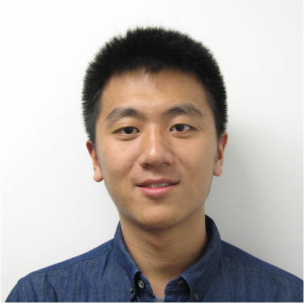
Maxway CRT: Improving the Robustness of Model-X Inference

Shuangning Li

Stanford University

Nov 2021

Joint work with

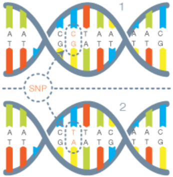


Molei Liu

Li & Liu (2021) "Maxway CRT: Improving the Robustness of Model-X Inference" *Draft in preparation*.

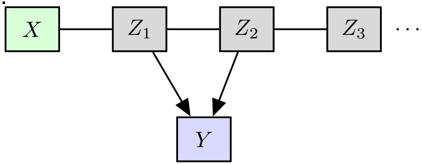
Motivation

- Imagine researchers are interested in whether a particular genetic variant influence a trait.



$\overset{?}{\Longrightarrow}$ risk of disease / body mass index

- Let X denote the genetic variant. Let Y be the trait.
- First idea: test for whether $X \perp\!\!\!\perp Y$. Not working because X can be correlated with other variables that influence Y.



Conditional Independence

- Idea: test for conditional independence!
- $\mathcal{H}_0 : X \perp\!\!\!\perp Y \mid Z$.
- Here, Y is the response variable of interest, X is an explanatory variable and Z are confounding variables (potentially high dimensional).
- There are n i.i.d. samples of (Y, X, Z) denoted as $(Y_i, X_i, Z_{i\cdot})$, and let $\boldsymbol{y} = (Y_1, Y_2, \ldots, Y_n)^\intercal \in \mathbb{R}^n$, $\boldsymbol{x} = (X_1, X_2, \ldots, X_n)^\intercal \in \mathbb{R}^n$, and $\boldsymbol{Z} = (Z_{1\cdot}, Z_{2\cdot}, \ldots, Z_{n\cdot})^\intercal \in \mathbb{R}^{n\times p}$.

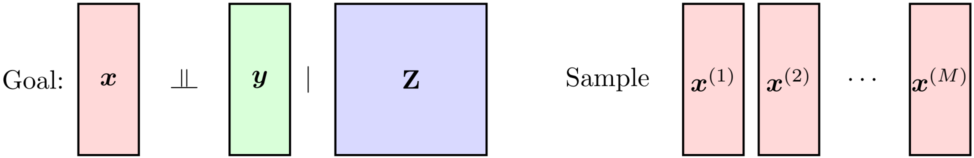
Conditional Randomization Test

- Introduced by Candès et al. (2018).

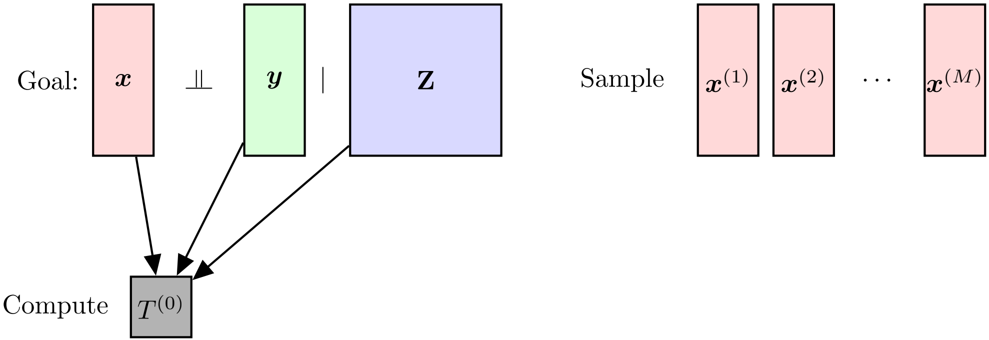
For $m \in \{1, \dots, M\}$
Sample $\boldsymbol{x}^{(m)}$ from the distribution of $\boldsymbol{x} | \boldsymbol{Z}$, independently of $\boldsymbol{x}$ and $\boldsymbol{y}$.
Output The p-value

$$p_{\mathrm{CRT}} = \frac{1}{M+1} \left(1 + \sum_{m=1}^{M} \mathbb{1} \left\{ T(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{Z}) \leq T(\boldsymbol{y}, \boldsymbol{x}^{(m)}, \boldsymbol{Z}) \right\} \right).$$

