Extracting Rare Dependence Patterns via Adaptive Sample Reweighting

Yiqing Li[1,*], Yewei Xia[1,2,*], Xiaofei Wang[1,3], Zhengming Chen[1,4], Liuhua Peng[5], Mingming Gong[1,5], Kun Zhang[1,6]

[1]Department of Machine Learning, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai, China
[3]KLAS and School of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin, China
[4]College of Mathematics and Computer, Shantou University, Shantou, Guangdong, China
[5]School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC, Australia
[6]Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Preliminary

Dependence testing

- Given: Samples from a distribution P_{XY}
- Goal: Are X and Y independent?

X	Y
	A large animal who slings slobber, exudes a distinctive houndy odor, and wants nothing more than to follow his nose.
	Their noses guide them through life, and they're never happier than when following an interesting scent.
	A responsive, interactive pet, one that will blow in your ear and follow you everywhere.

Hilbert-Schmidt Independence Criterion

$$\|\Sigma_{XY}\|^2_{\mathcal{HS}} = \|\mathbb{E}_{\mathbb{P}_{XY}}[(\psi_X - \mu_X) \otimes (\phi_Y - \mu_Y)]\|^2_{\mathcal{HS}}.$$

where $\mu_X \triangleq \mathbb{E}_{\mathbb{P}_X}[\psi(X)]$, $\mu_Y \triangleq \mathbb{E}_{\mathbb{P}_Y}[\phi(Y)]$

MMD

Hypothesis Testing

p-value: the probability of obtaining results as extreme as, or more extreme than, the observed results, assuming the null hypothesis is true.

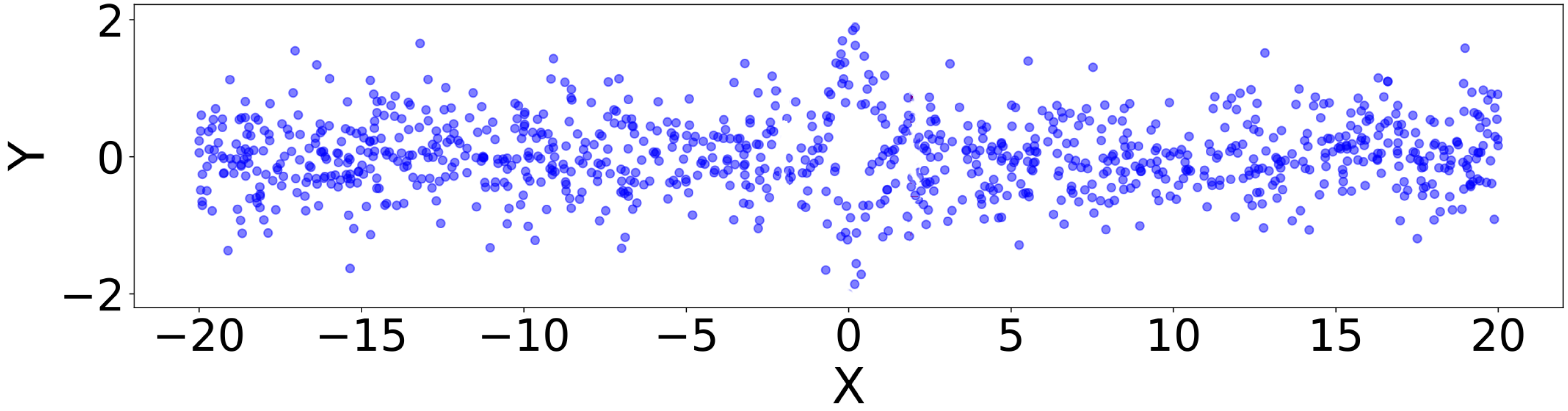
p-value $< \alpha$: reject.

p-value $> \alpha$: fail to reject.

[Credit to Arthur Gretton]

A Motivating Example

$X \sim U(-20,20), Y = s \cdot e^{-x^2} + \epsilon, \epsilon \sim \mathcal{N}(0,0.25), s \in \{-1,1\}$ with equal probability.



- p-value of HSIC with default settings on the whole sample is **0.1359** >0.05, fails to reject.
- Just a specific case, but it does reflect the shortcomings of HSIC in dealing with “extreme cases”.

Examples in Psychology

Time spent indulging in social media (t)

Probability of depressive disorders (p)

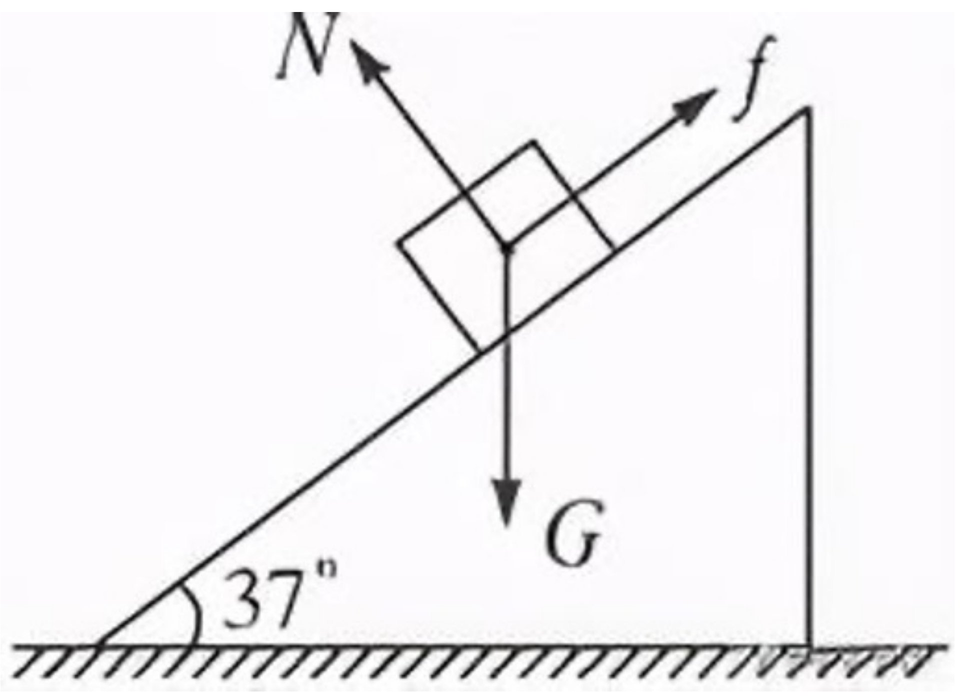
Only excessive usage (large t) ➡️ Probability of depressive 📈

But the percentage of people with excessive usage time is small, which makes it hard to detect dependence between t and p.

