children [2, 9]. Additionally, if there exists a direct edge between node X and node Y in the graph, they are never conditionally independent given any other set of variables. As a result, the conditioning set Z can be arbitrarily selected from the remaining nodes. The dataset consists of 4177 samples, and d_z varies from 1 to 6.

In order to evaluate the performance of various tests, we utilize precision, recall, and F-score as evaluation metrics. Precision is calculated as TP/(TP+FP), where TP represents the number of true CI instances correctly identified, and FP represents the number of non-CI instances incorrectly identified as CI. Recall is calculated as TP/(TP+FN), where FN represents the number of CI instances not identified. The F-score is then computed as the harmonic mean of precision and recall, given by 2 $\times$ precision $\times$ recall / (precision + recall) [3]. TN represents the number of correctly identified true non-CI instances. Table 1 presents the results for all methods. It should be noted that we do not record the results for GCIT as it does not correctly identify any CI relations. Our approach successfully identifies 31 CI relations and 32 non-CI relations, achieving the highest F-score among the testing methods, while maintaining high precision and recall.

C.2 Real Flow-Cytometry dataset

The Flow-Cytometry dataset is a widely used benchmark in the field of causal structure learning [7, 10]. This dataset captures the expression levels of proteins and phospholipids in human cells [8].

Table 1: The TP, TN, precision (pre), recall (rec) and F-score of our test and six SOTA methods for the real ABALONE dataset.

Method	TP	TN	Pre	Rec	F-score
KCIT	5	35	1	0.1429	0.2501
CCIT	12	34	0.9231	0.3429	0.5
CMIknn	22	35	1	0.6286	0.7720
LPCIT	5	35	1	0.1429	0.2501
NNSCIT	33	6	0.5323	0.9429	0.6805
Ours	31	32	0.9118	0.8857	0.8986

The data can be obtained from the website https://www.science.org/doi/10.1126/science.1105809. In our evaluation, we consider the consensus graph proposed in [5] as the ground truth, which has also been adopted by [9] for verifying CI relations. Figure 5(a) in [5] illustrates the causal relationships among the 11 proteins in the dataset. Following the philosophy outlined in Section C.1, we select 50 CI relations and 40 non-CI relations from this graph. The number of samples is 1755 and d_z varies from 1 to 9.

Table 2 presents the results for all tests. Our method outperforms other approaches by correctly identifying 47 CI relations and achieving the highest recall and F-score.

Table 2: The TP, TN, precision (pre), recall (rec) and F-score of our test and six SOTA methods for the real Flow-Cytometry dataset.

Method	TP	TN	Pre	Rec	F-score
KCIT	32	30	0.7619	0.64	0.6957
CCIT	33	29	0.75	0.66	0.7021
CMIknn	41	26	0.7455	0.82	0.7810
GCIT	40	24	0.7143	0.8	0.7547
LPCIT	38	25	0.7170	0.76	0.7379
NNSCIT	33	26	0.7021	0.66	0.6804
Ours	47	23	0.7344	0.94	0.8246

References

[1] Catherine L Blake and Christopher J Merz. Uci repository of machine learning databases, 1998.

[2] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[3] Honghao Li, Vincent Cabeli, Nadir Sella, and Hervé Isambert. Constraint-based causal structure learning with consistent separating sets. *In Advances in Neural Information Processing Systems*, 32, 2019.

[4] Dimitris Margaritis. Distribution-free learning of bayesian network structure in continuous domains. In *Proceedings of the 20th National Conference on Artificial Intelligence*, volume 5, pages 825–830, 2005.

[5] Joris M Mooij and Tom Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 431–439, 2013.

[6] Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, pages 1083–1093, 2020.

[7] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *In Advances in Neural Information Processing Systems*, 33, 2020.

[8] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[9] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. *In Advances in Neural Information Processing Systems*, 30, 2017.

[10] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020.