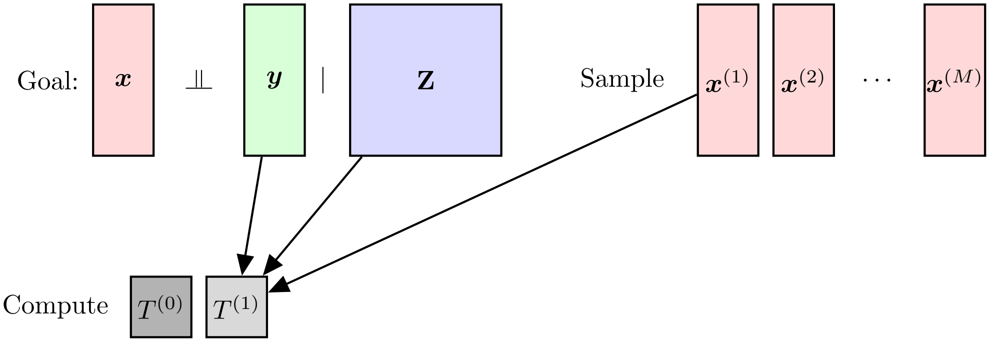
Conditional Randomization Test



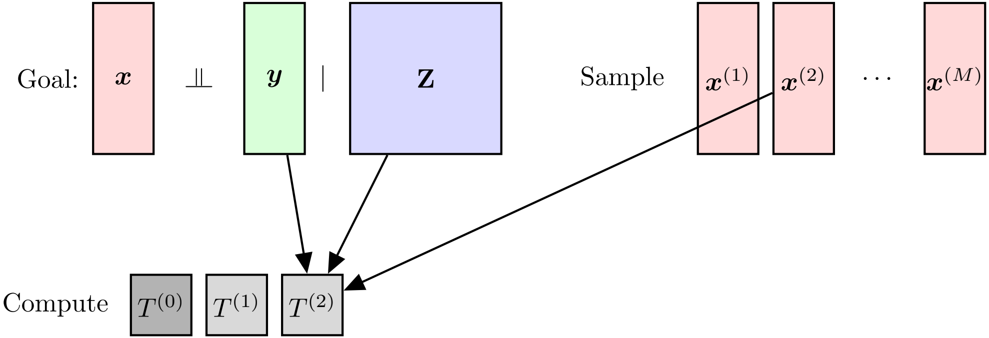
Conditional Randomization Test



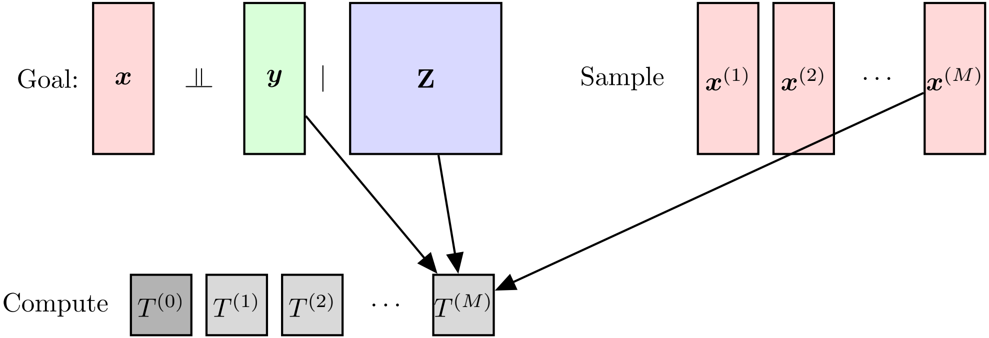
Conditional Randomization Test



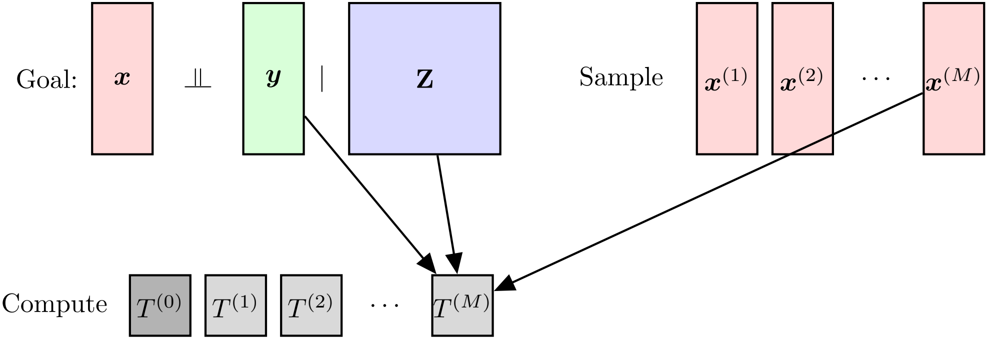
Conditional Randomization Test



Conditional Randomization Test



Conditional Randomization Test



$$p\text{-value } p_j = \frac{1}{M+1}\left(1 + \#\left\{ \boxed{T^{(m)}} : \boxed{T^{(m)}} \geq \boxed{T^{(0)}} \right\}\right)$$