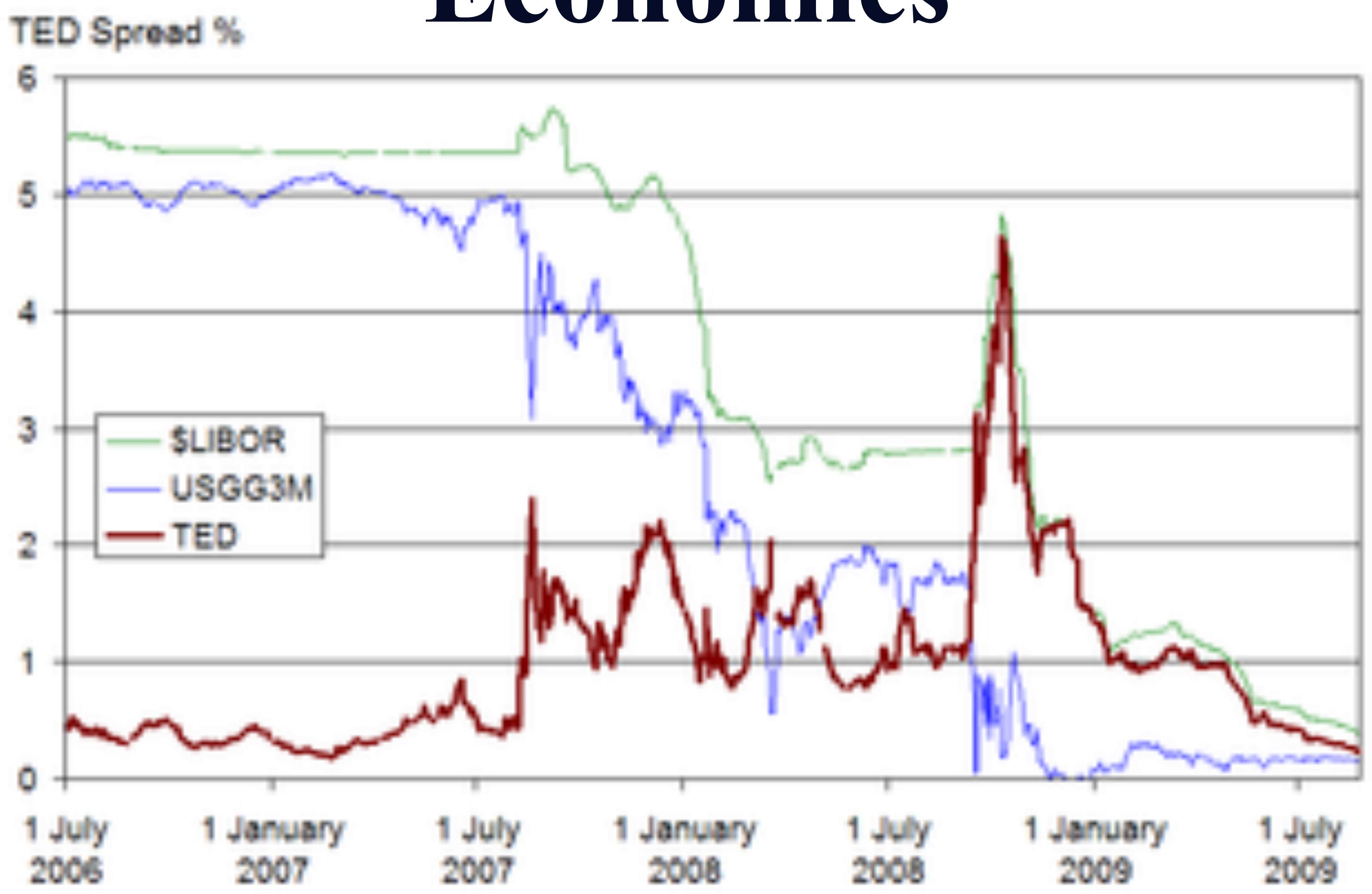
Keles, B., McCrae, N., and Grealish, A. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. International journal of adolescence and youth, 25(1):79–93, 2020.

