Double robustness of a model-X conditional independence test

Lawrence David Brown Student Workshop
March 22, 2023

Ziang Niu

Reference

Niu*, Chakraborty*, Dukes, & Katsevich. Reconciling model-X and doubly robust approaches to conditional independence testing. *Annals of Statistics,* 2024.

Abhinav Chakraborty

Oliver Dukes

Eugene Katsevich

Conditional independence testing

Conditional independence testing

Statistical task: Test whether a response variable $\mathbf{Y} \in \mathbb{R}$ is associated with a predictor variable $\mathbf{X} \in \mathbb{R}$ when controlling for covariates $\mathbf{Z} \in \mathbb{R}^p$, given n i.i.d. samples (X_i, Y_i, Z_i) from a joint distribution $\mathscr{L}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

Conditional independence testing

Statistical task: Test whether a response variable $\mathbf{Y} \in \mathbb{R}$ is associated with a predictor variable $\mathbf{X} \in \mathbb{R}$ when controlling for covariates $\mathbf{Z} \in \mathbb{R}^p$, given n i.i.d. samples (X_i, Y_i, Z_i) from a joint distribution $\mathscr{L}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

Hypothesis formulation: In the joint distribution $\mathscr{L}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, test the null hypothesis of conditional independence (CI):

Conditional independence testing

Statistical task: Test whether a response variable $\mathbf{Y} \in \mathbb{R}$ is associated with a predictor variable $\mathbf{X} \in \mathbb{R}$ when controlling for covariates $\mathbf{Z} \in \mathbb{R}^p$, given n i.i.d. samples (X_i, Y_i, Z_i) from a joint distribution $\mathscr{L}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

Hypothesis formulation: In the joint distribution $\mathscr{L}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, test the null hypothesis of conditional independence (CI):

$$H_0^{\text{CI}} : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}\,.$$

Conditional independence testing

Statistical task: Test whether a response variable $\mathbf{Y} \in \mathbb{R}$ is associated with a predictor variable $\mathbf{X} \in \mathbb{R}$ when controlling for covariates $\mathbf{Z} \in \mathbb{R}^p$, given n i.i.d. samples (X_i, Y_i, Z_i) from a joint distribution $\mathscr{L}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$.

Hypothesis formulation: In the joint distribution $\mathscr{L}_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, test the null hypothesis of conditional independence (CI):

$$H_0^{\text{CI}} : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}\,.$$

This turns out to be a very challenging problem!

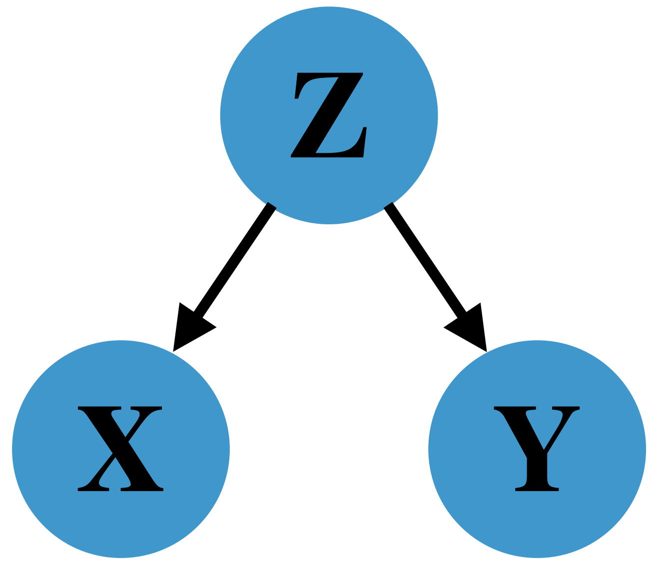
CI testing requires assumptions

If $\mathbf{Z}$ is continuous, any test with Type-I error control over the entire CI null $H_0^{\mathrm{CI}} : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ cannot have nontrivial power against any alternative. (Shah and Peters '20)

CI testing requires assumptions

If $\mathbf{Z}$ is continuous, any test with Type-I error control over the entire CI null $H_0^{\text{CI}} : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ cannot have nontrivial power against any alternative. (Shah and Peters '20)

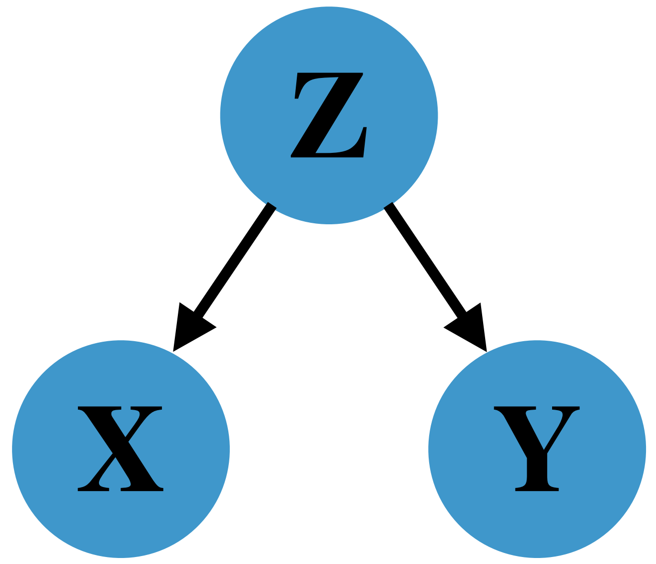
A test with Type-I error control must protect against too many sneaky ways $\mathbf{Z}$ can affect both $\mathbf{X}$ and $\mathbf{Y}$.



CI testing requires assumptions

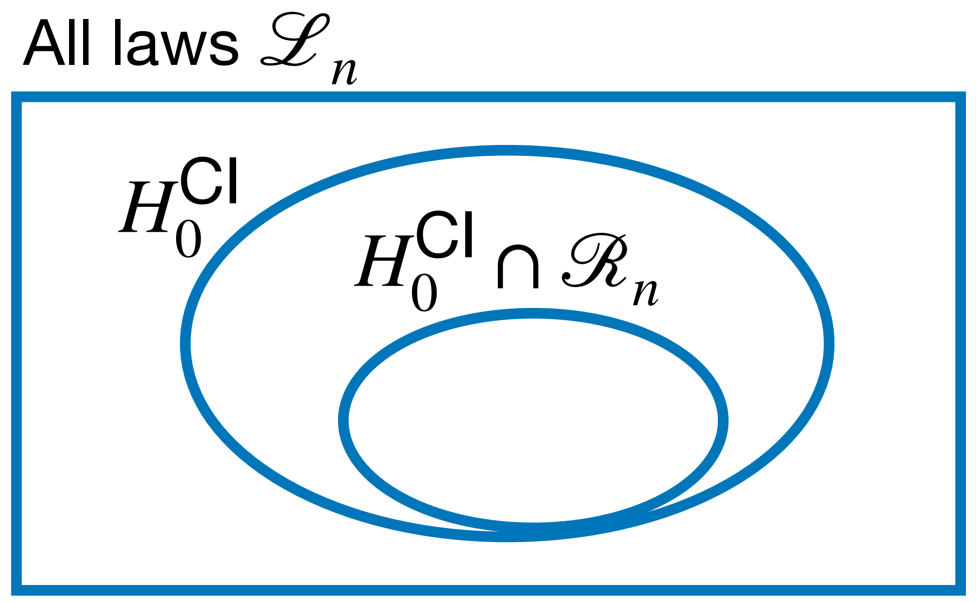
If $\mathbf{Z}$ is continuous, any test with Type-I error control over the entire CI null $H_0^{\mathrm{CI}} : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ cannot have nontrivial power against any alternative. (Shah and Peters '20)

A test with Type-I error control must protect against too many sneaky ways $\mathbf{Z}$ can affect both $\mathbf{X}$ and $\mathbf{Y}$.



Given a set of regularity conditions $\mathscr{R}_n$ on $\mathscr{L}_n$, one can only hope to control Type-I error over the smaller null hypothesis

$$H_0 : H_0^{\mathrm{CI}} \cap \mathscr{R}_n.$$



The model-X (MX) assumption

(Candès et al '18)

Assume we know the conditional distribution $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ exactly, i.e.

$$\mathscr{R}_n \equiv \{\mathscr{L}_n : \mathscr{L}_n(\mathbf{X} \mid \mathbf{Z}) = \mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})\},$$

where $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ is the given conditional distribution.

The model-X (MX) assumption

(Candès et al ’18)

Assume we know the conditional distribution $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ exactly, i.e.

$$\mathscr{R}_n \equiv \{\mathscr{L}_n : \mathscr{L}_n(\mathbf{X} \mid \mathbf{Z}) = \mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})\},$$

where $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ is the given conditional distribution.

Reasonable if conditional distribution $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ controlled by experimenter.[1]

[1]Ham et al ’22, Aufiero and Janson ’22

The model-X (MX) assumption

(Candès et al ’18)

Assume we know the conditional distribution $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ exactly, i.e.

$$\mathscr{R}_n \equiv \{\mathscr{L}_n : \mathscr{L}_n(\mathbf{X} \mid \mathbf{Z}) = \mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})\},$$

where $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ is the given conditional distribution.

Reasonable if conditional distribution $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ controlled by experimenter.[1]

Powerful CI tests available under the MX assumption: conditional randomization test (CRT) for single testing and MX knockoffs for multiple testing.[2]

[1]Ham et al ’22, Aufiero and Janson ’22

[2]Candes et al ’18

The conditional randomization test (CRT)

The conditional randomization test (CRT)

Conditional randomization test

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathscr{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathscr{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
   - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathscr{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
   - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;
3. Compute threshold
   $C_n \equiv \mathbb{Q}_{1-\alpha}[\{T_n, \tilde{T}_n^{(1)}, \ldots, \tilde{T}_n^{(B)}\}]$;

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
    - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathcal{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
    - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;
3. Compute threshold $C_n \equiv \mathbb{Q}_{1-\alpha}[\{T_n, \tilde{T}_n^{(1)}, \ldots, \tilde{T}_n^{(B)}\}]$;
4. Reject if $T_n > C_n$.

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathscr{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
   - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;
3. Compute threshold $C_n \equiv \mathbb{Q}_{1-\alpha}[\{T_n, \tilde{T}_n^{(1)}, \ldots, \tilde{T}_n^{(B)}\}]$;
4. Reject if $T_n > C_n$.

Properties

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathscr{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
   - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;
3. Compute threshold $C_n \equiv \mathbb{Q}_{1-\alpha}[\{T_n, \tilde{T}_n^{(1)}, \ldots, \tilde{T}_n^{(B)}\}]$;
4. Reject if $T_n > C_n$.

Properties

- Finite-sample Type-I error control

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathscr{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
   - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;
3. Compute threshold $C_n \equiv \mathbb{Q}_{1-\alpha}[\{T_n, \tilde{T}_n^{(1)}, \ldots, \tilde{T}_n^{(B)}\}]$;
4. Reject if $T_n > C_n$.

Properties

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathscr{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
   - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;
3. Compute threshold $C_n \equiv \mathbb{Q}_{1-\alpha}[\{T_n, \tilde{T}_n^{(1)}, \ldots, \tilde{T}_n^{(B)}\}]$;
4. Reject if $T_n > C_n$.

Properties

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

The conditional randomization test (CRT)

Conditional randomization test

1. Compute test stat $T_n \equiv T_n(X, Y, Z)$;
2. For $b = 1, \ldots, B$,
   - Draw $\tilde{X}_i^{(b)} \overset{\text{ind}}{\sim} \mathcal{L}_n^*(X_i \mid \mathbf{Z} = Z_i)$;
   - Recompute $\tilde{T}_n^{(b)} \equiv T_n(\tilde{X}^{(b)}, Y, Z)$;
3. Compute threshold $C_n \equiv \mathbb{Q}_{1-\alpha}[\{T_n, \tilde{T}_n^{(1)}, \ldots, \tilde{T}_n^{(B)}\}]$;
4. Reject if $T_n > C_n$.

Properties

- Finite-sample Type-I error control
- No assumptions on $\mathcal{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

Remark:

Test statistic choice impacts power;[1] often employs penalized regression or black-box machine learning.

[1]K. and Ramdas '22

Challenge: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ usually an approximation

Challenge: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ usually an approximation

Aside from controlled experiments, the MX assumption is typically too strong.

Challenge: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ usually an approximation

Aside from controlled experiments, the MX assumption is typically too strong.

MX motivated by genome-wide association studies, where plausible parametric model is available for $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$; parameters must still be learned from data.

Challenge: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ usually an approximation

Aside from controlled experiments, the MX assumption is typically too strong.

MX motivated by genome-wide association studies, where plausible parametric model is available for $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$; parameters must still be learned from data.

MX methods deployed by learning $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z}) \equiv \widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ from data:

Challenge: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ usually an approximation

Aside from controlled experiments, the MX assumption is typically too strong.

MX motivated by genome-wide association studies, where plausible parametric model is available for $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$; parameters must still be learned from data.

MX methods deployed by learning $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z}) \equiv \widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ from data:

1. Either out of sample, based on extra unlabeled pairs (X_i, Z_i),

Challenge: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ usually an approximation

Aside from controlled experiments, the MX assumption is typically too strong.

MX motivated by genome-wide association studies, where plausible parametric model is available for $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$; parameters must still be learned from data.

MX methods deployed by learning $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z}) \equiv \widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ from data:

1. Either out of sample, based on extra unlabeled pairs (X_i, Z_i),
2. Or in sample, based on the same data used for testing (more common).

Case 1: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on large auxiliary dataset

Case 1: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on large auxiliary dataset

Berrett et al '20:

If $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ obtained from well-specified OLS based on N auxiliary samples, then

$$\mathbb{P}[\text{false rejection}] \leq \alpha + O_p\left(\sqrt{\frac{n \cdot \text{dim}(\mathbf{Z})}{N}}\right)$$

Case 1: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on large auxiliary dataset

Berrett et al '20:

If $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ obtained from well-specified OLS based on N auxiliary samples, then

$$\mathbb{P}[\text{false rejection}] \leq \alpha + O_p\left(\sqrt{\frac{n \cdot \text{dim}(\mathbf{Z})}{N}}\right)$$

CRT properties if $N \gg n \cdot \text{dim}(\mathbf{Z})$

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

Case 1: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on large auxiliary dataset

Berrett et al ’20:

If $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ obtained from well-specified OLS based on N auxiliary samples, then

$$\mathbb{P}[\text{false rejection}] \leq \alpha + O_p\left(\sqrt{\frac{n \cdot \text{dim}(\mathbf{Z})}{N}}\right)$$

CRT properties if $N \gg n \cdot \text{dim}(\mathbf{Z})$

- ~~Finite-sample~~ Asymptotic Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

Case 2: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

Case 2: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

MX method applied as if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ were known (more common in practice).

Case 2: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

MX method applied as if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ were known (more common in practice).

- **Theory:** For worst-case test statistics, no hope for Type-I error control.[1] Beyond that, few existing results.

[1]Berrett et al '20

Case 2: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

MX method applied as if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ were known (more common in practice).

- **Theory:** For worst-case test statistics, no hope for Type-I error control.[1] Beyond that, few existing results.

- **Simulations:** MX methods robust to in-sample learning of $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$.[2]

[1]Berrett et al ’20
[2]Candes et al ’18, Liu et al ’22

Case 2: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

MX method applied as if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ were known (more common in practice).