Conditional Randomization Test

Theorem (Candès et al. (2018))

If $X \perp\!\!\!\perp Y \mid Z$, then the p-values from CRT satisfy $\mathbb{P}\left[p_j \leq \alpha\right] \leq \alpha$, for any $\alpha \in [0, 1]$. This holds regardless of the test statistic $T(\cdot)$.

- Requires perfect knowledge of the distribution of $\boldsymbol{x} \mid \boldsymbol{Z}$.
- Let $\rho^{\star n}$ be the distribution of $\boldsymbol{x} \mid \boldsymbol{Z}$. Assume in CRT, $\boldsymbol{x}^{(b)}$ are generated instead from ρ^n. Berrett et al. (2020) showed that

$$P\left(p_{\text{CRT}} \leq \alpha\right) \leq \alpha + \underbrace{d_{\text{TV}}(\rho^{\star n}, \rho^n)}_{\text{Model-X error}}.$$

The bound is tight when M is large.

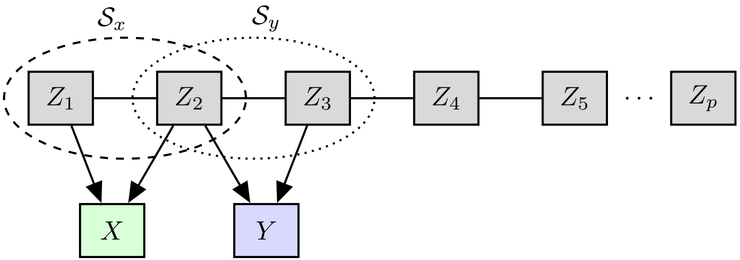
This work

- In addition to the knowledge of the distribution of $X \mid Z$, we also have knowledge of the distribution of $Y \mid Z$. Can we make use of this additional knowledge and make CRT more robust?
- We propose a Maxway (**M**odel and **A**djust **X** **W**ith the **A**ssistance of **Y**) CRT.
- Type-I error inflation of Model-X CRT is Δ_x.
  Type-I error inflation of Maxway CRT is $\Delta_x \Delta_y + \Delta_{x|g}$.
  Here Δ_x, Δ_y and $\Delta_{x|g}$ are the estimation errors for the distributions of $\boldsymbol{x} \mid \boldsymbol{Z}$, $\boldsymbol{y} \mid \boldsymbol{Z}$ and $\boldsymbol{x} \mid g(\boldsymbol{Z})$ respectively.
- "double robustness" in type-I error control.

A simple model

Suppose Z is a high dimensional random variable, but X and Y only depend on a small subset of the Z_j's. Assume that $Z = (Z_1, \ldots, Z_p)$, and

$$X = \phi(Z_1, Z_2) + \varepsilon, \quad Y = \psi(Z_2, Z_3) + \eta.$$



- To implement the original Model-X CRT, need to know $\mathcal{S}_x$ and the distribution of $X \mid Z_{\mathcal{S}_x}$.
- Assume for now that given a set $\mathcal{S}$ whose cardinality is not huge, we are able to learn the distribution of $X \mid Z_{\mathcal{S}}$ accurately.

A simple model

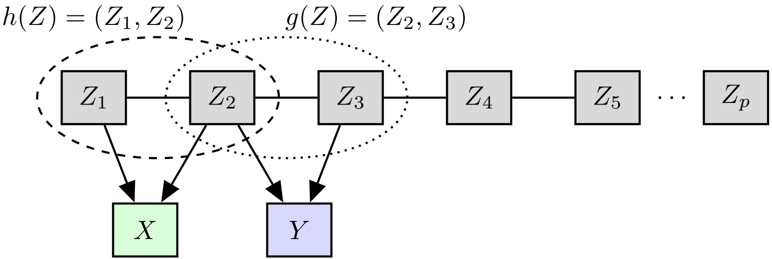
- A guess of the set: $\mathcal{S}$. Then (a special version of) the CRT becomes
  1. For $m = 1, 2, ..., M$:
     Sample $\boldsymbol{x}^{(m)}$ from the distribution of $\boldsymbol{x} \mid \boldsymbol{Z}_{\cdot\mathcal{S}}$ independently of $(\boldsymbol{x}, \boldsymbol{y})$.
  2. Output CRT p-value

$$p_{\text{CRT}} = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbf{1}\{T(\boldsymbol{y}, \boldsymbol{x}^{(m)}, \boldsymbol{Z}_{\cdot\mathcal{S}}) \geq T(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{Z}_{\cdot\mathcal{S}})\}\right).$$