Examples in Other Fields

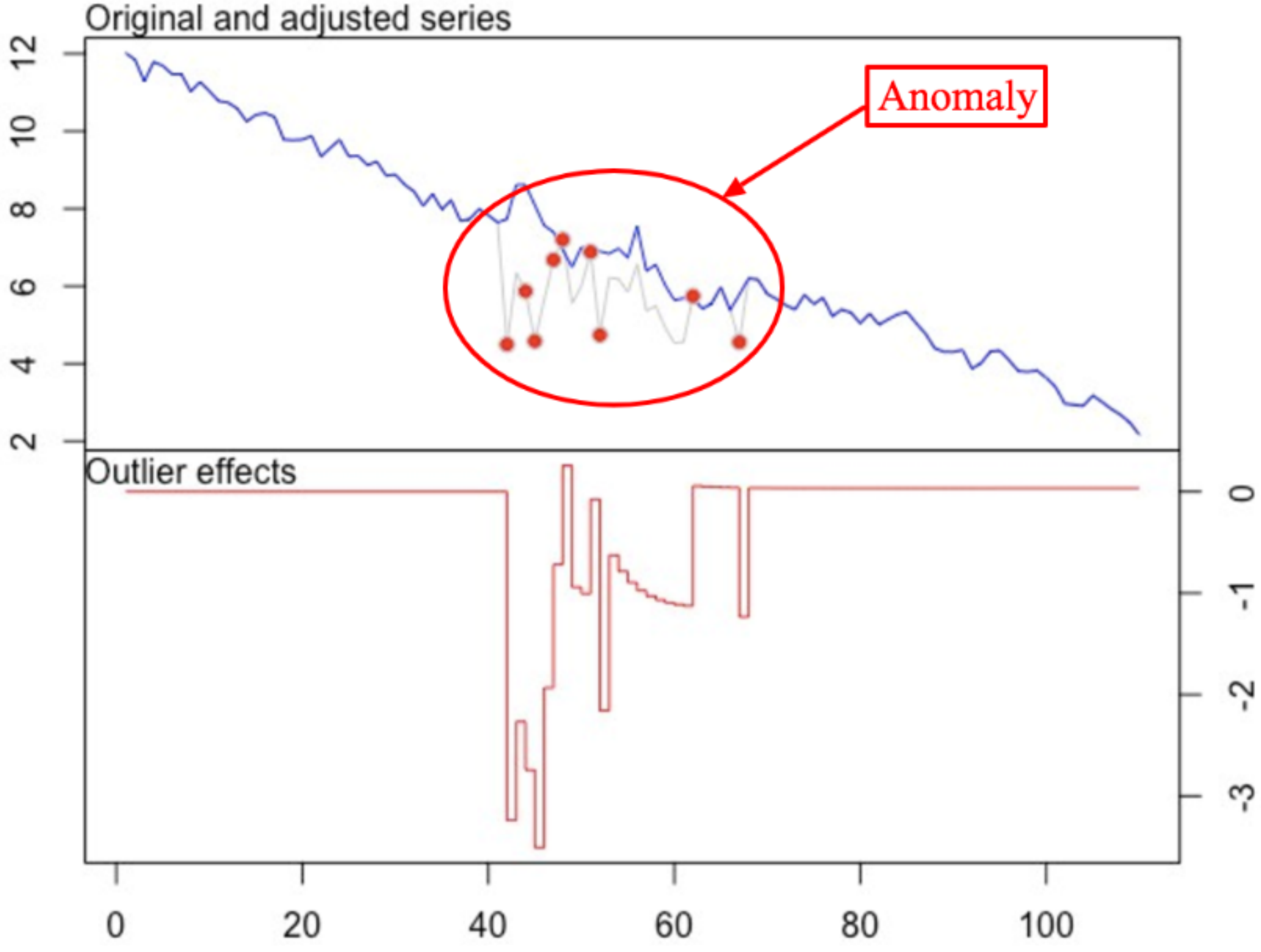
Physics



Economics

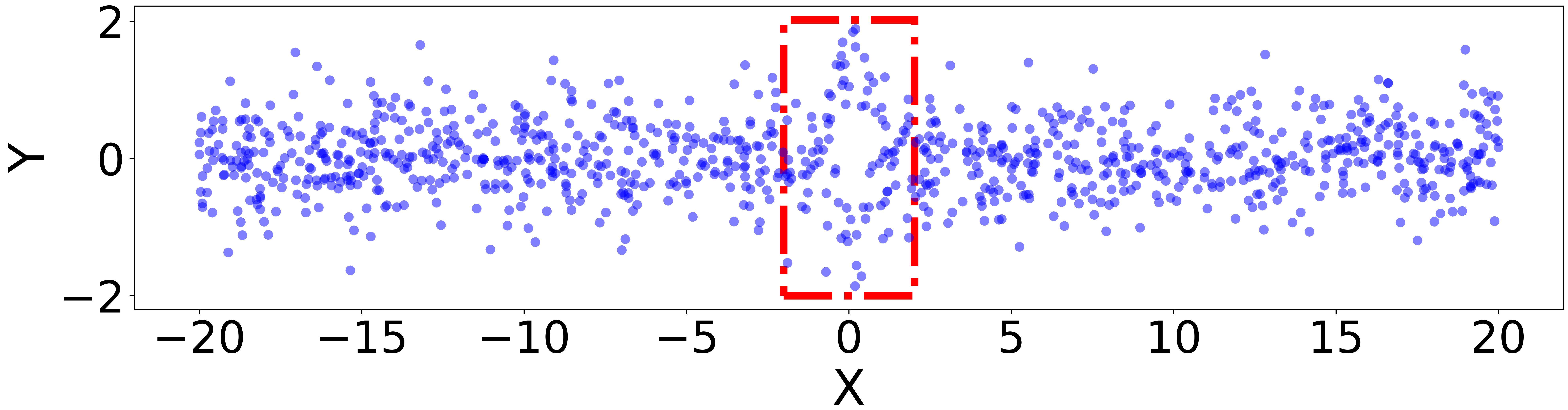


Anomaly detection



A Motivating Example

$X \sim U(-20,20), Y = s \cdot e^{-x^2} + \epsilon, \epsilon \sim \mathcal{N}(0,0.25), s \in \{-1,1\}$ with equal probability.



- p-value of HSIC with default settings on the whole sample is **0.1359** >0.05, fails to reject.
- p-value on the samples within the red rectangle is $\mathbf{6.8 * 10^{-11}}$ <0.05, reject.

Rare Dependence

- **Definition** 📖: The dependence patterns between two variables are **significant only within a small range** of the entire distribution's support.

- **Goal** 🎯: How to detect dependence even in the presence of rare dependence.

- **Idea** 💡: **Automatically** identifies and amplifies the **significantly dependent sub-population** to make the dependence pattern obvious and easier to detect.