- **Theory:** For worst-case test statistics, no hope for Type-I error control.[1] Beyond that, few existing results.
- **Simulations:** MX methods robust to in-sample learning of $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$.[2]

Open question:

[1]Berrett et al ’20
[2]Candes et al ’18, Liu et al ’22

Case 2: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

MX method applied as if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ were known (more common in practice).

- **Theory:** For worst-case test statistics, no hope for Type-I error control.[1] Beyond that, few existing results.
- **Simulations:** MX methods robust to in-sample learning of $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$.[2]

Open question:

How robust are MX methods when $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample?

[1]Berrett et al ’20
[2]Candes et al ’18, Liu et al ’22

Case 2: $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

MX method applied as if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ were known (more common in practice).

- **Theory:** For worst-case test statistics, no hope for Type-I error control.[1] Beyond that, few existing results.
- **Simulations:** MX methods robust to in-sample learning of $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$.[2]

Open question:

> How robust are MX methods when $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample?

We study this question in the context of a specific MX method: the dCRT.

[1]Berrett et al ’20
[2]Candes et al ’18, Liu et al ’22

The distilled CRT and its in-sample approximation

The distilled CRT and its in-sample approximation

Let $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ and $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$.

The distilled CRT and its in-sample approximation

Let $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ and $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$.

The dCRT[1] is an instance of the CRT, with

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)),$$

where $\hat{\mu}_{n,y}(\mathbf{Z})$ is obtained via in-sample machine learning of Y on Z and $\mu_{n,x}(\mathbf{Z})$ is known by the MX assumption.

[1]Liu et al ’22

The distilled CRT and its in-sample approximation

Let $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ and $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$.

The dCRT[1] is an instance of the CRT, with

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)),$$

where $\hat{\mu}_{n,y}(\mathbf{Z})$ is obtained via in-sample machine learning of Y on Z and $\mu_{n,x}(\mathbf{Z})$ is known by the MX assumption.

The approximate dCRT with $\widehat{\mathcal{L}}_n(\mathbf{X} \mid \mathbf{Z})$ learned in sample is the same, except

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)).$$

[1]Liu et al ’22

The distilled CRT and its in-sample approximation

Let $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ and $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$.

The dCRT[1] is an instance of the CRT, with

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)),$$

where $\hat{\mu}_{n,y}(\mathbf{Z})$ is obtained via in-sample machine learning of Y on Z and $\mu_{n,x}(\mathbf{Z})$ is known by the MX assumption.

The approximate dCRT with $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ learned in sample is the same, except

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)).$$

Resampling distribution changed to $\hat{\mathscr{L}}(X_i \mid Z_i)$

[1]Liu et al '22

The distilled CRT and its in-sample approximation

Let $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ and $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$.

The dCRT[1] is an instance of the CRT, with

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)),$$

where $\hat{\mu}_{n,y}(\mathbf{Z})$ is obtained via in-sample machine learning of Y on Z and $\mu_{n,x}(\mathbf{Z})$ is known by the MX assumption.

The approximate dCRT with $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ learned in sample is the same, except

$$T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i)).$$

methodology under examination, henceforth just "dCRT"

Resampling distribution changed to $\hat{\mathscr{L}}(X_i \mid Z_i)$

[1]Liu et al '22

Is dCRT robust to in-sample learning?

Intuition from simulations

Intuition from simulations

Simple numerical simulation:

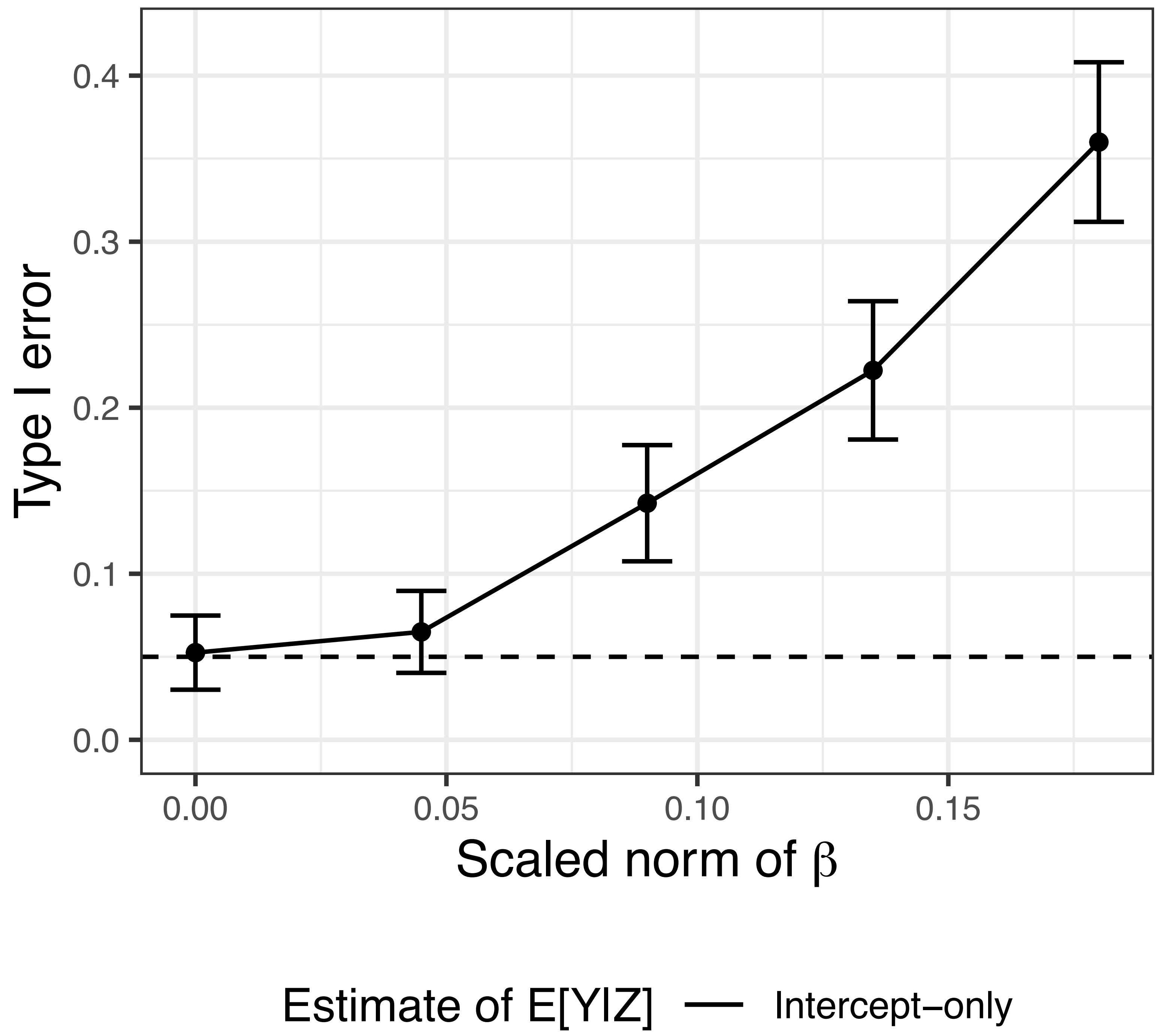
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Intuition from simulations

Simple numerical simulation:

- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\widehat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.



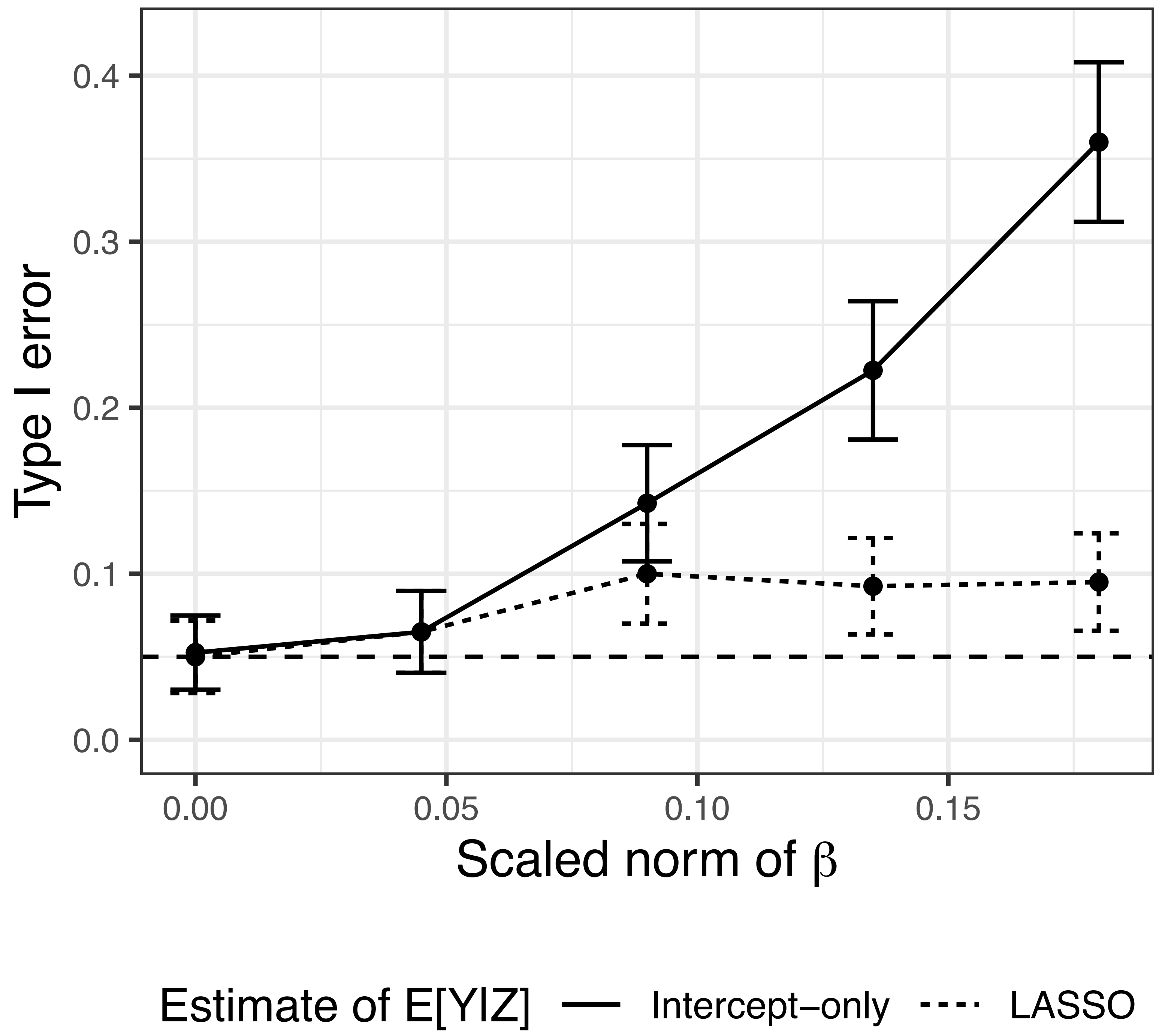
Intuition from simulations

Simple numerical simulation:

- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\widehat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Case 2: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated decently:
$\widehat{\mu}_{n,y}(\mathbf{Z})$ obtained via lasso of Y on Z.



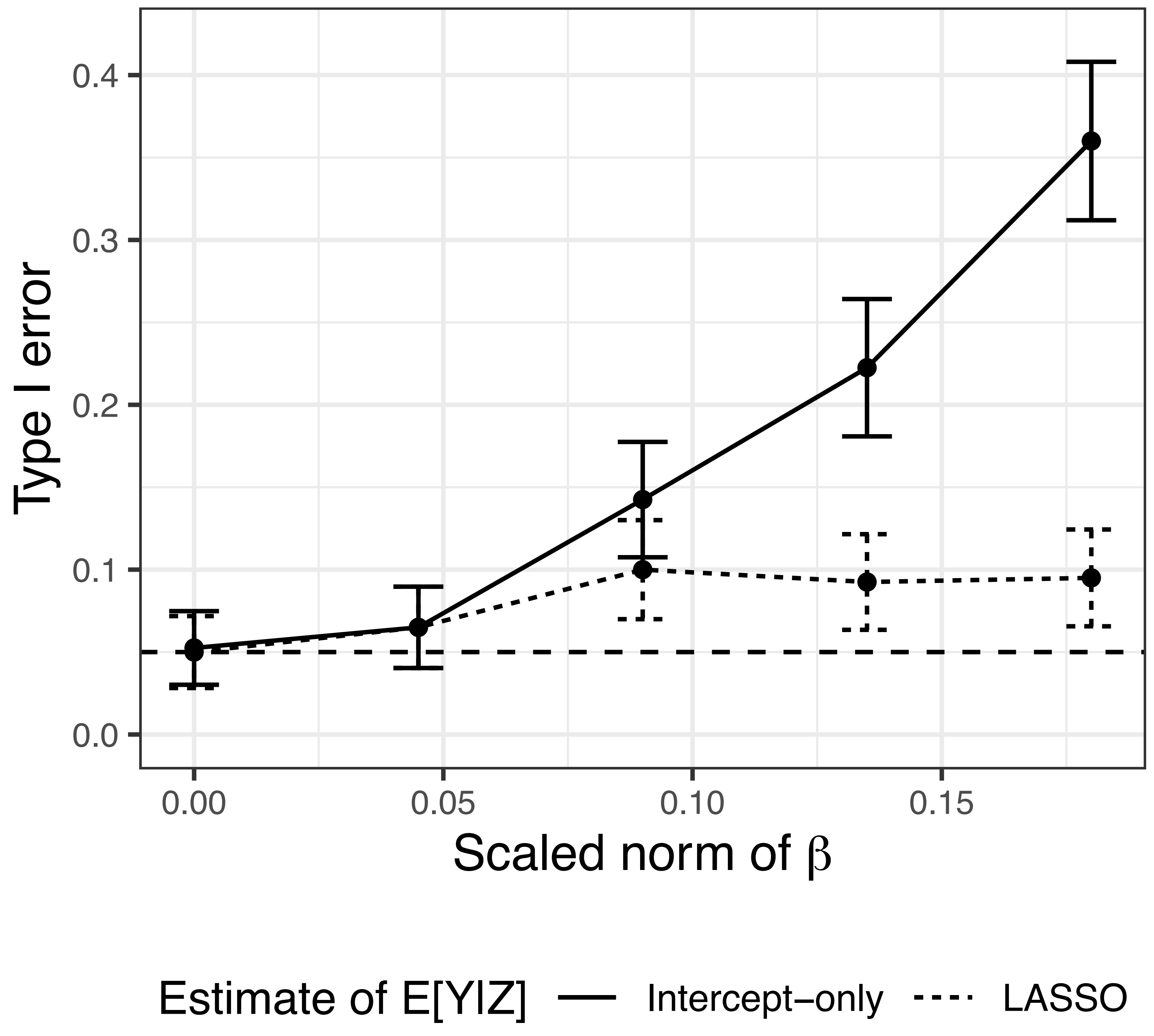
Intuition from simulations

Simple numerical simulation:

- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\hat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Case 2: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated decently:
$\hat{\mu}_{n,y}(\mathbf{Z})$ obtained via lasso of Y on Z.



No hope for good inference with poor estimate for $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$.