- If $\mathcal{S}$ contains $\{1, 2\}$, the p-value is valid.
- If $\mathcal{S}$ contains $\{2, 3\}$, the p-value is valid as well. This is because $X \perp\!\!\!\perp Y \mid Z_{\mathcal{S}}$. The above procedure can be treated alternatively as a CRT for $(X, Y, Z_{\mathcal{S}})$.
- Some knowledge of how Y depends on Z can be useful in enhancing robustness in CRT. Without any information or prior knowledge on Y, to achieve validity, the best we can hope for is the set $\mathcal{S}$ to contain $\mathcal{S}_x$. Extra information on the distribution of Y **relaxes** the condition of the validity of the p-value.

Maxway CRT

- Assume that
$$X \overset{\text{approximately}}{\perp\!\!\!\perp} Z \mid h(Z) \quad , \quad Y \overset{\text{approximately}}{\perp\!\!\!\perp} Z \mid g(Z),$$
for some low dimensional functions of g and h.



Maxway CRT

- Assume that
$$X \overset{\text{approximately}}{\perp\!\!\!\perp} Z \mid h(Z) \quad , \quad Y \overset{\text{approximately}}{\perp\!\!\!\perp} Z \mid g(Z),$$
  for some low dimensional functions of g and h.
- Sparse linear model: $X = \alpha^\top Z_{\mathcal{S}_x} + \varepsilon$, where $\mathcal{S}_x \subset \{1, \ldots, p\}$. $h(Z) = Z_{\mathcal{S}_x}$ or $h(Z) = \alpha^\top Z_{\mathcal{S}_x}$.
- More general setting: the distribution of X given Z is very complicated.
  - Do a transformation: $R(X, Z) = F_X(X \mid Z)$. If X has a continuous distribution conditional on Z, then $R(X, Z) \sim \text{Unif}[0, 1]$ both marginally and conditionally on Z. Thus $R(X, Z) \perp\!\!\!\perp Z$. We can take h to be a null set.
  - Test whether $R(X, Z)$ is independent of Y conditional on Z.
  - Extracting the residual of X after removing the influence of Z.

Maxway CRT

Let $\rho^\star(\cdot \mid g(Z), h(Z))$ be the conditional distribution of X given $g(Z)$ and $h(Z)$, and let ρ be an estimate of $\rho^\star$.

Model and adjust X with the assistance of Y (Maxway) CRT

1. Sample $\boldsymbol{x}^{(m)}$ from the distribution of $\rho^n(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z}))$ independently of $(\boldsymbol{x}, \boldsymbol{y})$.
2. Output Maxway CRT p-value

$$p_{\text{maxway}} = \frac{1}{M+1}\left(1 + \sum_{m=1}^{M} \mathbf{1}\left\{T(\boldsymbol{x}^{(m)}, \boldsymbol{y}, g(\boldsymbol{Z}), h(\boldsymbol{Z})) \geq T(\boldsymbol{x}, \boldsymbol{y}, g(\boldsymbol{Z}), h(\boldsymbol{Z}))\right\}\right).$$

Exact Inference

Theorem

Suppose that either of the following conditions holds: (i) each $\boldsymbol{x}^{(m)}$ *is exchangeable with* $\boldsymbol{x}$ *given* $\boldsymbol{Z}$*; (ii)* $\boldsymbol{x}^{(m)}$ *is exchangeable with* $\boldsymbol{x}$ *given* $\{h(\boldsymbol{Z}), g(\boldsymbol{Z})\}$*, and* $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{y} \mid g(\boldsymbol{Z})$*. Then the Maxway CRT p-value defined is valid, i.e.,* $\mathbb{P}\left[p_{\text{maxway}} \leq \alpha\right] \leq \alpha$ *for any* $\alpha \in [0, 1]$ *under* $\mathcal{H}_0$*.*

- For (i), proof from standard CRT. $T(\boldsymbol{x}^{(m)}, \boldsymbol{y}, g(\boldsymbol{Z}), h(\boldsymbol{Z}))$ is exchangeable with $T(\boldsymbol{x}, \boldsymbol{y}, g(\boldsymbol{Z}), h(\boldsymbol{Z}))$.
- For (ii), $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid (g(\boldsymbol{Z}), h(\boldsymbol{Z})) \Rightarrow T(\boldsymbol{x}^{(m)}, \boldsymbol{y}, g(\boldsymbol{Z}), h(\boldsymbol{Z}))$ is exchangeable with $T(\boldsymbol{x}, \boldsymbol{y}, g(\boldsymbol{Z}), h(\boldsymbol{Z}))$.