Reweighting Function and Reweighted Distribution

- **Idea**💡: **Automatically** identifies and amplifies the **significantly dependent sub-population** to make the dependence pattern obvious and easier to detect.
- ➡️ Change the original distribution! Resampling/Reweighting
- Reweighting function: $\mathscr{B} \triangleq \left\{ \beta : \mathscr{C} \to \mathbb{R}^{\geq 0} \mid \mathbb{E}_{\mathbb{P}_{XY}}[\beta(C)] = 1 \right\}$. $\tilde{\mathbb{P}}(X, Y) = \beta(C)\mathbb{P}(X, Y)$.
- C is a reference variable that can be either X or Y.
- If X and Y are independent and C is either X or Y **but not both**, then X and Y are still independent in the reweighted distribution of (X, Y) with weight $\beta(C)$.

Reweighted HSIC

- Reweighting function: $\mathscr{B} \triangleq \left\{\beta : \mathscr{C} \to \mathbb{R}^+ \mid \mathbb{E}_{\mathbb{P}_{XY}}[\beta(C)] = 1\right\}$. $\tilde{\mathbb{P}}(X, Y) = \beta(C)\mathbb{P}(X, Y)$.

- Question: What is a good reweighting function for us?
- A possible criterion: maximize the dependence pattern in $\tilde{\mathbb{P}}(X, Y)$.

$$\mathrm{HSIC}(X, Y) \triangleq \|\Sigma_{XY}\|_{HS}^2 = \|\mathbb{E}_{\mathbb{P}_{XY}}[(\psi_X - \mu_X) \otimes (\phi_Y - \mu_Y)]\|_{HS}^2.$$

$$\begin{aligned}\mathrm{HSIC}^\beta(X, Y) &\triangleq \left\| \mathbb{E}_{\tilde{\mathbb{P}}} \left[(\psi_X - \mathbb{E}_{\tilde{\mathbb{P}}}[\psi_X]) \otimes (\phi_Y - \mathbb{E}_{\tilde{\mathbb{P}}}[\phi_Y])\right] \right\|_{HS}^2 \\ &= \left\| \mathbb{E}_{\mathbb{P}} \left[\beta(X)(\psi_X - \mathbb{E}_{\mathbb{P}}[\beta(X)\psi_X]) \otimes (\phi_Y - \mathbb{E}_{\mathbb{P}}[\beta(X)\phi_Y])\right] \right\|_{HS}^2\end{aligned}$$

Reweighting Function and Reweighted Distribution

- Sample version: $$\mathrm{HSIC}_b^{\beta}(\mathscr{D}) = \frac{1}{n^2}\mathrm{Tr}\left[\mathbf{K}_X\mathbf{H}_\beta\mathbf{K}_Y\mathbf{H}_\beta\right],$$

$$\mathbf{H}_\beta \triangleq \mathbf{D}_\beta(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{D}_\beta) \qquad \mathbf{D}_\beta \triangleq \mathrm{diag}(\beta_1, \ldots, \beta_n)$$

- V-statistics:

$$\mathrm{HSIC}_b^{\beta}(\mathscr{D}) = \frac{1}{n^4}\sum_{i,j,q,r}^{n} h_{ijqr}^{\beta} \qquad h_{ijqr}^{\beta} \triangleq \frac{1}{4!}\sum_{(s,t,u,v)}^{(i,j,q,r)} (\beta_s\beta_t k_X^{st}k_Y^{st} + \beta_s\beta_t\beta_u\beta_v k_X^{st}k_Y^{uv} - 2\beta_s\beta_t\beta_u k_X^{st}k_Y^{su}).$$

Theorem 3.4 (Null distribution). *Under $\mathcal{H}_0$, we have $\mathbb{E}_i h_{ijqr}^{\beta} = 0$. In this case, $\mathrm{HSIC}_b^{\beta}(\mathcal{D})$ converges in distribution to a weighted sum of $\mathcal{X}^2$ variables, i.e.,*

$$n\mathrm{HSIC}_b^{\beta}(\mathscr{D}) \xrightarrow{d} \sum_{l=1}^{\infty} \lambda_l^{\beta}\chi_{1l}^2,$$

Theorem 3.5. *When $\mathrm{HSIC}^{\beta}(X,Y) > 0$, $\mathrm{HSIC}_b^{\beta}(\mathcal{D})$ converges in distribution to a Gaussian according to:*

$$\sqrt{n}\left(\mathrm{HSIC}_b^{\beta}(\mathscr{D}) - \mathrm{HSIC}^{\beta}(X,Y)\right) \xrightarrow{d} \mathscr{N}(0,\sigma_\beta^2).$$

$$\sigma_\beta^2 = 16(\mathbb{E}_i(\mathbb{E}_{j,q,r}h_{ijqr}^{\beta})^2 - \mathrm{HSIC}^{\beta}(X,Y)^2)$$

Reweighted HSIC

- Optimization Problem:

$$\arg\min_{\beta} -\log \hat{J}_{\beta}^{UI} + \lambda_1 \|\omega\|^2_{\mathscr{F}_X} + \frac{\lambda_2}{n} \sum_{i=1}^{n} (\beta_i - 1)^2,$$

$$\text{s.t. } \beta_i \geq 0, \ \sum_{i=1}^{n} \beta_i = n,$$

- $$\beta(X) = \langle \psi_X^T, \omega \rangle_{\mathscr{F}_X}, \text{ where } \omega \triangleq \psi_X^T \alpha = \sum_{i=1}^{n} \alpha_i \psi(x_i)^T$$

$$\|\omega\|^2_{\mathscr{F}_X} = \alpha^T \mathbf{K}_X \alpha$$

Reweighted HSIC

Algorithm 1 Reweighted HSIC (RHSIC)