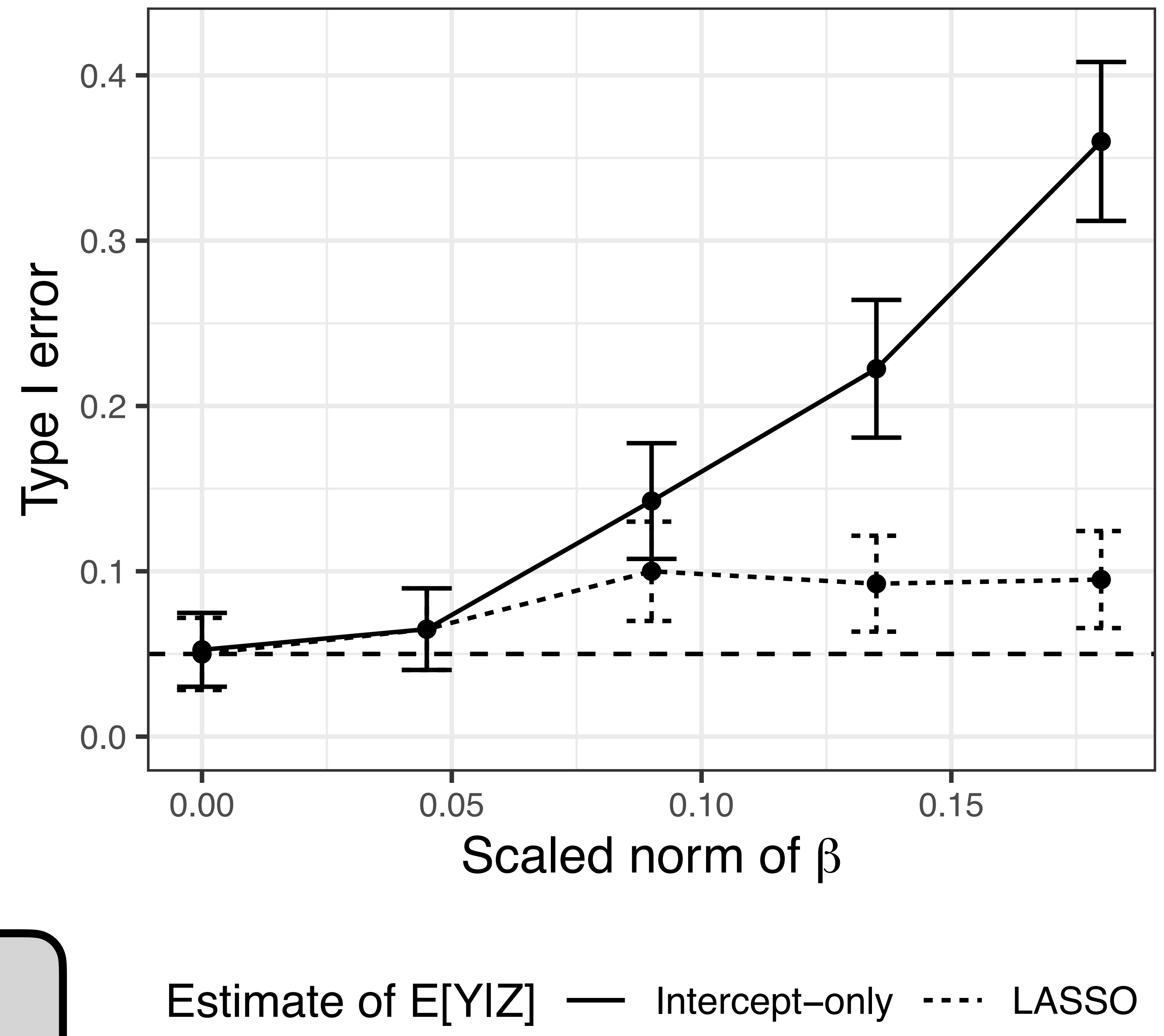
Intuition from simulations

Simple numerical simulation:

- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\widehat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Case 2: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated decently:
$\widehat{\mu}_{n,y}(\mathbf{Z})$ obtained via lasso of Y on Z.



No hope for good inference with poor estimate for $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$.

Better estimate of $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ improves robustness of dCRT.

Doubly robust conditional independence tests

Doubly robust conditional independence tests

We found that for dCRT, better estimation of $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z})$ compensates for errors in the estimation of $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$. This is a double robustness phenomenon!

Doubly robust conditional independence tests

We found that for dCRT, better estimation of $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z})$ compensates for errors in the estimation of $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$. This is a double robustness phenomenon!

We claim that the dCRT itself is doubly robust!

Doubly robust conditional independence tests

We found that for dCRT, better estimation of $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z})$ compensates for errors in the estimation of $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$. This is a double robustness phenomenon!

We claim that the dCRT itself is doubly robust!

In fact, we claim that the dCRT is asymptotically equivalent to the doubly robust generalized covariance measure (GCM) test (Shah and Peters, 2020).

The GCM test
(Shah and Peters ’20)

The GCM test

(Shah and Peters ’20)

1. Fit an approximation $\hat{\mu}_{n,x}(\mathbf{Z})$ of $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ via machine learning;

The GCM test
(Shah and Peters '20)

1. Fit an approximation $\hat{\mu}_{n,x}(\mathbf{Z})$ of $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ via machine learning;
2. Fit an approximation $\hat{\mu}_{n,y}(\mathbf{Z})$ of $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$ via machine learning;

The GCM test

(Shah and Peters ’20)

1. Fit an approximation $\hat{\mu}_{n,x}(\mathbf{Z})$ of $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ via machine learning;
2. Fit an approximation $\hat{\mu}_{n,y}(\mathbf{Z})$ of $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$ via machine learning;
3. Compute $T_n(X, Y, Z) \equiv \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))$

The GCM test

(Shah and Peters '20)

1. Fit an approximation $\hat{\mu}_{n,x}(\mathbf{Z})$ of $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ via machine learning;
2. Fit an approximation $\hat{\mu}_{n,y}(\mathbf{Z})$ of $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$ via machine learning;
3. Compute $T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))$
4. Compute $(S_n^{\text{GCM}})^2(X, Y, Z) \equiv \mathbb{V}\{(X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))\}$

The GCM test
(Shah and Peters '20)

1. Fit an approximation $\hat{\mu}_{n,x}(\mathbf{Z})$ of $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ via machine learning;
2. Fit an approximation $\hat{\mu}_{n,y}(\mathbf{Z})$ of $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$ via machine learning;
3. Compute $T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))$
4. Compute $(S_n^{\text{GCM}})^2(X, Y, Z) \equiv \mathbb{V}\{(X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))\}$ ← Sample variance

The GCM test
(Shah and Peters ’20)

1. Fit an approximation $\hat{\mu}_{n,x}(\mathbf{Z})$ of $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ via machine learning;
2. Fit an approximation $\hat{\mu}_{n,y}(\mathbf{Z})$ of $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$ via machine learning;
3. Compute $T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))$
4. Compute $(S_n^{\text{GCM}})^2(X, Y, Z) \equiv \mathbb{V}\{(X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))\}$ ← Sample variance
5. Reject if $\frac{T_n(X, Y, Z)}{S_n^{\text{GCM}}(X, Y, Z)} > z_{1-\alpha}$.

The GCM test

(Shah and Peters ’20)

1. Fit an approximation $\hat{\mu}_{n,x}(\mathbf{Z})$ of $\mu_{n,x}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{X} \mid \mathbf{Z}]$ via machine learning;
2. Fit an approximation $\hat{\mu}_{n,y}(\mathbf{Z})$ of $\mu_{n,y}(\mathbf{Z}) \equiv \mathbb{E}_{\mathscr{L}_n}[\mathbf{Y} \mid \mathbf{Z}]$ via machine learning;
3. Compute $T_n(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))$
4. Compute $(S_n^{\text{GCM}})^2(X, Y, Z) \equiv \mathbb{V}\{(X_i - \hat{\mu}_{n,x}(Z_i))(Y_i - \hat{\mu}_{n,y}(Z_i))\}$ ← Sample variance
5. Reject if $\frac{T_n(X, Y, Z)}{S_n^{\text{GCM}}(X, Y, Z)} > z_{1-\alpha}$. ← Asymptotic threshold, rather than resampling-based.

Double robustness of the GCM test

Double robustness of the GCM test

The GCM test is doubly robust (Shah and Peters ’20)

Double robustness of the GCM test

The GCM test is doubly robust (Shah and Peters '20)

If RMSE$(\hat{\mu}_{n,x}) = o_P(1)$, RMSE$(\hat{\mu}_{n,y}) = o_P(1)$, RMSE$(\hat{\mu}_{n,x}) \cdot$ RMSE$(\hat{\mu}_{n,y}) = o_P(n^{-1/2})$ for each $\mathscr{L}_n \in H_0 \equiv H_0^{\text{CI}} \cap \mathscr{R}_n$, then GCM test has asymptotic Type-I error control.

Double robustness of the GCM test

The GCM test is doubly robust (Shah and Peters ’20)

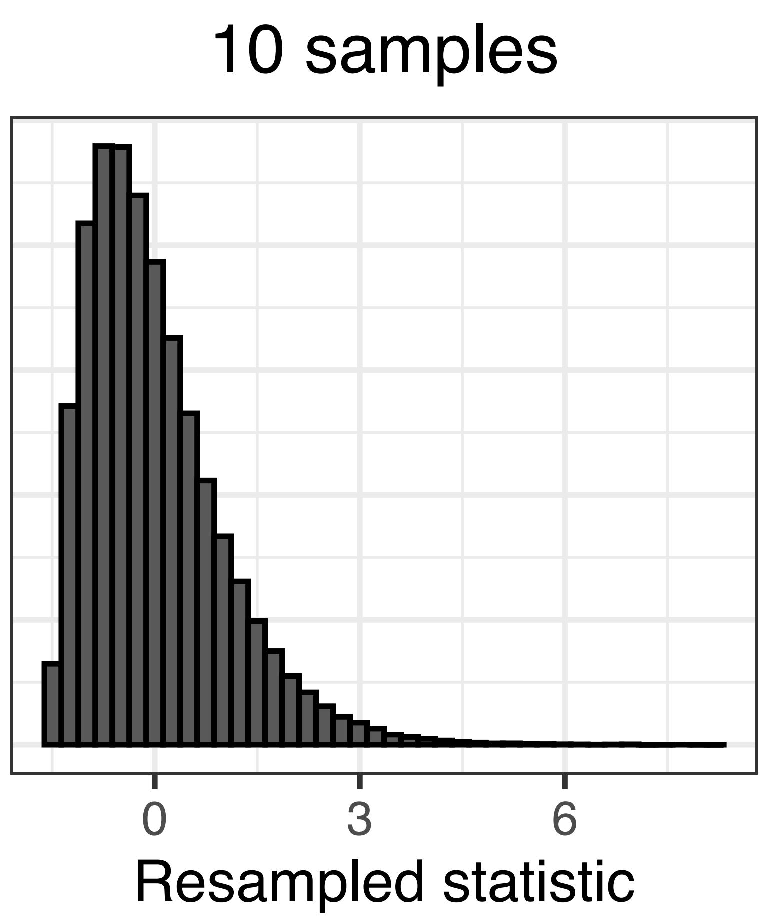
If RMSE$(\hat{\mu}_{n,x}) = o_P(1)$, RMSE$(\hat{\mu}_{n,y}) = o_P(1)$, RMSE$(\hat{\mu}_{n,x}) \cdot$ RMSE$(\hat{\mu}_{n,y}) = o_P(n^{-1/2})$ for each $\mathscr{L}_n \in H_0 \equiv H_0^{\text{CI}} \cap \mathscr{R}_n$, then GCM test has asymptotic Type-I error control.

These rates allow for high-dimensional regressions in the “consistency regime,” e.g. $s = o(\sqrt{n}/\log p)$.

Convergence of dCRT resampling distribution to normal

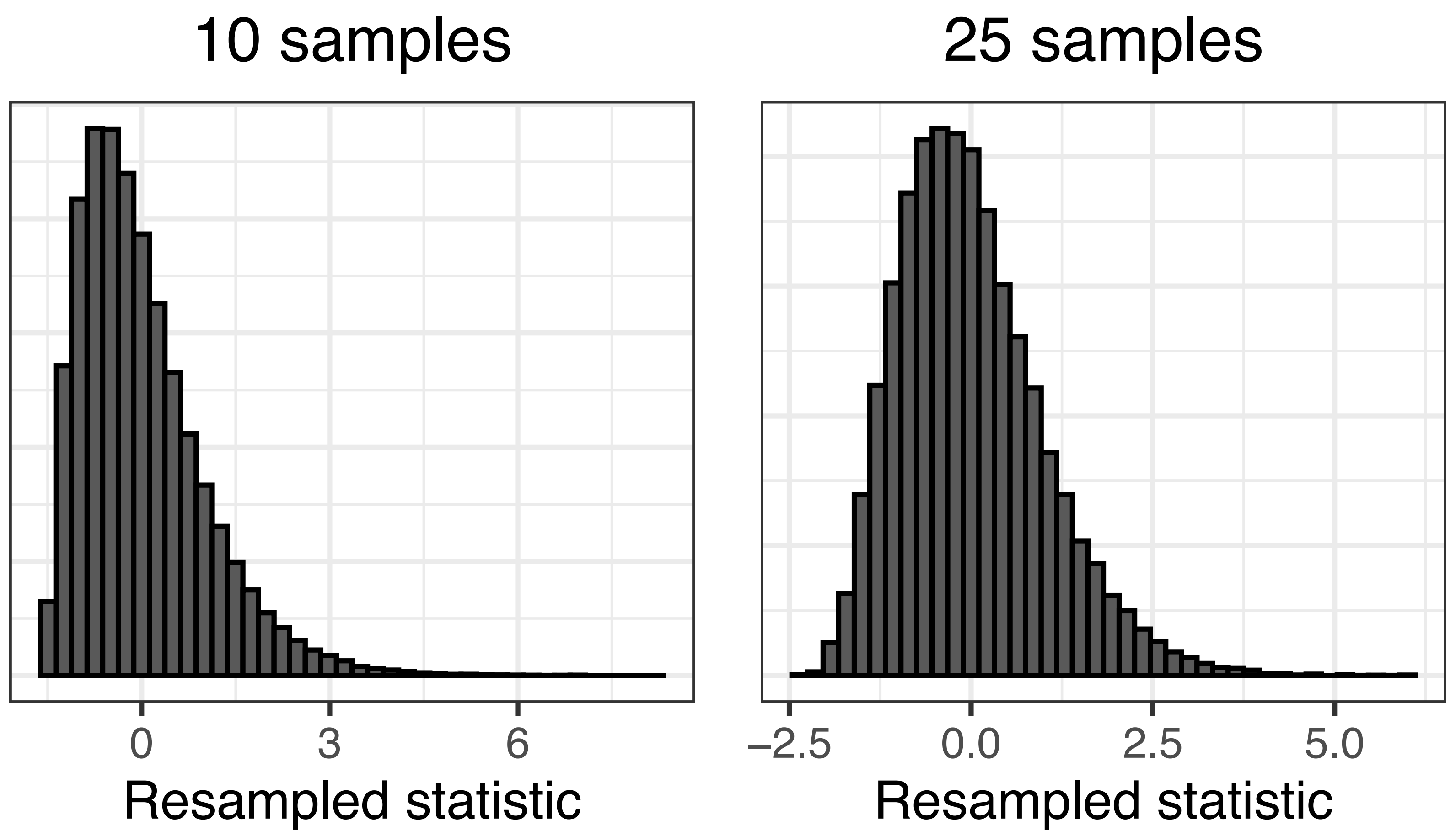
Convergence of dCRT resampling distribution to normal

For small n, dCRT resampling distribution need not be normal.