Exact Inference

Theorem

Suppose that either of the following conditions holds: (i) each $\boldsymbol{x}^{(m)}$ *is exchangeable with* $\boldsymbol{x}$ *given* $\boldsymbol{Z}$*; (ii)* $\boldsymbol{x}^{(m)}$ *is exchangeable with* $\boldsymbol{x}$ *given* $\{h(\boldsymbol{Z}), g(\boldsymbol{Z})\}$*, and* $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{y} \mid g(\boldsymbol{Z})$*. Then the Maxway CRT p-value defined is valid, i.e.,* $\mathbb{P}[p_{\text{maxway}} \le \alpha] \le \alpha$ *for any* $\alpha \in [0,1]$ *under* $\mathcal{H}_0$*.*

- Compared to the conditions for the Model-X CRT p-value to be valid, the conditions stated in this theorem is strictly weaker.
- Condition (i) requires that the knowledge of the distribution of X given Z is perfect. Condition (ii) requires that $g(Z)$ contains all the information about Y that Z can possibly provide, and that the distribution of X given the low dimensional $g(Z)$ and $h(Z)$ is known.

Robustness of the Maxway CRT

- When g, h and ρ are not perfect anymore...

Theorem

For any $\alpha \in (0,1)$,

$$\mathbb{P}\left[p_{\text{maxway}} \leq \alpha\right] \leq \alpha + 2\mathbb{E}\left[d_x(\boldsymbol{Z})d_y(\boldsymbol{Z})\right] + \mathbb{E}\left[d_\rho(g(\boldsymbol{Z}), h(\boldsymbol{Z}))\right],$$

where

$d_\rho(g(\boldsymbol{Z}), h(\boldsymbol{Z})) = d_{\text{TV}}\left(\rho^{\star n}(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z})), \rho^n(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z}))\right)$,

$d_x(\boldsymbol{Z}) = d_{\text{TV}}\left(\rho^{\star n}(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z})), f_{\boldsymbol{x}|\boldsymbol{Z}}(\cdot \mid \boldsymbol{Z})\right)$, *and*

$d_y(\boldsymbol{Z}) = d_{\text{TV}}\left(f_{\boldsymbol{y}|\boldsymbol{Z}}(\cdot \mid \boldsymbol{Z}), f_{\boldsymbol{y}|g(\boldsymbol{Z})}(\cdot \mid g(\boldsymbol{Z}))\right)$.

Robustness of the Maxway CRT

Theorem

For any $\alpha \in (0,1)$,

$$\mathbb{P}\left[p_{\text{maxway}} \leq \alpha\right] \leq \alpha + 2\mathbb{E}\left[d_x(\boldsymbol{Z})\, d_y(\boldsymbol{Z})\right] + \mathbb{E}\left[d_\rho(g(\boldsymbol{Z}), h(\boldsymbol{Z}))\right],$$

where $d_x(\boldsymbol{Z}) = d_{\text{TV}}\left(\rho^{\star n}(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z})), f_{\boldsymbol{x}\mid\boldsymbol{Z}}(\cdot \mid \boldsymbol{Z})\right)$.

- Recall $\rho^{\star n}(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z})) = f_{x\mid h(Z), g(Z)}(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z}))$.
- Captures how independent X is to Z conditional on $h(Z)$.
- When $X \perp\!\!\!\perp Z \mid h(Z)$, then the conditional distribution of $X \mid Z$ would be the same as $X \mid h(Z)$, thus the term is zero.

Robustness of the Maxway CRT

Theorem

For any $\alpha \in (0,1)$,

$$\mathbb{P}\left[p_{\text{maxway}} \le \alpha\right] \le \alpha + 2\mathbb{E}\left[d_x(\boldsymbol{Z})\, d_y(\boldsymbol{Z})\right] + \mathbb{E}\left[d_\rho(g(\boldsymbol{Z}), h(\boldsymbol{Z}))\right],$$

where $d_y(\boldsymbol{Z}) = d_{\text{TV}}\left(f_{\boldsymbol{y}|\boldsymbol{Z}}(\cdot \mid \boldsymbol{Z}), f_{\boldsymbol{y}|g(\boldsymbol{Z})}(\cdot \mid g(\boldsymbol{Z}))\right)$.