1: **Input:** $\mathcal{D}$: samples. C: reference variable. α: significance level. B: the number of permutations.
2: **Output:** p-value and test statistics value.
3: Split $\mathcal{D}$ into $\mathcal{D}_{tr} = \{x_{tr}, y_{tr}\}$ and $\mathcal{D}_{te} = \{x_{te}, y_{te}\}$.
4: Optimize the constrained problem (9) on $\mathcal{D}_{tr}$, to obtain the reweighting function $\hat{\beta}(\cdot)$.
5: Use $\hat{\beta} = \hat{\beta}(x_{te})$ to calculate $T_{obs} = \text{HSIC}_b^{\hat{\beta}}(\mathcal{D}_{te})$.
6: **for** all $k \in \{1, \ldots, B\}$ **do**
7: Permute y_{te} to get $\tilde{y}_{te}^k$ and $\tilde{\mathcal{D}}_{te}^k = x_{te} \cup \tilde{y}_{te}^k$.
8: Calculate k-th statistics $T_k = \text{HSIC}_b^{\hat{\beta}}(\tilde{\mathcal{D}}_{te}^k)$.
9: **end for**
10: Compute p-value by $p = \frac{1}{B} \sum_{k=1}^{B} \mathbb{I}[T_k \geq T_{obs}]$ where $\mathbb{I}$ denotes the indicator function.

Generalization Guarantee

$$
\begin{aligned}
&\mathrm{HSIC}^{\beta^*}(X,Y) - \mathrm{HSIC}^{\hat{\beta}}(X,Y) \\
&= \underbrace{\left[\mathrm{HSIC}^{\beta^*}(X,Y) - \mathrm{HSIC}_b^{\beta^*}(\mathscr{D})\right]}_{A} + \underbrace{\left[\mathrm{HSIC}_b^{\beta^*}(\mathscr{D}) - \mathrm{HSIC}_b^{\hat{\beta}}(\mathscr{D})\right]}_{B} + \underbrace{\left[\mathrm{HSIC}_b^{\hat{\beta}}(\mathscr{D}) - \mathrm{HSIC}^{\hat{\beta}}(X,Y)\right]}_{C} \\
&\le \underbrace{\sup_{\beta\in\mathscr{B}} \left[\mathrm{HSIC}^{\beta}(X,Y) - \mathrm{HSIC}_b^{\beta}(\mathscr{D})\right]}_{A'} + 0 + \underbrace{\left[\mathrm{HSIC}_b^{\hat{\beta}}(\mathscr{D}) - \mathrm{HSIC}^{\hat{\beta}}(X,Y)\right]}_{C} \\
&\le \underbrace{2\sup_{\beta\in\mathscr{B}} \left|\mathrm{HSIC}^{\beta}(X,Y) - \mathrm{HSIC}_b^{\beta}(\mathscr{D})\right|}_{A''}
\end{aligned}
$$

Theorem 3.7 (Uniform Bound) Suppose $\mathcal{X} \subset \mathbb{R}^d$ is a closed and bounded space and the values of the kernels k_X and k_Y are also bounded. Assume that the reweighting functions $\beta \in \mathcal{B}$ are continuous and Lipschitz. Then with probability at least 1-δ, we have

$$
\sup_{\beta\in\mathscr{B}} \left|\mathrm{HSIC}_b^{\beta}(\mathscr{D}) - \mathrm{HSIC}^{\beta}(X,Y)\right| \sim \mathscr{O}\left(\sqrt{\frac{1}{n}\log\frac{1}{\delta} + \frac{\log n}{n^{\frac{2}{3}}}} + \frac{1}{n^{\frac{1}{3}}}\right).
$$

Conditional Independence Version

$$\mathscr{B} = \left\{ \beta : \mathscr{C} \times \mathscr{Z} \to \mathbb{R}^{\geq 0} \mid \mathbb{E}_{\mathbb{P}_{XY|Z}}[\beta(C,Z)] = 1 \right\}. \qquad \tilde{\mathbb{P}}(X, Y \mid Z) = \beta(C,Z)\mathbb{P}(X, Y \mid Z).$$

- Population version: $J_\beta^{CI} \triangleq \|\Sigma^\beta_{\ddot{X}Y|Z}\|^2_{HS} = \left\| \mathbb{E}_{\tilde{\mathbb{P}}} \left[(\psi^\beta_{\ddot{X}|Z} - \mathbb{E}_{\tilde{\mathbb{P}}}[\psi^\beta_{\ddot{X}|Z}]) \otimes (\phi^\beta_{Y|Z} - \mathbb{E}_{\tilde{\mathbb{P}}}[\phi^\beta_{Y|Z}]) \right] \right\|^2_{HS}$

  where $\psi^\beta_{\ddot{X}|Z} \triangleq \psi_{\ddot{X}} - \mathbb{E}_{\tilde{\mathbb{P}}}[\psi_{\ddot{X}} | Z]$, $\phi^\beta_{Y|Z} \triangleq \phi_Y - \mathbb{E}_{\tilde{\mathbb{P}}}[\phi_Y | Z]$.

- Sample version: $\hat{J}_\beta^{CI} = \frac{1}{n^2} \operatorname{Tr} \left[\widetilde{\mathbf{K}}^\beta_{\ddot{X}|Z} \widetilde{\mathbf{K}}^\beta_{Y|Z} \right]$

$$\widetilde{\mathbf{K}}^\beta_{\ddot{X}|Z} := \mathbf{R}^\beta_Z \widetilde{\mathbf{K}}^\beta_{\ddot{X}} \mathbf{R}^{\beta^T}_Z \mathbf{D}_\beta, \ \widetilde{\mathbf{K}}^\beta_{Y|Z} := \mathbf{R}^\beta_Z \widetilde{\mathbf{K}}^\beta_{Y} \mathbf{R}^{\beta^T}_Z \mathbf{D}_\beta \qquad \mathbf{R}^\beta_Z = \epsilon \left[\widetilde{\mathbf{K}}^\beta_Z \mathbf{D}_\beta + \epsilon \mathbf{I} \right]^{-1}$$

- Threshold estimation: conditional permutation [Runge, 2018].