Convergence of dCRT resampling distribution to normal

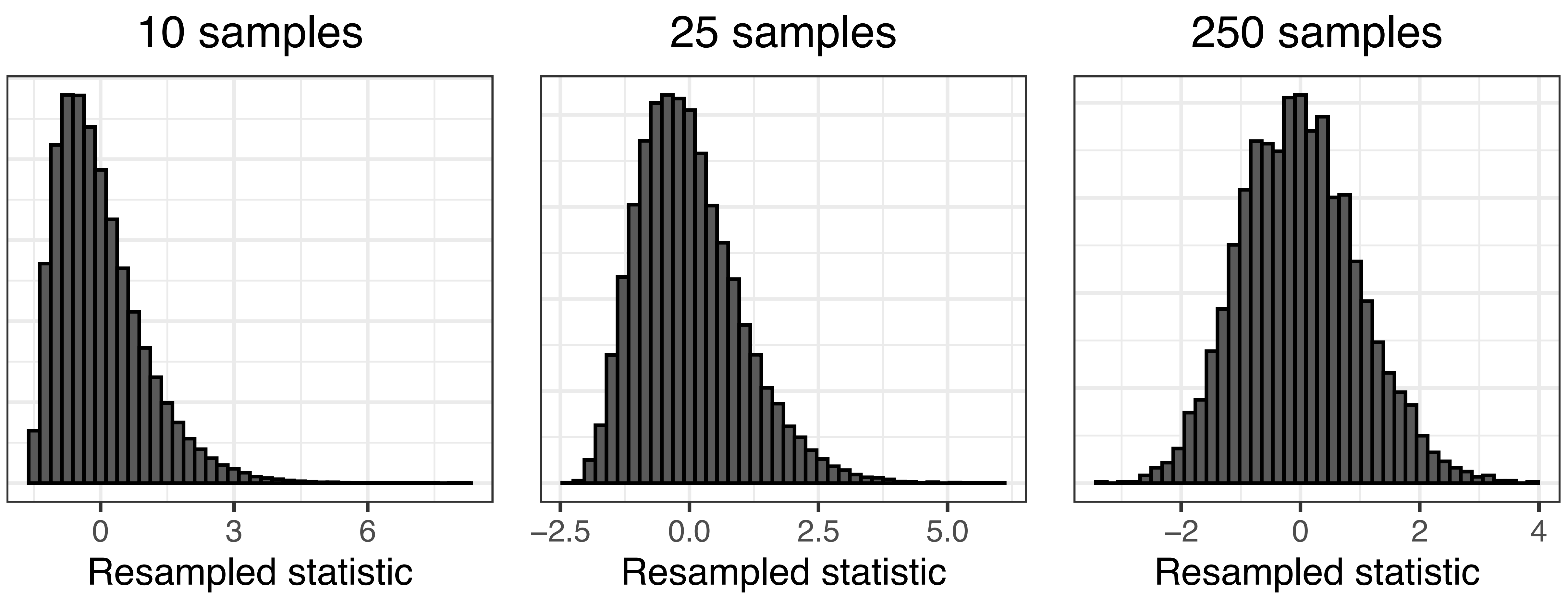
For small n, dCRT resampling distribution need not be normal.



Convergence of dCRT resampling distribution to normal

For small n, dCRT resampling distribution need not be normal.

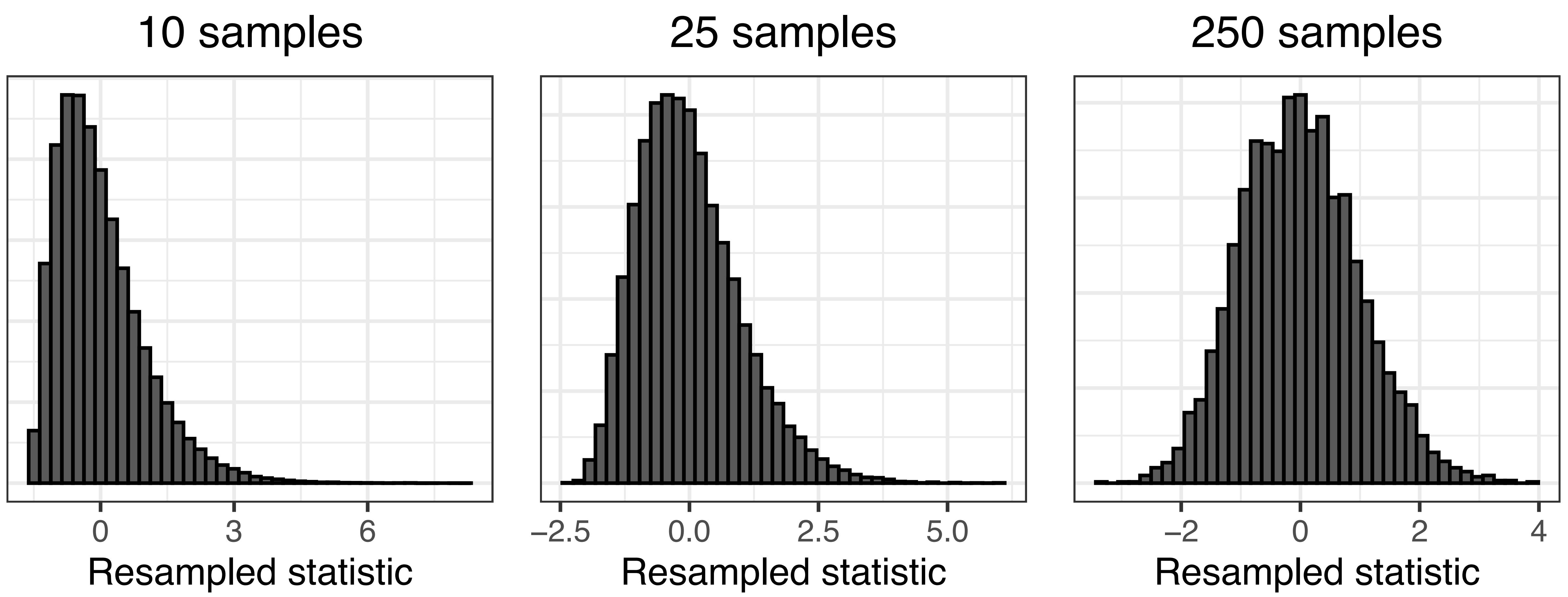
For large n, resampling recapitulates the asymptotic normal distribution.



Convergence of dCRT resampling distribution to normal

For small n, dCRT resampling distribution need not be normal.

For large n, resampling recapitulates the asymptotic normal distribution.

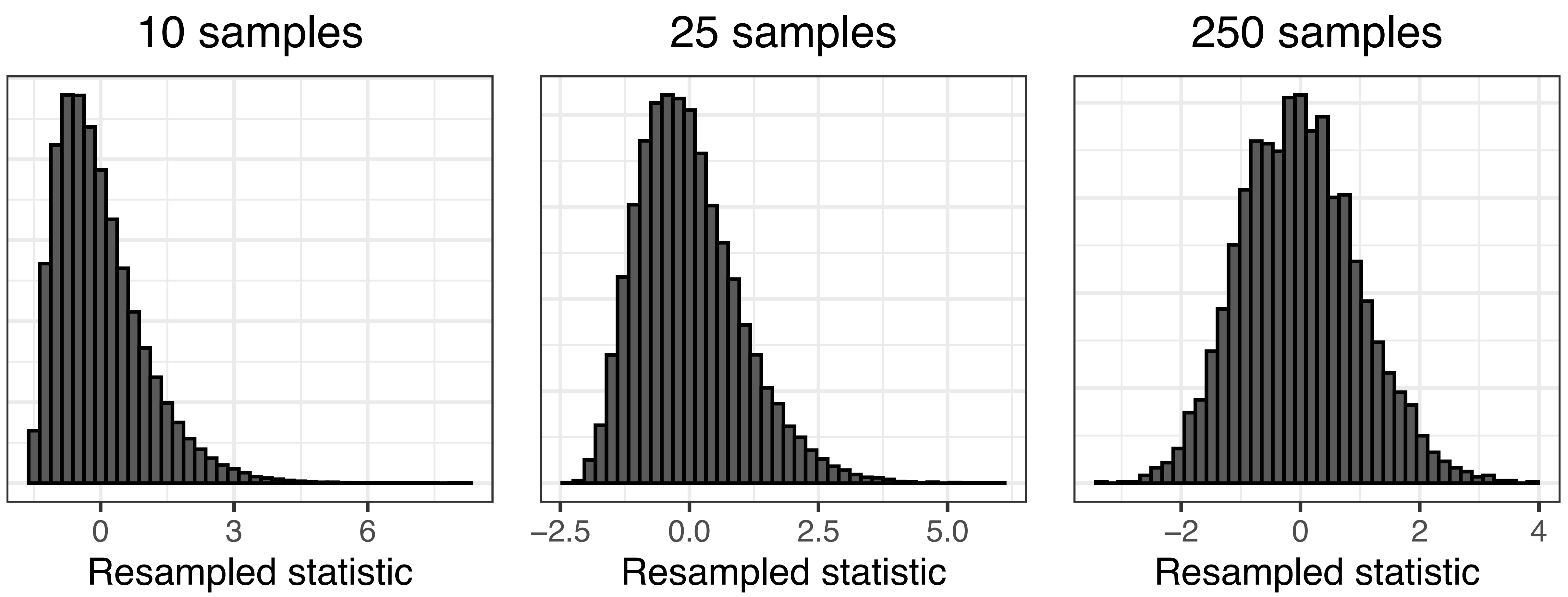


We show that (normalized) dCRT resampling distribution converges to $N(0,1)$ conditionally on the data: conditional CDF converges to $\Phi(x)$ in probability.

Convergence of dCRT resampling distribution to normal

For small n, dCRT resampling distribution need not be normal.

For large n, resampling recapitulates the asymptotic normal distribution.



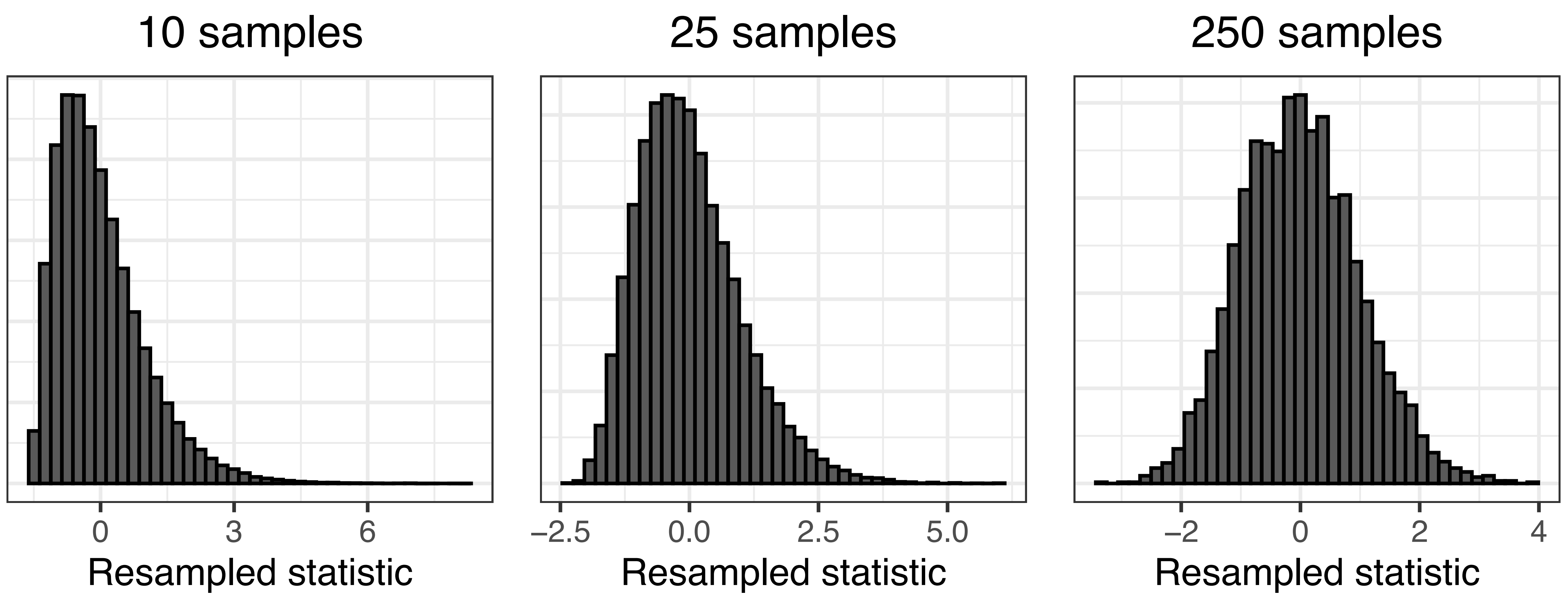
We show that (normalized) dCRT resampling distribution converges to $N(0,1)$ conditionally on the data: conditional CDF converges to $\Phi(x)$ in probability.

Technical challenge: the resampling distribution $\hat{\mathscr{L}}(X_i \mid Z_i)$ is based on estimate which is varying across i and n and resamples are only conditionally independent.

Convergence of dCRT resampling distribution to normal

For small n, dCRT resampling distribution need not be normal.

For large n, resampling recapitulates the asymptotic normal distribution.



We show that (normalized) dCRT resampling distribution converges to $N(0,1)$ conditionally on the data: conditional CDF converges to $\Phi(x)$ in probability.

Technical challenge: the resampling distribution $\hat{\mathscr{L}}(X_i \mid Z_i)$ is based on estimate which is varying across i and n and resamples are only conditionally independent.

We proved a new conditional CLT for triangular arrays!

dCRT-GCM equivalence and dCRT double robustness

dCRT-GCM equivalence and dCRT double robustness

Theorem (Niu et al '24; informal). Assume

1. $\text{RMSE}(\hat{\mu}_{n,x}) = o_P(1)$, $\text{RMSE}(\hat{\mu}_{n,y}) = o_P(1)$, $\text{RMSE}(\hat{\mu}_{n,x}) \cdot \text{RMSE}(\hat{\mu}_{n,y}) = o_P(n^{-1/2})$.
2. The estimated variances are "consistent" in some sense.

Then, for any $\mathscr{L}_n \in H_0$, the dCRT is asymptotically equivalent to the GCM test.

dCRT-GCM equivalence and dCRT double robustness

Theorem (Niu et al ’24; informal). Assume

1. RMSE($\hat{\mu}_{n,x}$) $= o_P(1)$, RMSE($\hat{\mu}_{n,y}$) $= o_P(1)$, RMSE($\hat{\mu}_{n,x}$) $\cdot$ RMSE($\hat{\mu}_{n,y}$) $= o_P(n^{-1/2})$.
2. The estimated variances are “consistent” in some sense.

Then, for any $\mathscr{L}_n \in H_0$, the dCRT is asymptotically equivalent to the GCM test.

Corollary (Niu et al ’24; informal). The dCRT is doubly robust, in the sense that it controls Type-I error under assumptions 1 and 2.

dCRT-GCM equivalence and dCRT double robustness

Theorem (Niu et al ’24; informal). Assume

1. RMSE$(\hat{\mu}_{n,x}) = o_P(1)$, RMSE$(\hat{\mu}_{n,y}) = o_P(1)$, RMSE$(\hat{\mu}_{n,x}) \cdot$ RMSE$(\hat{\mu}_{n,y}) = o_P(n^{-1/2})$.
2. The estimated variances are “consistent” in some sense.