- Captures how independent Y is to Z conditional on $g(Z)$.
- When $Y \perp\!\!\!\perp Z \mid g(Z)$, then the conditional distribution of $Y \mid Z$ would be the same as $Y \mid g(Z)$, thus the term is zero.

Robustness of the Maxway CRT

Theorem

For any $\alpha \in (0,1)$,

$$\mathbb{P}\left[p_{\text{maxway}} \le \alpha\right] \le \alpha + 2\mathbb{E}\left[d_x(\boldsymbol{Z})d_y(\boldsymbol{Z})\right] + \mathbb{E}\left[\, d_\rho(g(\boldsymbol{Z}), h(\boldsymbol{Z})) \,\right],$$

where $d_\rho(g(\boldsymbol{Z}), h(\boldsymbol{Z})) = d_{\text{TV}}\left(\rho^{\star n}(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z})), \rho^n(\cdot \mid g(\boldsymbol{Z}), h(\boldsymbol{Z}))\right)$.

- The d_ρ term is about the accuracy of ρ, i.e., how accurate we can estimate the distribution of X given low dimensional objects $h(Z)$ and $g(Z)$.

Compared to the Model-X CRT

- Model-X CRT

$$\mathbb{P}\left[p_{\mathrm{mx}} \leq \alpha\right] \leq \alpha + d_x' \approx \alpha + d_\rho + d_x.$$

- Maxway CRT

$$\mathbb{P}\left[p_{\mathrm{maxway}} \leq \alpha\right] \leq \alpha + d_\rho + 2 d_x d_y.$$

- For maxway CRT, the test statistic can only be a function of $\boldsymbol{x}, \boldsymbol{y}, g(\boldsymbol{Z}), h(\boldsymbol{Z})$. Not the most general form. But has a computational advantage (Liu et al., 2020), and it is typically powerful (Katsevich and Ramdas, 2020)

Example (Gaussian linear example, estimation)

$Y_i = Z_{i\cdot}^\top \alpha^\star + \varepsilon_i$, $X_i = Z_{i\cdot}^\top \beta^\star + \eta_i$. Take $g(Z)$ to be an estimate of the mean function $Z_{i\cdot}^\top \alpha^\star$. Take $h(Z)$ to be an estimate of the mean function $Z_{i\cdot}^\top \beta^\star$. Estimate the parameters with lasso.

$$\text{Type-I error inflation of the Maxway CRT} \lesssim \sqrt{\frac{n}{N_\rho}} + \sqrt{\frac{s_\alpha \log(p) n}{N_y}} \sqrt{\frac{s_\beta \log(p) n}{N_x}}.$$

$$\text{Type-I error inflation of the Model-X CRT} \lesssim \sqrt{\frac{s_\beta \log(p) n}{N_x}}.$$

Examples

Example (Gaussian linear example, variable selection)

$Y_i = Z_{i\cdot}^\top \alpha^\star + \varepsilon_i$, $X_i = Z_{i\cdot}^\top \beta^\star + \eta_i$. Take $g(Z)$ to be an estimate of $Z_{\mathcal{S}^\star}$, where $\mathcal{S}^\star$ is the support of $\alpha^\star$. Take $h(Z)$ to be an estimate of the mean function $Z_{i\cdot}^\top \beta^\star$. Estimate the parameters/support with lasso.

$$\text{Type-I error inflation of the Maxway CRT} \lesssim \sqrt{\frac{ns_\alpha}{N_\rho}} + \delta\sqrt{\frac{s_\beta \log(p) n}{N_x}},$$

where $1 - \delta$ is the probability of exact recovery of the support of $\alpha^\star$.

$$\text{Type-I error inflation of the Model-X CRT} \lesssim \sqrt{\frac{s_\beta \log(p) n}{N_x}}.$$

Examples

Example (Binary X, smooth mean functions)

$Y_i = g^\star(Z_{i\cdot}) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma_y^2)$ independently and $X_i \sim \mathrm{Bern}(h^\star(Z_{i\cdot}))$ independently. In the Maxway CRT, we take $g(Z_{i\cdot})$ to be an estimate of $g^\star(Z_{i\cdot})$ and take $h(Z_{i\cdot})$ to be an estimate of $h^\star(Z_{i\cdot})$. Assume $g^\star$ is α-smooth, $h^\star$ is β-smooth. Estimate the mean functions with a kernel method.