Runge, J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information.
In International Conference on Artificial Intelligence and Statistics, pp. 938–947. PMLR, 2018.

Causal Discovery in the Presence of Rare Dependence

Assumption B.1. $\forall X, Y \in \mathbf{V}$, $Z \subseteq \mathbf{V} \setminus \{X, Y, C\}$, if $\mathrm{KCIT}(X, Y|Z)$ rejects the null hypothesis, then $X \not\perp\!\!\!\perp Y|Z$. Besides, if both $\mathrm{KCIT}(X, Y|Z)$ and $\mathrm{RKCIT}^{\beta(C)}(X, Y|Z)$ fails to reject the null hypothesis, then $X \perp\!\!\!\perp Y|Z$.

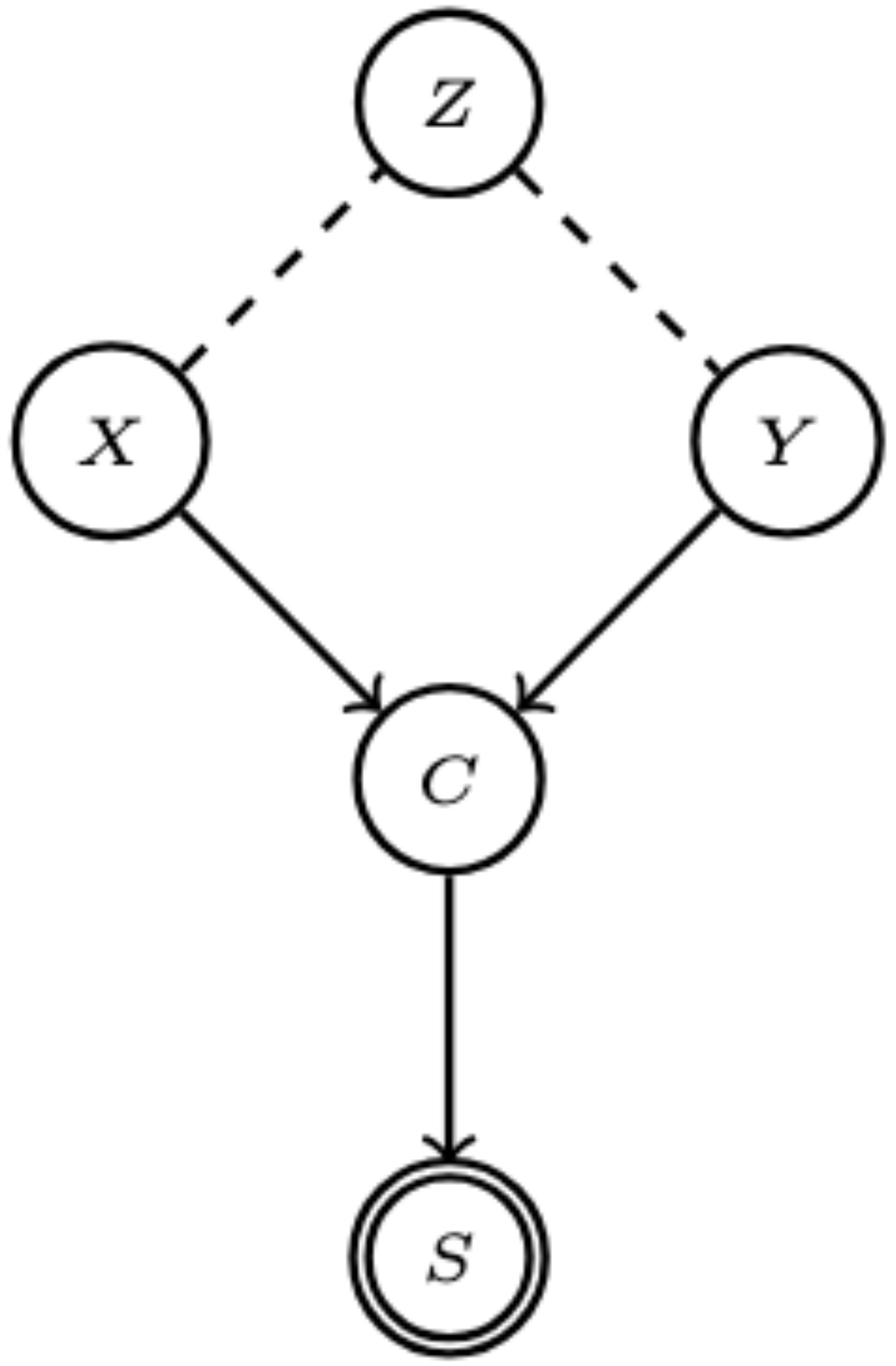
Rule 1. *$\forall\ X, Y \in \mathbf{V}$, if $\exists Z \subseteq \mathbf{V} \setminus \{X, Y, C\}$ s.t. both $\mathrm{KCIT}(X, Y|Z)$ and $\mathrm{RKCIT}^{\beta(C)}(X, Y|Z)$ fail to reject the null hypothesis, then X and Y are not adjacent in G.*

Proposition B.2. *For a pair of variables $X, Y \in V$, suppose that $\exists Z \subseteq V \setminus \{X, Y, C\}$ s.t. $\mathrm{KCIT}(X, Y|Z)$ fails to reject the null hypothesis. Besides, for all these Z, we have that $\mathrm{RKCIT}^{\beta(C)}(X, Y|Z)$ rejects the null hypothesis. Then, under Assumption 4.1, i) X and Y are adjacent with a rare dependence, or ii) X and Y are not adjacent in G and C must be the direct common effect of X and Y.*

Causal Discovery in the Presence of Rare Dependence

Assumption B.1. $\forall X, Y \in \mathbf{V}$, $Z \subseteq \mathbf{V} \setminus \{X, Y, C\}$, if $\mathrm{KCIT}(X, Y|Z)$ rejects the null hypothesis, then $X \not\perp\!\!\!\perp Y|Z$. Besides, if both $\mathrm{KCIT}(X, Y|Z)$ and $\mathrm{RKCIT}^{\beta(C)}(X, Y|Z)$ fails to reject the null hypothesis, then $X \perp\!\!\!\perp Y|Z$.