Then, for any $\mathscr{L}_n \in H_0$, the dCRT is asymptotically equivalent to the GCM test.

Corollary (Niu et al ’24; informal). The dCRT is doubly robust, in the sense that it controls Type-I error under assumptions 1 and 2.

Fitting $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ improves not just power; it improves robustness as well.

Type-I error of CRT under different assumptions

Type-I error of CRT under different assumptions

(Previous work)

Type-I error of CRT under different assumptions

(Previous work)

CRT properties under
MX assumption ($\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ known)

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

Type-I error of CRT under different assumptions

(Previous work)

CRT properties under MX assumption ($\mathcal{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ known)

- Finite-sample Type-I error control
- No assumptions on $\mathcal{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

CRT properties if $\mathcal{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on $N \gg n \cdot \text{dim}(\mathbf{Z})$ samples

- Asymptotic Type-I error control
- No assumptions on $\mathcal{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

Type-I error of CRT under different assumptions

(Previous work)

CRT properties under MX assumption ($\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ known)

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on $N \gg n \cdot \text{dim}(\mathbf{Z})$ samples

- Asymptotic Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

Type-I error of CRT under different assumptions

(Previous work)

CRT properties under MX assumption ($\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ known)

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on $N \gg n \cdot \text{dim}(\mathbf{Z})$ samples

- Asymptotic Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

(Our work)

Type-I error of CRT under different assumptions

(Previous work)

CRT properties under MX assumption ($\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ known)

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on $N \gg n \cdot \dim(\mathbf{Z})$ samples

- Asymptotic Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

(Our work)

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

- Asymptotic Type-I error control
- $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$ cannot be too complex
- Type-I error control for only certain T_n

Type-I error of CRT under different assumptions

Theoretical
(Previous work)

CRT properties under MX assumption ($\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ known)

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on $N \gg n \cdot \dim(\mathbf{Z})$ samples

- Asymptotic Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

(Our work)

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

- Asymptotic Type-I error control
- $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$ cannot be too complex
- Type-I error control for only certain T_n

Type-I error of CRT under different assumptions

Theoretical
(Previous work)

CRT properties under MX assumption ($\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ known)

- Finite-sample Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned on $N \gg n \cdot \dim(\mathbf{Z})$ samples

- Asymptotic Type-I error control
- No assumptions on $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$
- Test statistic T_n can be arbitrary

Practical
(Our work)

CRT properties if $\mathscr{L}_n^*(\mathbf{X} \mid \mathbf{Z})$ learned in sample

- Asymptotic Type-I error control
- $\mathscr{L}_n(\mathbf{Y} \mid \mathbf{Z})$ cannot be too complex
- Type-I error control for only certain T_n

Then, why do we use dCRT?

Then, why do we use dCRT?

- When effective sample size is small, dCRT p-value has better calibration

Then, why do we use dCRT?

- When effective sample size is small, dCRT p-value has better calibration

$$\mathbf{X} \sim \text{Bern}(\text{expit}(-4 + \mathbf{Z})),\ \mathbf{Y} \sim \text{Pois}(\exp(-3 + \mathbf{Z})),\ \mathbf{Z} \sim N(0,1),\ n = 1000$$

Then, why do we use dCRT?

- When effective sample size is small, dCRT p-value has better calibration

$$\mathbf{X} \sim \text{Bern}(\text{expit}(-4 + \mathbf{Z})),\ \mathbf{Y} \sim \text{Pois}(\exp(-3 + \mathbf{Z})),\ \mathbf{Z} \sim N(0,1),\ n = 1000$$

Then, why do we use dCRT?

- When effective sample size is small, dCRT p-value has better calibration

$$\mathbf{X} \sim \text{Bern}(\text{expit}(-4 + \mathbf{Z})),\ \mathbf{Y} \sim \text{Pois}(\exp(-3 + \mathbf{Z})),\ \mathbf{Z} \sim N(0,1),\ n = 1000$$

Then, why do we use dCRT?

- When effective sample size is small, dCRT p-value has better calibration

$$\mathbf{X} \sim \text{Bern}(\text{expit}(-4 + \mathbf{Z})),\ \mathbf{Y} \sim \text{Pois}(\exp(-3 + \mathbf{Z})),\ \mathbf{Z} \sim N(0,1),\ n = 1000$$

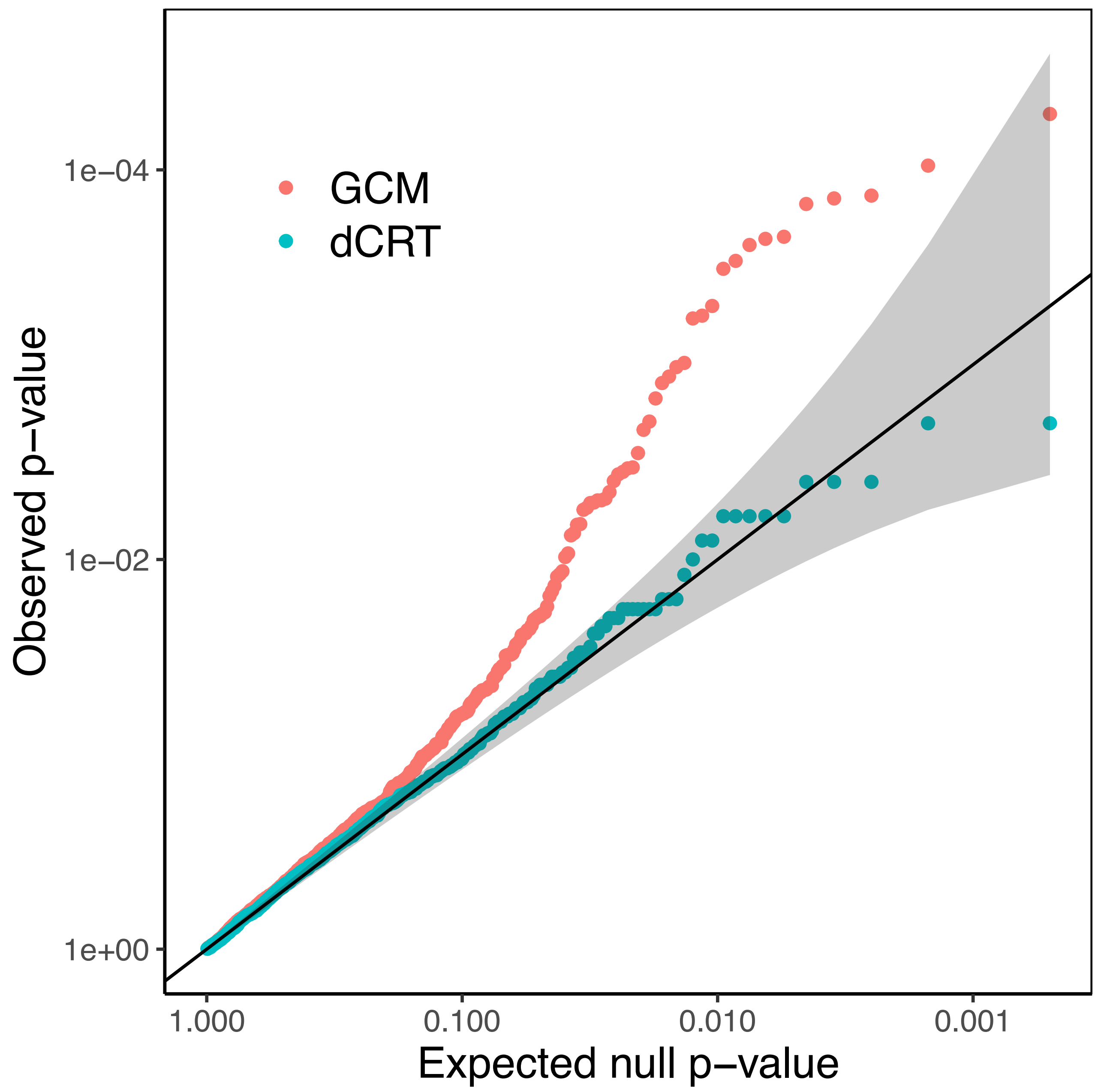
Sparse $\mathbf{X}$ and $\mathbf{Y}$, as is common in single-cell genomics

Then, why do we use dCRT?

- When effective sample size is small, dCRT p-value has better calibration

$$\mathbf{X} \sim \text{Bern}(\text{expit}(-4 + \mathbf{Z})),\ \mathbf{Y} \sim \text{Pois}(\exp(-3 + \mathbf{Z})),\ \mathbf{Z} \sim N(0,1),\ n = 1000$$

Sparse $\mathbf{X}$ and $\mathbf{Y}$, as is common in single-cell genomics

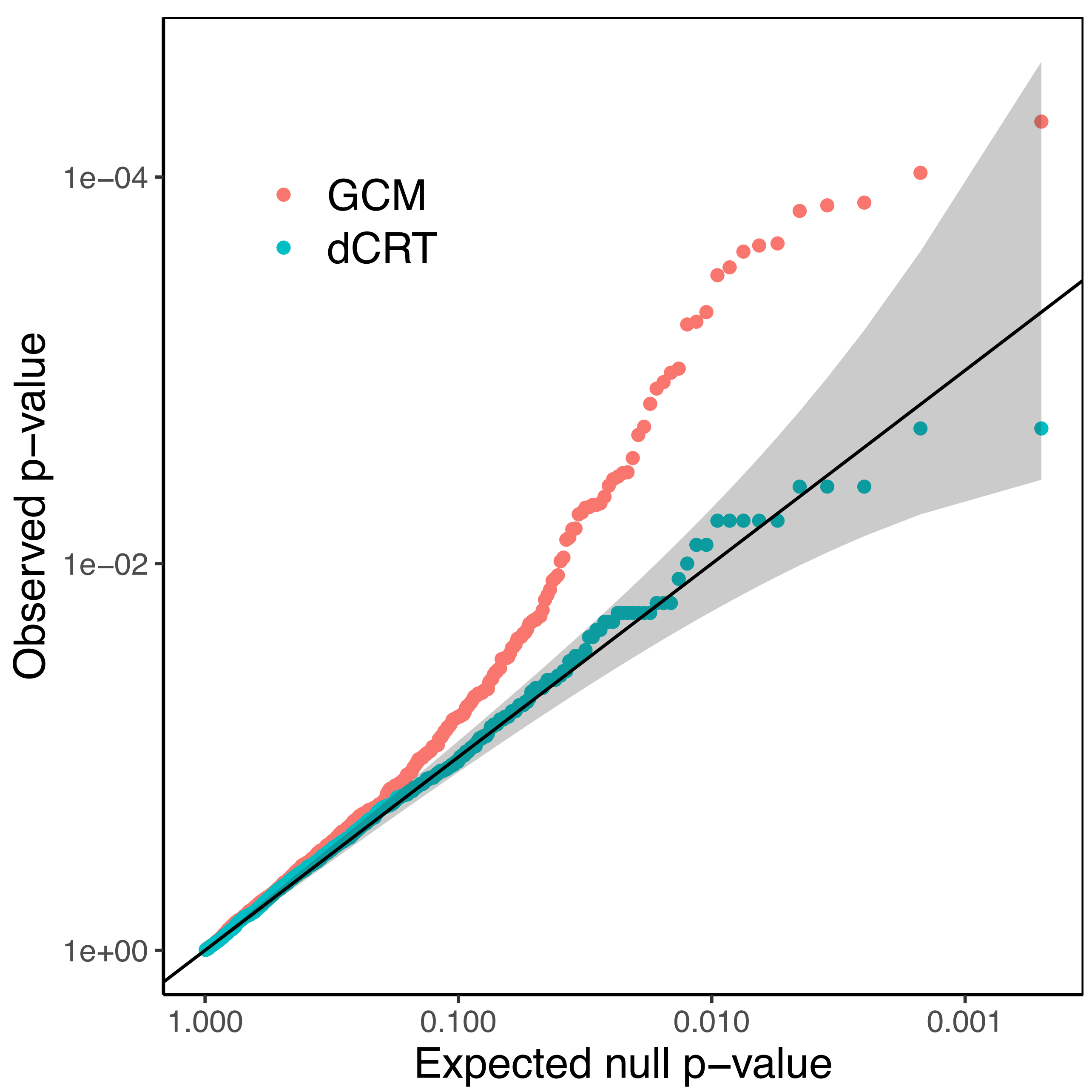


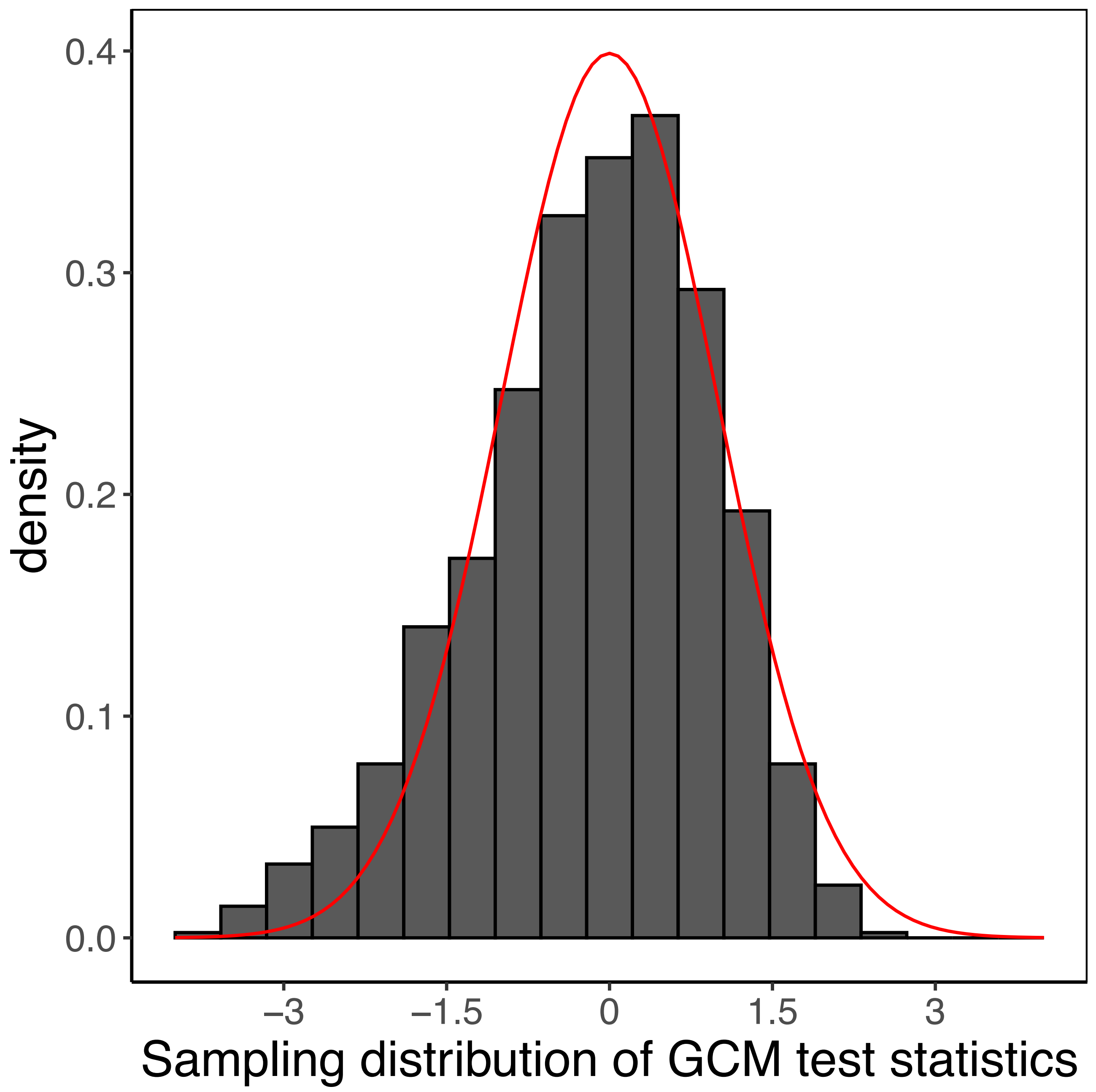
Then, why do we use dCRT?