$$\text{Type-I error inflation of the Maxway CRT} \lesssim \sqrt{n} N_\rho^{-\frac{\gamma}{2\gamma+2}} + n N_y^{-\frac{\alpha}{2\alpha+p}} N_x^{-\frac{\beta}{2\beta+p}}.$$

$$\text{Type-I error inflation of the Model-X CRT} \lesssim \sqrt{n} N_x^{-\frac{\beta}{2\beta+p}}.$$

Semi-supervised Scenario

- Labelled data $\boldsymbol{D} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{Z})$ of sample size n
- Unlabelled data $\boldsymbol{D}^u = (\boldsymbol{x}^u, \boldsymbol{Z}^u) \rightarrow$ train h and ρ
- How to train g? What if we don't have a external dataset of $(\boldsymbol{y}, \boldsymbol{Z})$?
  - Train g on $\boldsymbol{D}$.
  - Theory cannot be directly applied. Still hope to avoid overfitting.
  - "Cross-fitting". Divide the data $\boldsymbol{D}$ into K fold. Train g on the $K-1$ folds and evaluate $g(Z_i)$ on the other fold.

Simulations

Gaussian Linear Model

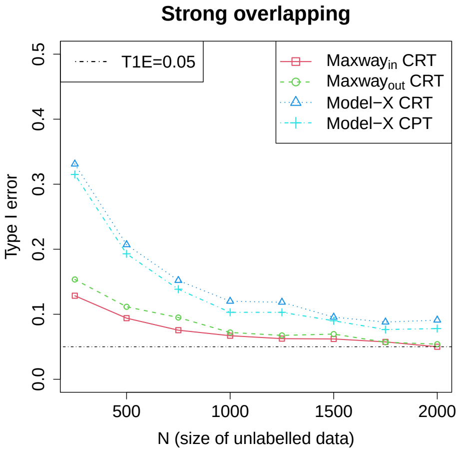
Generate $Z \in \mathbb{R}^p$ from $\mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$ where $p = 500$ and $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. Then generate the conditional gaussian X and Y following:

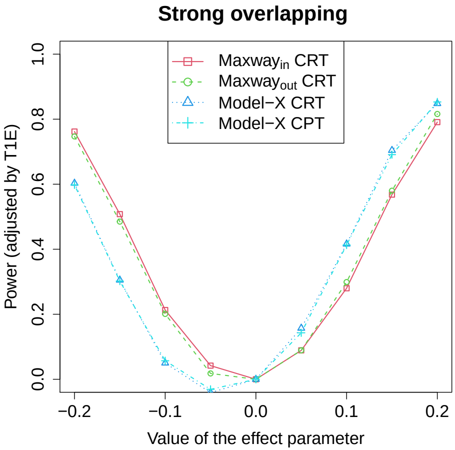
$$X = 0.3\sum_{j=1}^{5} \nu_j Z_j + \eta \sum_{\ell \in \mathcal{I}_1} \nu_\ell Z_\ell + \epsilon_1; \quad Y = \gamma h(X, Z) + 0.3\sum_{j=1}^{5} \nu_j Z_j + \eta \sum_{\ell \in \mathcal{I}_2} \nu_\ell Z_\ell + \epsilon_2,$$

where $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, 1)$, each ν_j is randomly picked from $\{-1, 1\}$, and $\mathcal{I}_1, \mathcal{I}_2$ are two disjoint sets of indices randomly drawn from $\{6, 7, \ldots, p\}$ satisfying $|\mathcal{I}_1| = |\mathcal{I}_2| = 25$.

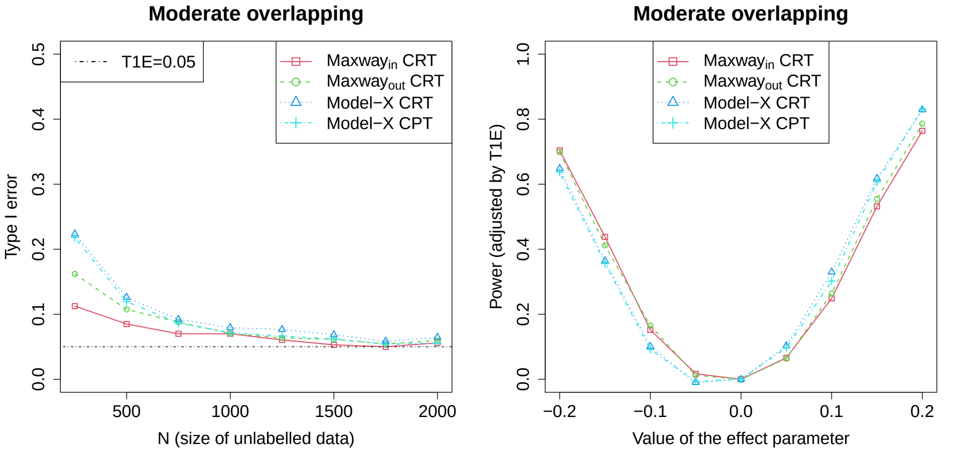
- η: how strong is the confounding? $\eta = 0, 0.1, 0.2$: strong, moderate, weak overlapping/confounding.
- $\gamma = 0$: evaluate type I error.
- $\gamma \neq 0$: power curve.