Rule 2. *For two variables $X, Y \in \mathbf{V}$ that satisfy the condition in Proposition B.2, if there exists $Z \subseteq \mathbf{V} \backslash \{X, Y, C\}$, such that $\mathrm{RKCIT}^{\beta(C^{perm})}(X, Y|Z)$ fail to reject the null hypothesis, then X and Y are not adjacent in G. Here C^{perm} denotes the shuffled C in dataset $\mathcal{D}$.*



Causal Discovery in the Presence of Rare Dependence

Algorithm

Rule 1. *$\forall\ X, Y \in \mathbf{V}$, if $\exists Z \subseteq \mathbf{V} \backslash \{X, Y, C\}$ s.t. both $\mathrm{KCIT}(X, Y|Z)$ and $\mathrm{RKCIT}^{\beta(C)}(X, Y|Z)$ fail to reject the null hypothesis, then X and Y are not adjacent in G.*

Rule 2. *For two variables $X, Y \in \mathbf{V}$ that satisfy the condition in Proposition B.2, if there exists $Z \subseteq \mathbf{V} \backslash \{X, Y, C\}$, such that $\mathrm{RKCIT}^{\beta(C^{perm})}(X, Y|Z)$ fail to reject the null hypothesis, then X and Y are not adjacent in G. Here C^{perm} denotes the shuffled C in dataset $\mathcal{D}$.*

Theorem 4.3. *With Assumption 4.1, the causal Markov assumption and faithfulness assumption, Algorithm 2 correctly recovers the underlying causal graph structure up to its Markov equivalence class.*

Algorithm 2 Rare Dependence PC (RD-PC)

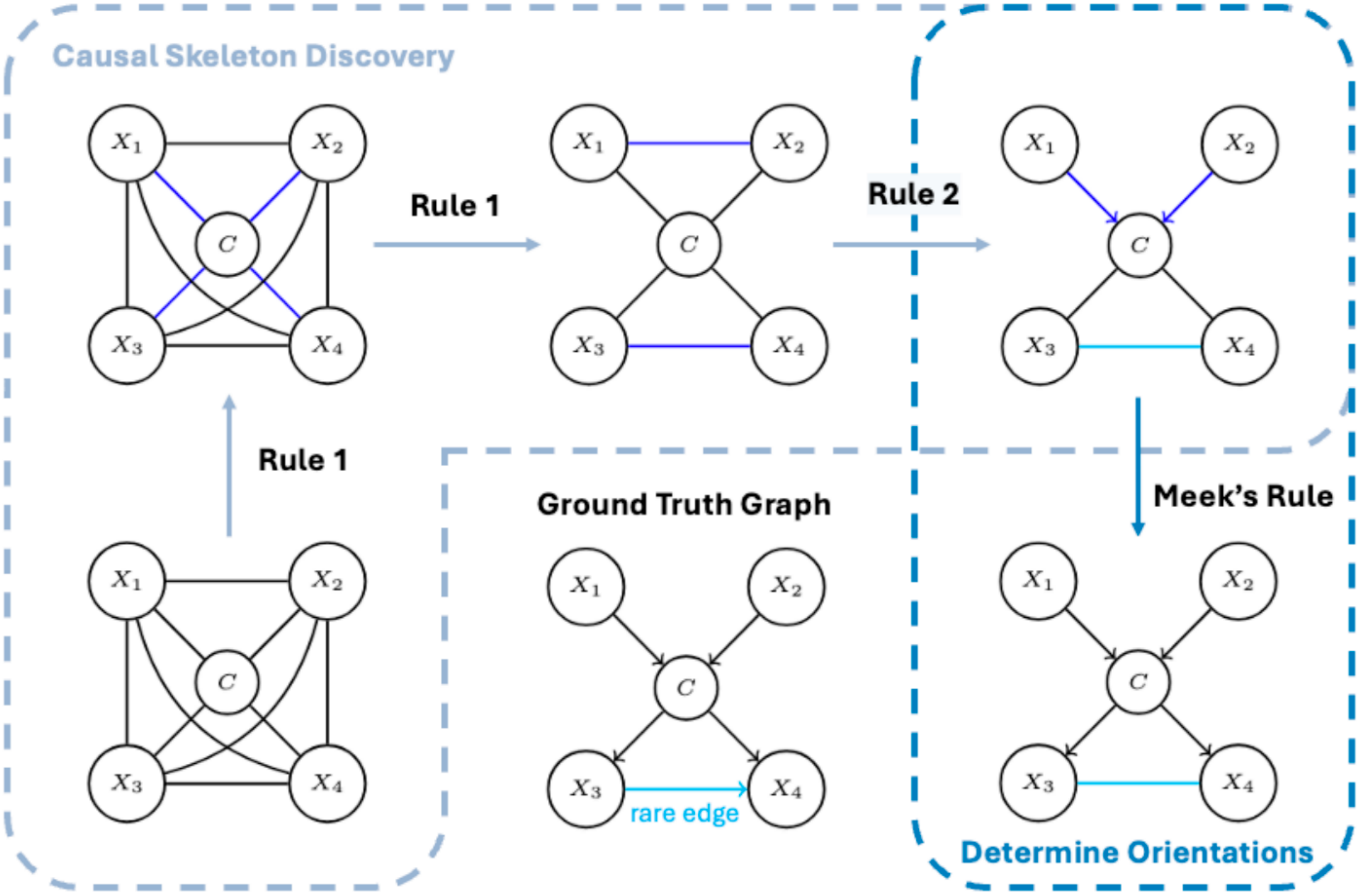
Input: $\mathcal{D}$: dataset. $\mathbf{V}$: node set. C: reference variable.
Output: causal graph G.
Stage 1: Causal skeleton discovery.
Initialize a complete undirected graph G on $\mathbf{V}$.
Remove the edge connected to C in G by **Rule 1**.
For $X, Y \in \mathbf{V} \setminus \{C\}$, remove the edge (X, Y) in G by **Rule 1**. If both X and Y are not adjacent to C, using KCIT only is enough.
Stage 2: Eliminating extraneous edges.
For $X, Y \in \mathbf{V} \backslash \{C\}$, if both X and Y are adjacent to C, check whether (X, Y) are the extraneous edge. Shuffle data of C in $\mathcal{D}$ as C^{perm}, if **Rule 2** is satisfied, remove the edge (X, Y), and orient $X \to C$ and $Y \to C$.
Stage 3: Determining the orientation.
Orient edges in G with the same orientation procedure as the PC algorithm (Meek, 1995).