- When effective sample size is small, dCRT p-value has better calibration

$$\mathbf{X} \sim \text{Bern}(\text{expit}(-4 + \mathbf{Z})),\ \mathbf{Y} \sim \text{Pois}(\exp(-3 + \mathbf{Z})),\ \mathbf{Z} \sim N(0,1),\ n = 1000$$

Sparse $\mathbf{X}$ and $\mathbf{Y}$, as is common in single-cell genomics





Preview of follow-up work: A resampling-free dCRT

Preview of follow-up work: A resampling-free dCRT

A computational challenge:

dCRT requires a large number of resamples to obtain accurate small p-values.

A statistical solution:

We extend saddlepoint approximation theory to approximate the conditional tail probability of the resampling distribution.

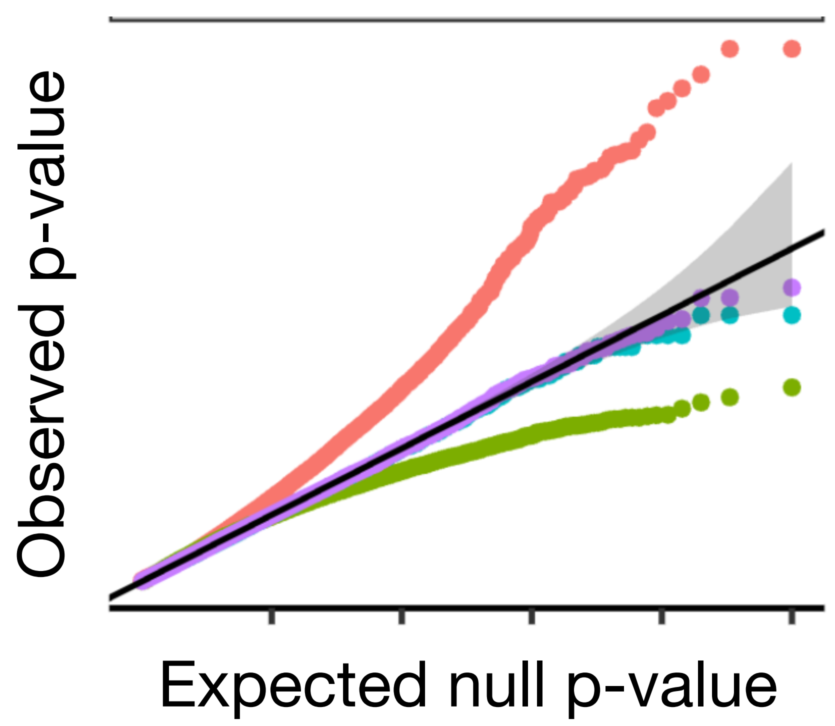
Preview of follow-up work: A resampling-free dCRT

A computational challenge:

dCRT requires a large number of resamples to obtain accurate small p-values.

A statistical solution:

We extend saddlepoint approximation theory to approximate the conditional tail probability of the resampling distribution.



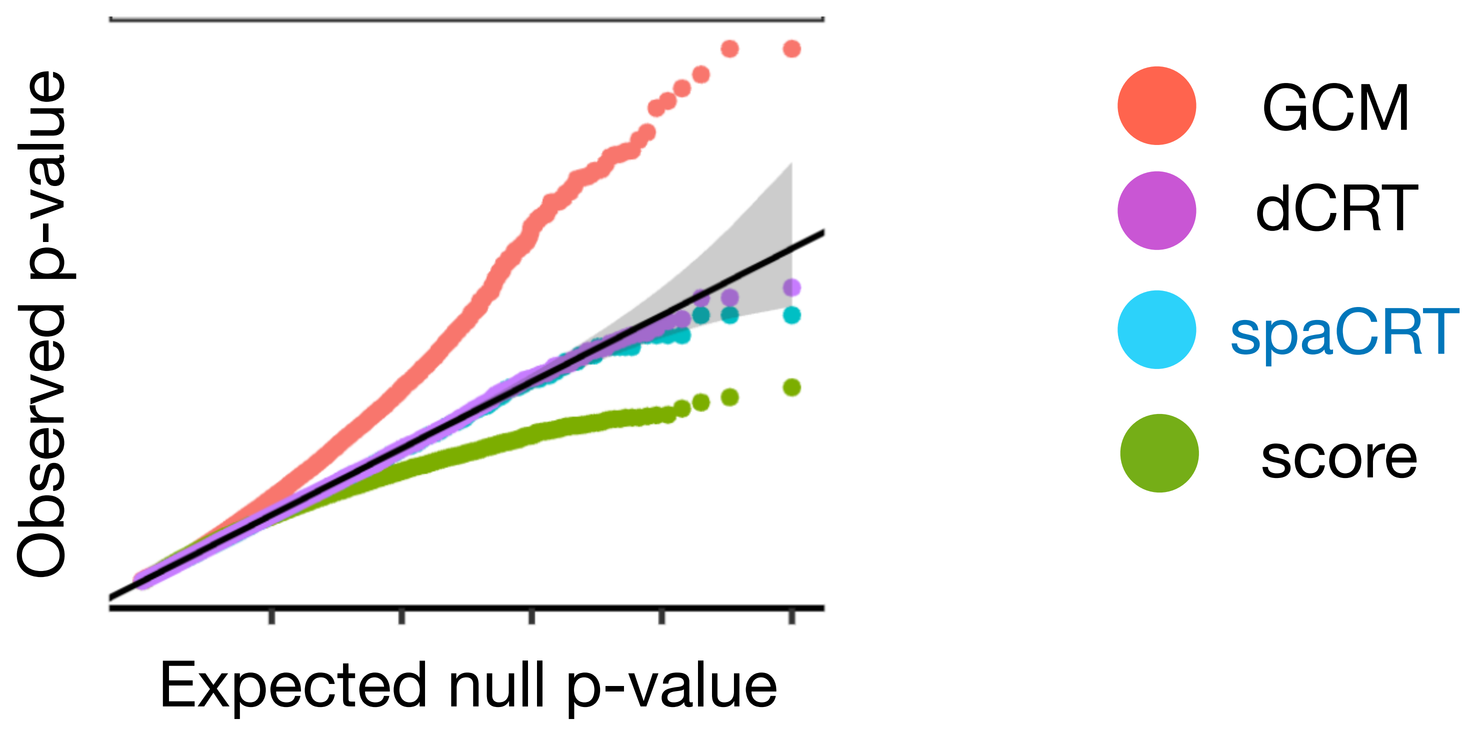
Preview of follow-up work: A resampling-free dCRT

A computational challenge:

dCRT requires a large number of resamples to obtain accurate small p-values.

A statistical solution:

We extend saddlepoint approximation theory to approximate the conditional tail probability of the resampling distribution.



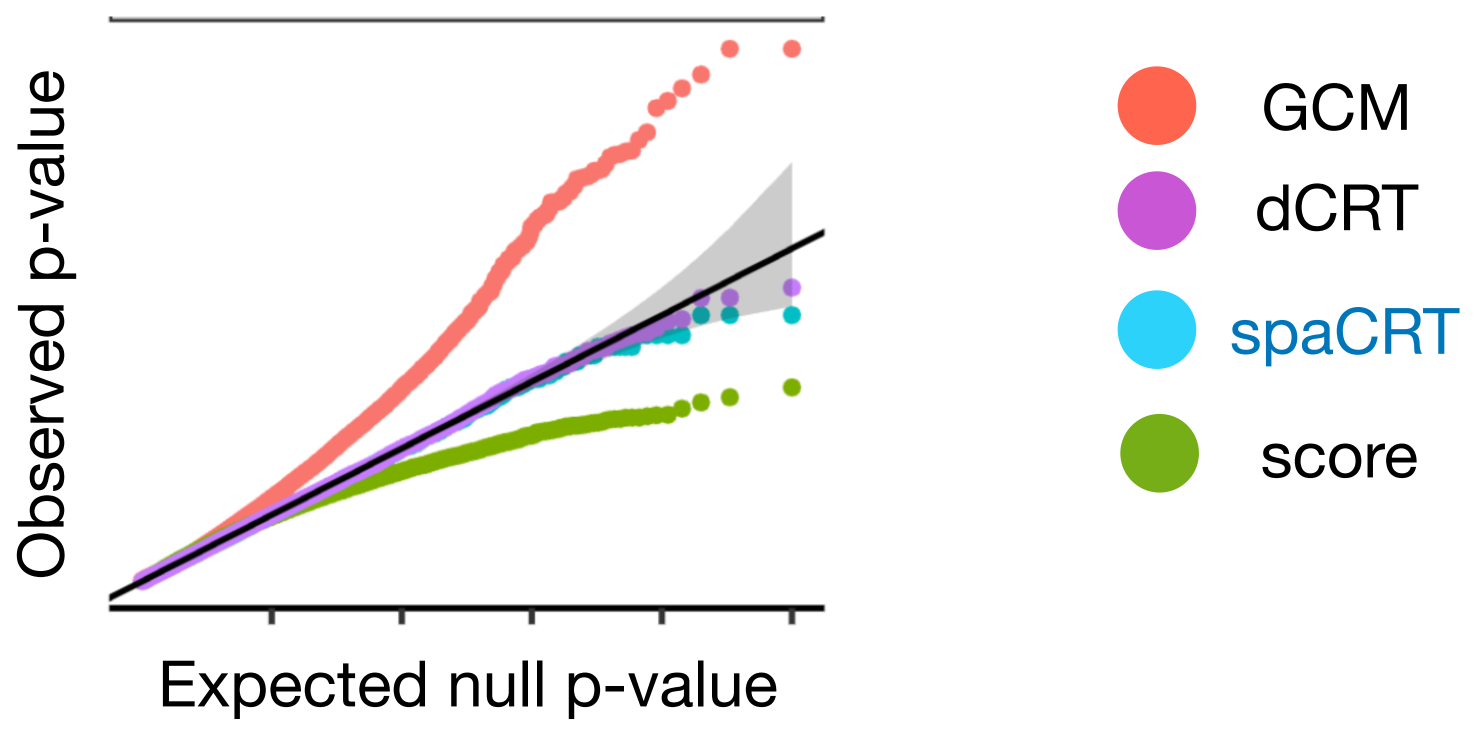
Preview of follow-up work: A resampling-free dCRT

A computational challenge:

dCRT requires a large number of resamples to obtain accurate small p-values.

A statistical solution:

We extend saddlepoint approximation theory to approximate the conditional tail probability of the resampling distribution.



spaCRT is completely resampling-free
and almost as fast as GCM!

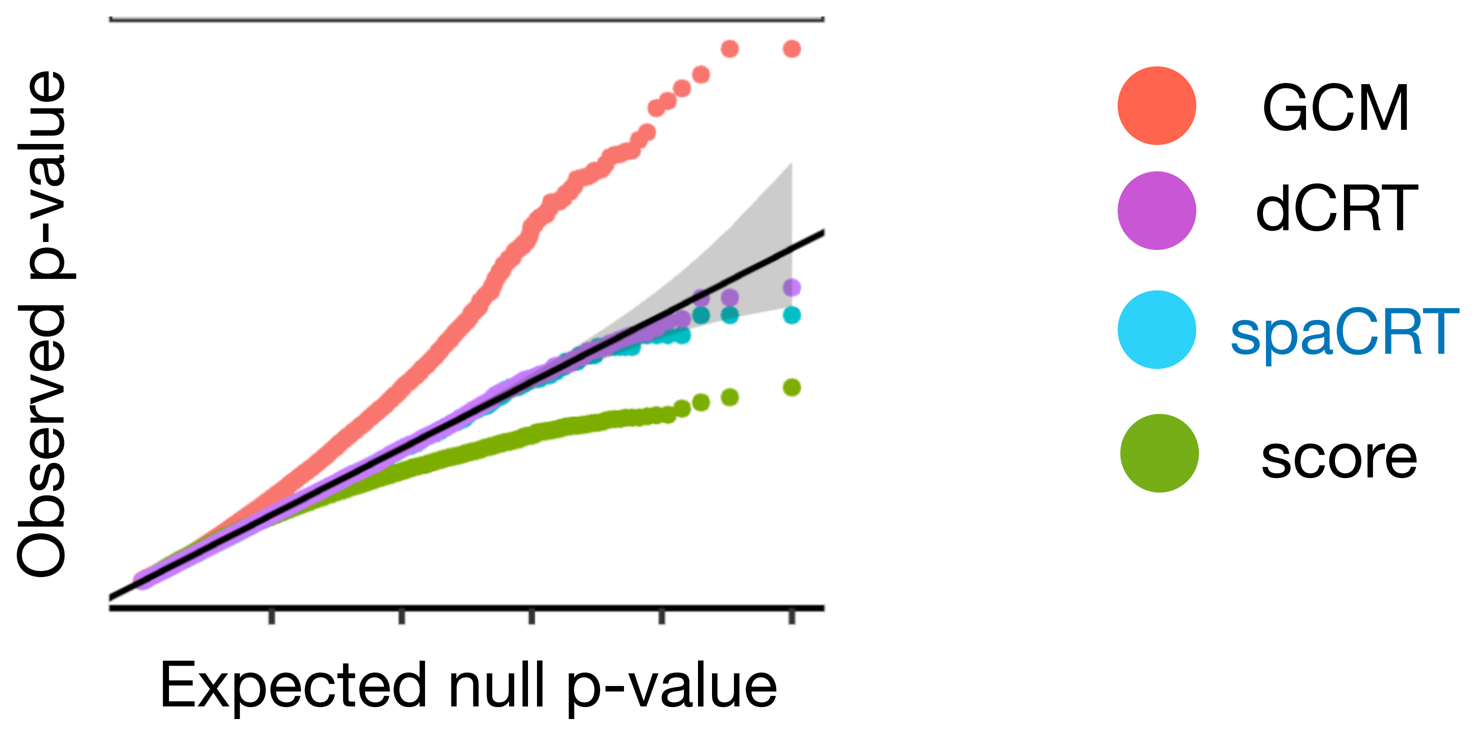
Preview of follow-up work: A resampling-free dCRT

A computational challenge:

dCRT requires a large number of resamples to obtain accurate small p-values.

A statistical solution:

We extend saddlepoint approximation theory to approximate the conditional tail probability of the resampling distribution.



spaCRT is completely resampling-free and almost as fast as GCM!

Stay tuned for the incoming paper!