Simulations

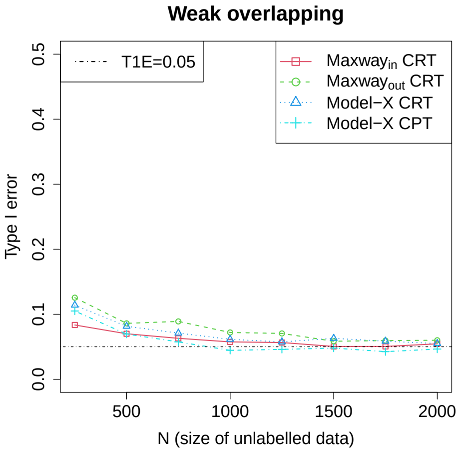


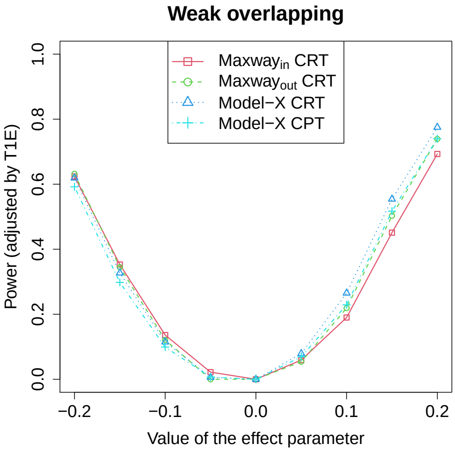


Simulations

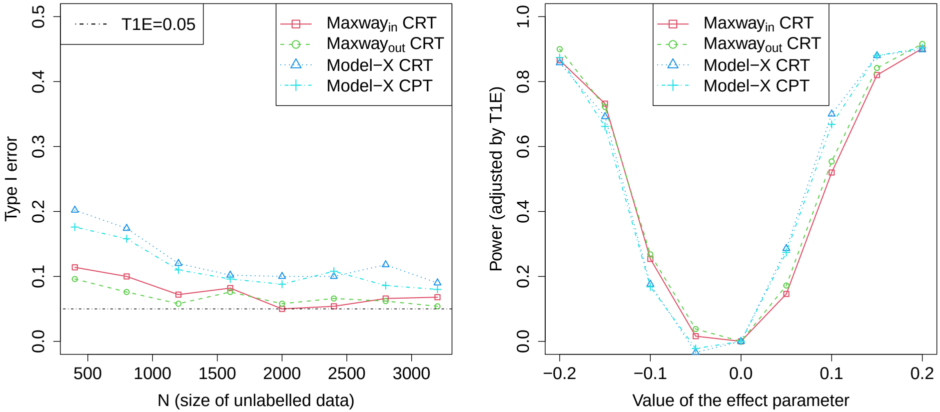


Simulations





Simulations: nonlinear model



Real Data Example

- Statins are one of the most commonly used drug in the United States for lowering the level of low–density lipoprotein (LDL) and the risk of cardiovascular disease (CVD).
- Working mechanism: HMGCR inhibition
- Evidences showing that the use of statins could increase the risk for type II diabetes mellitus (DM).

$\overset{?}{\Longrightarrow}$

- Unethical/expensive to conduct randomized control trial.
- Statins = absense of certain SNP in HMGCR.
- Test whether SNP $\perp\!\!\!\perp$ diabetes $\mid$ other variables—gives a biological perspective

Real Data Example

- UK Biobank
- Z includes age, gender and genetic variants associated with DM or its related phenotypes including high LDL, high–density lipoprotein (HDL) and BMI.
- p-values

Statistic	CRT	CPT	Maxway CRT
d_0	0.06	0.06	0.04
d_I	0.16	0.18	0.13

Table: The d_0 and d_I p-values for the dependence of the risk of DM on the treatment of statins functionally represented by the variant rs17238484-G.

- The Maxway CRT is not generally more conservative than the original Model-X CRT.

Thank you!