Causal Discovery in the Presence of Rare Dependence

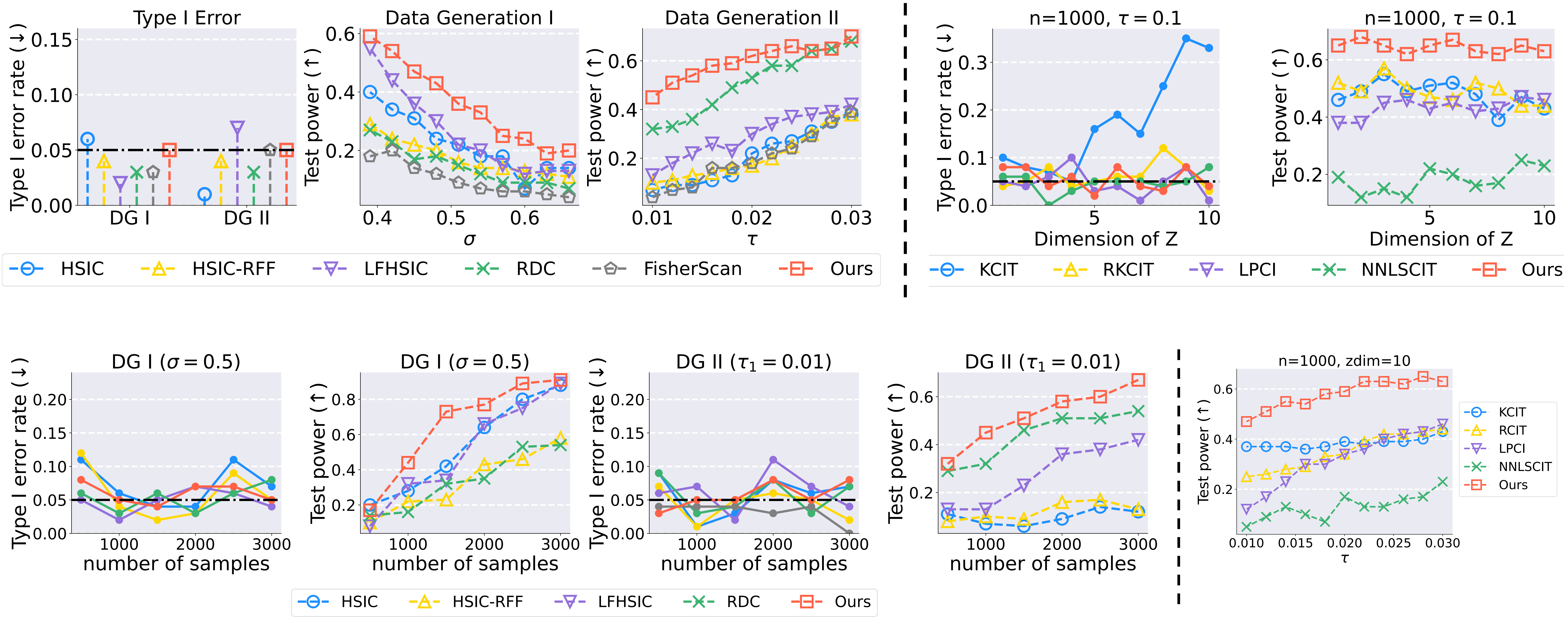
Algorithm

Rule 1. $\forall\ X, Y \in \mathbf{V}$, *if* $\exists Z \subseteq \mathbf{V}\backslash\{X, Y, C\}$ *s.t. both* $\mathrm{KCIT}(X, Y|Z)$ *and* $\mathrm{RKCIT}^{\beta(C)}(X, Y|Z)$ *fail to reject the null hypothesis, then* X *and* Y *are not adjacent in* G.

Rule 2. *For two variables* $X, Y \in \mathbf{V}$ *that satisfy the condition in Proposition B.2, if there exists* $Z \subseteq \mathbf{V}\backslash\{X, Y, C\}$, *such that* $\mathrm{RKCIT}^{\beta(C^{perm})}(X, Y|Z)$ *fail to reject the null hypothesis, then* X *and* Y *are not adjacent in* G. *Here* C^{perm} *denotes the shuffled* C *in dataset* $\mathcal{D}$.



Experimental Results



Conclusion and Future Work

- We portray the problem of rare dependence.
- We propose a novel testing method that combines kernel-based independence tests with adaptive sample importance reweighting.
- We also extend the idea to detect conditional rare independence. In addition, we integrate our reweighting CI tests into the PC algorithm for causal discovery in the presence of rare dependence.

- Extension: distribution & bound for CI statistics, RDPC with less assumptions
- Extension: without data splitting/ towards high-dimensional variable.