Discussion

Discussion

Take-home message:

Discussion

Take-home message:

- When $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ fit in sample, MX inference is like doubly robust inference; Type-I error control possible, but both $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ and $\widehat{\mathscr{L}}_n(\mathbf{Y} \mid \mathbf{Z})$ matter.

Discussion

Take-home message:

- When $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ fit in sample, MX inference is like doubly robust inference; Type-I error control possible, but both $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ and $\widehat{\mathscr{L}}_n(\mathbf{Y} \mid \mathbf{Z})$ matter.
- We bridge the double robustness literature (asymptotic framework) and model-X literature (finite-sample framework), filling a gap between theory and practice in CRT.

Discussion

Take-home message:

- When $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ fit in sample, MX inference is like doubly robust inference; Type-I error control possible, but both $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ and $\widehat{\mathscr{L}}_n(\mathbf{Y} \mid \mathbf{Z})$ matter.
- We bridge the double robustness literature (asymptotic framework) and model-X literature (finite-sample framework), filling a gap between theory and practice in CRT.
- dCRT is particularly useful under low effective sample size regime.

Discussion

Take-home message:

- When $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ fit in sample, MX inference is like doubly robust inference; Type-I error control possible, but both $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ and $\widehat{\mathscr{L}}_n(\mathbf{Y} \mid \mathbf{Z})$ matter.
- We bridge the double robustness literature (asymptotic framework) and model-X literature (finite-sample framework), filling a gap between theory and practice in CRT.
- dCRT is particularly useful under low effective sample size regime.

Open questions:

Discussion

Take-home message:

- When $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ fit in sample, MX inference is like doubly robust inference; Type-I error control possible, but both $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ and $\widehat{\mathscr{L}}_n(\mathbf{Y} \mid \mathbf{Z})$ matter.
- We bridge the double robustness literature (asymptotic framework) and model-X literature (finite-sample framework), filling a gap between theory and practice in CRT.
- dCRT is particularly useful under low effective sample size regime.

Open questions:

- Extensions to other test statistics beyond dCRT

Discussion

Take-home message:

- When $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ fit in sample, MX inference is like doubly robust inference; Type-I error control possible, but both $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ and $\widehat{\mathscr{L}}_n(\mathbf{Y} \mid \mathbf{Z})$ matter.
- We bridge the double robustness literature (asymptotic framework) and model-X literature (finite-sample framework), filling a gap between theory and practice in CRT.
- dCRT is particularly useful under low effective sample size regime.

Open questions:

- Extensions to other test statistics beyond dCRT
- Extensions to knockoffs

Discussion

Take-home message:

- When $\mathscr{L}_n(\mathbf{X} \mid \mathbf{Z})$ fit in sample, MX inference is like doubly robust inference; Type-I error control possible, but both $\widehat{\mathscr{L}}_n(\mathbf{X} \mid \mathbf{Z})$ and $\widehat{\mathscr{L}}_n(\mathbf{Y} \mid \mathbf{Z})$ matter.
- We bridge the double robustness literature (asymptotic framework) and model-X literature (finite-sample framework), filling a gap between theory and practice in CRT.
- dCRT is particularly useful under low effective sample size regime.

Open questions:

- Extensions to other test statistics beyond dCRT
- Extensions to knockoffs
- Moving beyond the “consistency regime,” e.g. to proportional asymptotics

Thank you!

Thank you!

Questions?

Is dCRT robust to in-sample learning?

Is dCRT robust to in-sample learning?

Simple numerical simulation:

- $\mathscr{L}(\mathbf{Z}) \sim N(0,I)$; $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$; $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$;
- $n = 1600$, $p = 400$, β has 5 nonzero elements;
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Is dCRT robust to in-sample learning?

Simple numerical simulation:

- $\mathscr{L}(\mathbf{Z}) \sim N(0,I)$; $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$; $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$;
- $n = 1600$, $p = 400$, β has 5 nonzero elements;
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

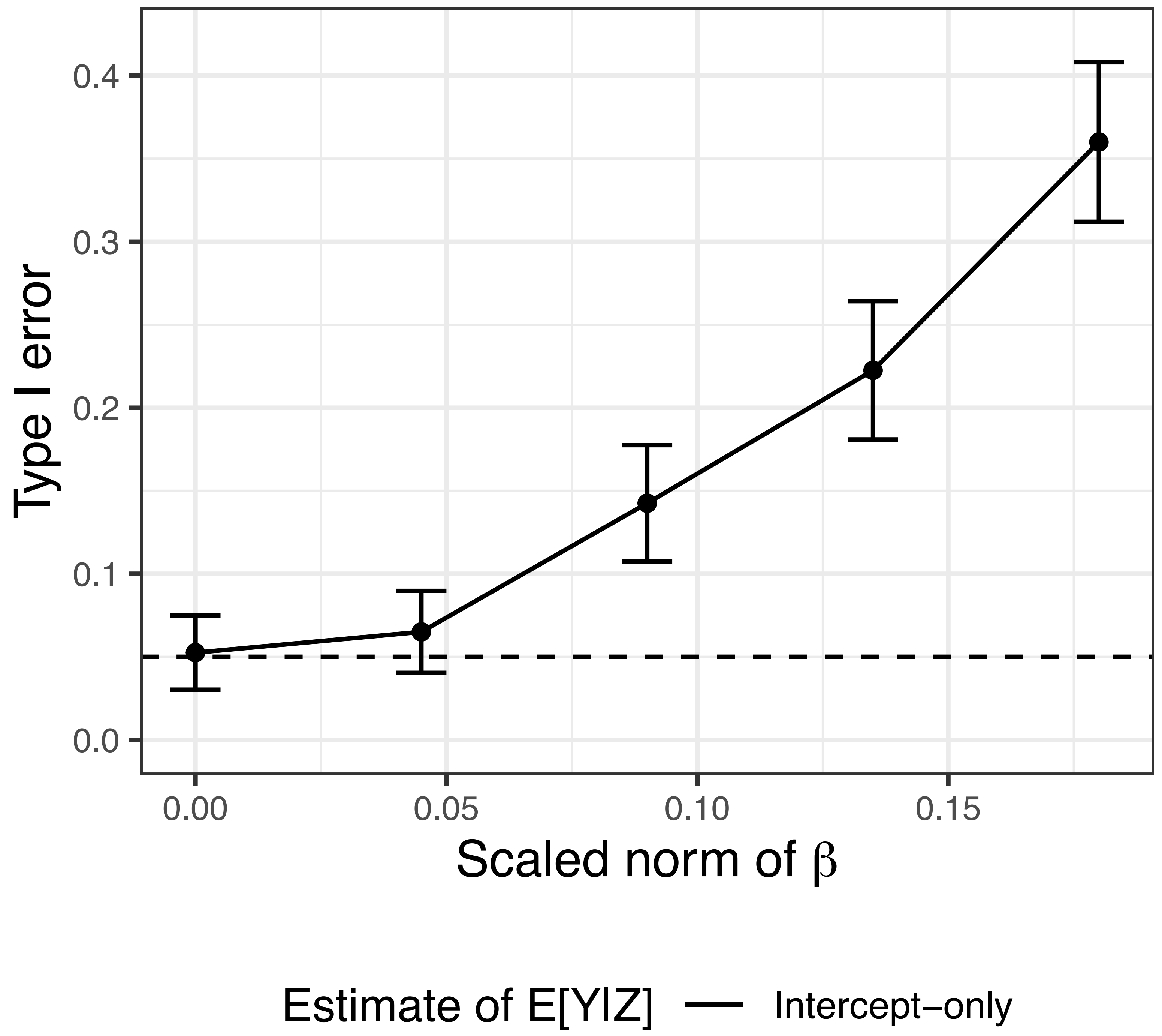
Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\widehat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Is dCRT robust to in-sample learning?

Simple numerical simulation:

- $\mathscr{L}(\mathbf{Z}) \sim N(0, I)$; $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta, 1)$; $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta, 1)$;
- $n = 1600$, $p = 400$, β has 5 nonzero elements;
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\hat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.



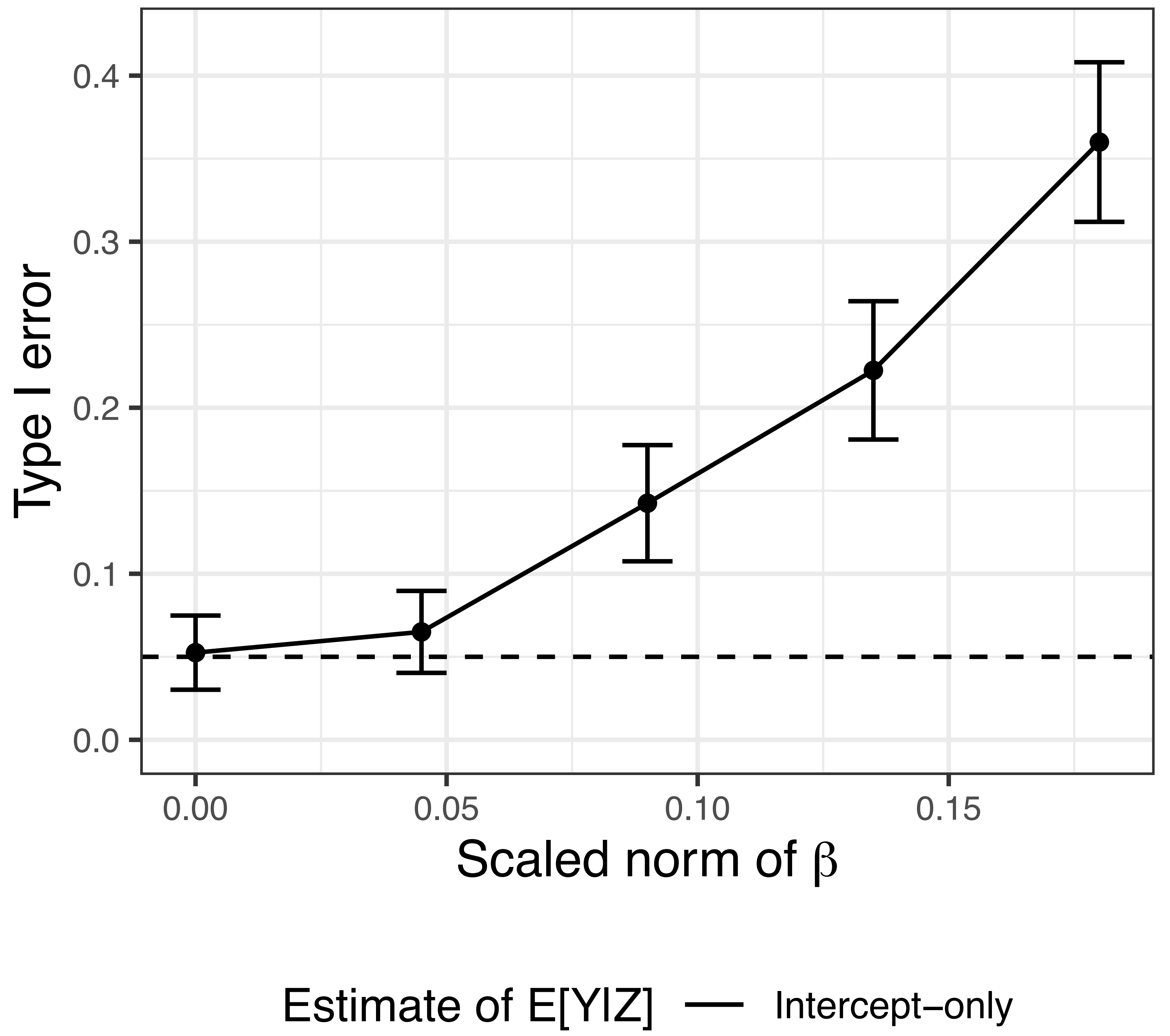
Is dCRT robust to in-sample learning?

Simple numerical simulation:

- $\mathscr{L}(\mathbf{Z}) \sim N(0, I)$; $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta, 1)$; $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta, 1)$;
- $n = 1600$, $p = 400$, β has 5 nonzero elements;
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\hat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Case 2: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated decently:
$\hat{\mu}_{n,y}(\mathbf{Z})$ obtained via lasso of Y on Z.



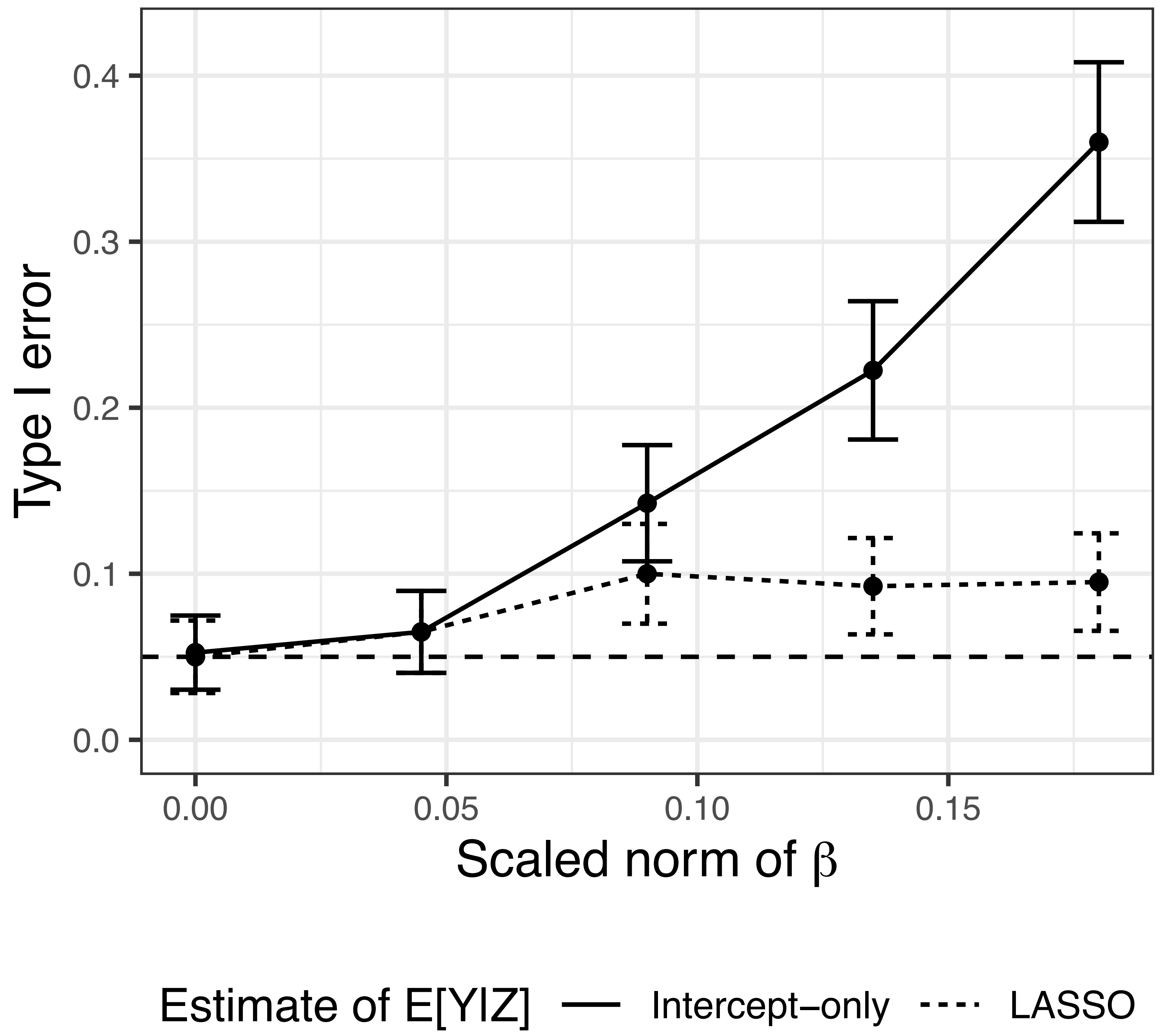
Is dCRT robust to in-sample learning?

Simple numerical simulation:

- $\mathscr{L}(\mathbf{Z}) \sim N(0,I)$; $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$; $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$;
- $n = 1600$, $p = 400$, β has 5 nonzero elements;
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\hat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Case 2: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated decently:
$\hat{\mu}_{n,y}(\mathbf{Z})$ obtained via lasso of Y on Z.



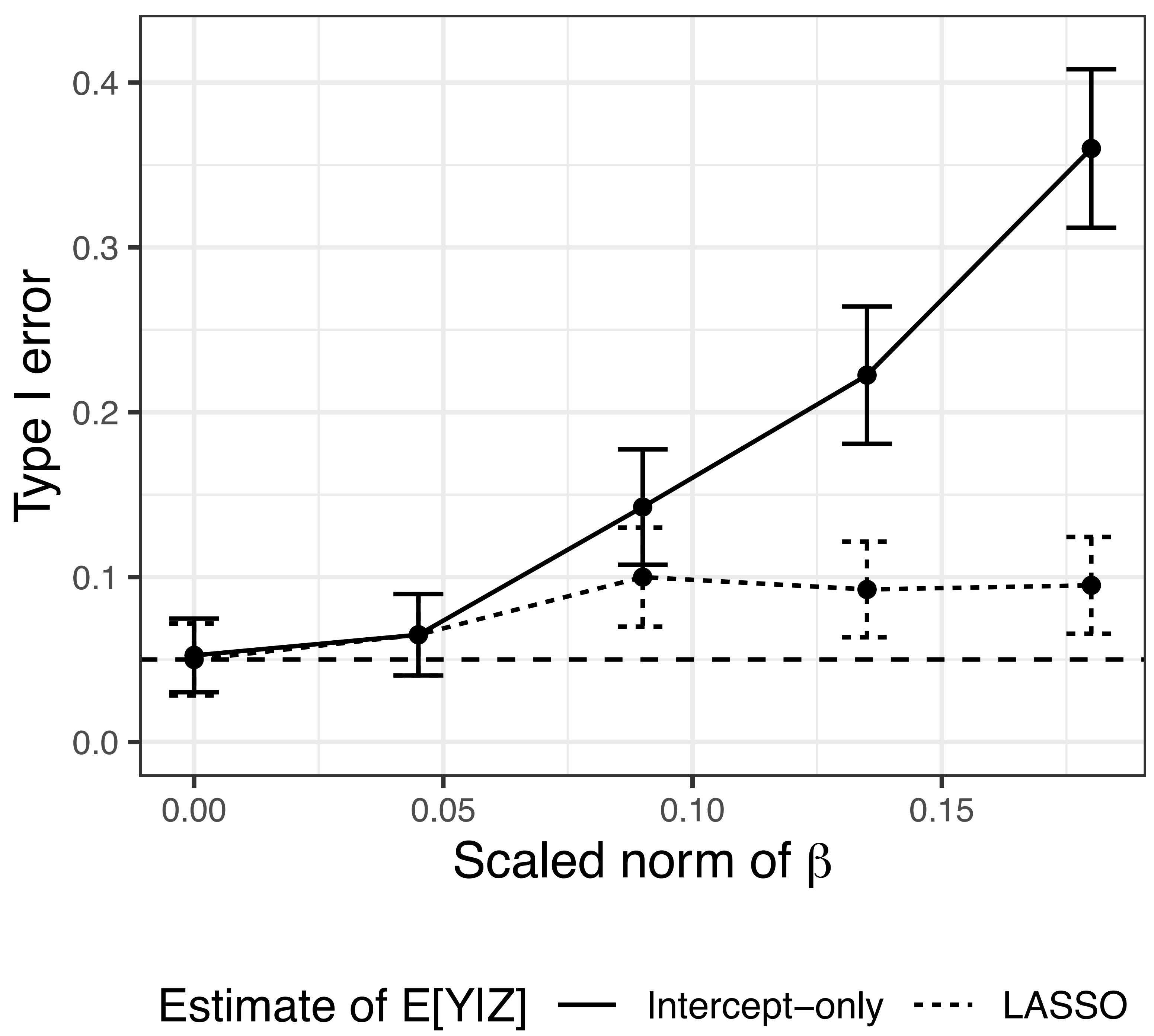
Is dCRT robust to in-sample learning?

Simple numerical simulation:

- $\mathscr{L}(\mathbf{Z}) \sim N(0,I)$; $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$; $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$;
- $n = 1600$, $p = 400$, β has 5 nonzero elements;
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\widehat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Case 2: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated decently:
$\widehat{\mu}_{n,y}(\mathbf{Z})$ obtained via lasso of Y on Z.



No hope for good inference with poor estimate for $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$.

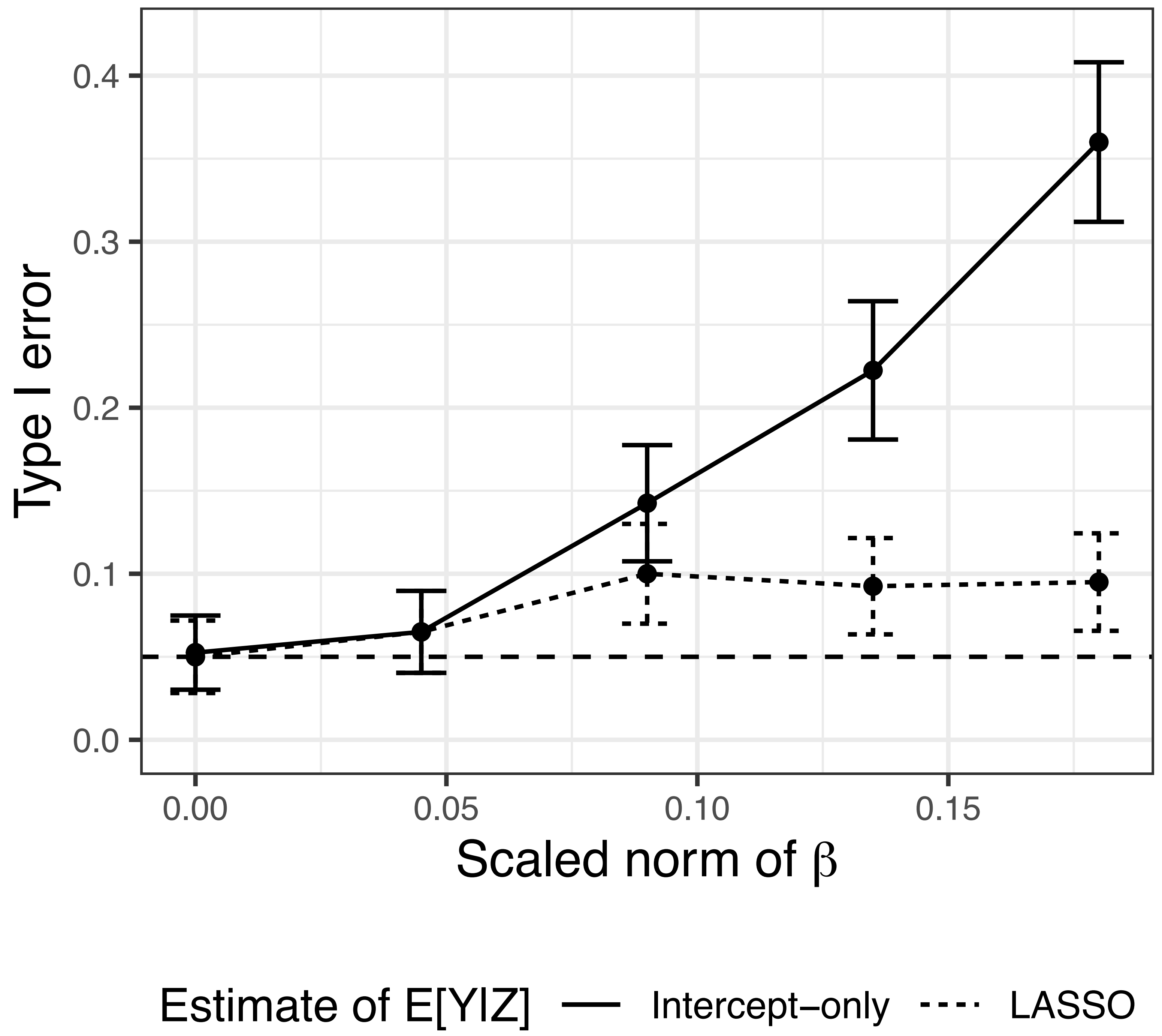
Is dCRT robust to in-sample learning?

Simple numerical simulation:

- $\mathscr{L}(\mathbf{Z}) \sim N(0,I)$; $\mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$; $\mathscr{L}(\mathbf{Y} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1)$;
- $n = 1600$, $p = 400$, β has 5 nonzero elements;
- $\mathscr{L}(\mathbf{X} \mid \mathbf{Z})$ estimated via the lasso of X on Z.

Case 1: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated poorly: $\widehat{\mu}_{n,y}(\mathbf{Z}) \equiv 0$.

Case 2: $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ estimated decently:
$\widehat{\mu}_{n,y}(\mathbf{Z})$ obtained via lasso of Y on Z.



No hope for good inference with poor estimate for $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$.

Better estimate of $\mathbb{E}[\mathbf{Y} \mid \mathbf{Z}]$ improves robustness of dCRT.

dCRT and GCM have same asymptotic power

dCRT and GCM have same asymptotic power

Corollary (Niu et al '24; informal). Assume

1. RMSE($\hat{\mu}_{n,x}$) $= o_P(1)$, RMSE($\hat{\mu}_{n,y}$) $= o_P(1)$, RMSE($\hat{\mu}_{n,x}$) $\cdot$ RMSE($\hat{\mu}_{n,y}$) $= o_P(n^{-1/2})$.
2. The estimated variances are consistent in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n}(\text{Var}_{\widehat{\mathscr{L}}_n}[X_i \mid Z_i] - \text{Var}_{\mathscr{L}_n}[X_i \mid Z_i])\text{Var}_{\mathscr{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0.$$

dCRT and GCM have same asymptotic power

Corollary (Niu et al ’24; informal). Assume

1. RMSE$(\hat{\mu}_{n,x}) = o_P(1)$, RMSE$(\hat{\mu}_{n,y}) = o_P(1)$, RMSE$(\hat{\mu}_{n,x}) \cdot$ RMSE$(\hat{\mu}_{n,y}) = o_P(n^{-1/2})$.
2. The estimated variances are consistent in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n}(\text{Var}_{\widehat{\mathscr{L}}_n}[X_i \mid Z_i] - \text{Var}_{\mathscr{L}_n}[X_i \mid Z_i])\text{Var}_{\mathscr{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0.$$

Then, for any sequence $\mathscr{L}_n$ of local alternatives, the dCRT is asymptotically equivalent to the GCM test, i.e.

$$\lim_{n\to\infty} \mathbb{P}_{\mathscr{L}_n}[\text{GCM test and dCRT coincide}] = 1.$$

dCRT and GCM have same asymptotic power

Corollary (Niu et al '24; informal). Assume

1. RMSE$(\hat{\mu}_{n,x}) = o_P(1)$, RMSE$(\hat{\mu}_{n,y}) = o_P(1)$, RMSE$(\hat{\mu}_{n,x}) \cdot$ RMSE$(\hat{\mu}_{n,y}) = o_P(n^{-1/2})$.
2. The estimated variances are consistent in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n}(\text{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i] - \text{Var}_{\mathcal{L}_n}[X_i \mid Z_i])\text{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0.$$

Then, for any sequence $\mathcal{L}_n$ of local alternatives, the dCRT is asymptotically equivalent to the GCM test, i.e.

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[\text{GCM test and dCRT coincide}] = 1.$$

- In large samples, GCM is preferable to dCRT because it avoids resampling.

dCRT and GCM have same asymptotic power

Corollary (Niu et al ’24; informal). Assume

1. $\text{RMSE}(\hat{\mu}_{n,x}) = o_P(1)$, $\text{RMSE}(\hat{\mu}_{n,y}) = o_P(1)$, $\text{RMSE}(\hat{\mu}_{n,x}) \cdot \text{RMSE}(\hat{\mu}_{n,y}) = o_P(n^{-1/2})$.
2. The estimated variances are consistent in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n}(\text{Var}_{\widehat{\mathscr{L}}_n}[X_i \mid Z_i] - \text{Var}_{\mathscr{L}_n}[X_i \mid Z_i])\text{Var}_{\mathscr{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0.$$

Then, for any sequence $\mathscr{L}_n$ of local alternatives, the dCRT is asymptotically equivalent to the GCM test, i.e.

$$\lim_{n\to\infty} \mathbb{P}_{\mathscr{L}_n}[\text{GCM test and dCRT coincide}] = 1.$$

- In large samples, GCM is preferable to dCRT because it avoids resampling.
- In small samples, dCRT may still be preferable to GCM.

Can other CRT variants be more powerful than GCM?

Can other CRT variants be more powerful than GCM?

Consider alternatives specified by the partially linear semiparametric model

$$\mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

Can other CRT variants be more powerful than GCM?

Consider alternatives specified by the partially linear semiparametric model

$$\mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

For this model, the oracle product-of-residuals test statistic

$$T_n^{\text{oracle}}(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \mu_{n,y}(Z_i))$$

is efficient score statistic, with optimal asymptotic power against local alternatives.

Can other CRT variants be more powerful than GCM?

Consider alternatives specified by the partially linear semiparametric model

$$\mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

For this model, the oracle product-of-residuals test statistic

$$T_n^{\text{oracle}}(X,Y,Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \mu_{n,y}(Z_i))$$

is efficient score statistic, with optimal asymptotic power against local alternatives.

Then, a test based on $T_n^{\text{oracle}}(X,Y,Z)$ is the most powerful test of

$$H_0^{\text{SP}} : \beta = 0 \quad \text{versus} \quad H_1^{\text{SP}} : \beta = h/\sqrt{n}.$$

Can other CRT variants be more powerful than GCM?

Consider alternatives specified by the partially linear semiparametric model

$$\mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0, \sigma^2).$$

For this model, the oracle product-of-residuals test statistic

$$T_n^{\text{oracle}}(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \mu_{n,y}(Z_i))$$

is efficient score statistic, with optimal asymptotic power against local alternatives.

Then, a test based on $T_n^{\text{oracle}}(X, Y, Z)$ is the most powerful test of

$$H_0^{\text{SP}} : \beta = 0 \quad \text{versus} \quad H_1^{\text{SP}} : \beta = h/\sqrt{n}.$$

The GCM statistic $T_n(X, Y, Z)$ is asymptotically equivalent to $T_n^{\text{oracle}}(X, Y, Z)$ under Shah and Peters's conditions, so GCM is also most powerful.

Can other CRT variants be more powerful than GCM?

Can other CRT variants be more powerful than GCM?

Testing the CI null $H_0^{\text{CI}} \cap \mathscr{R}_n$ is not the same as testing the semiparametric null

$$H_0^{\text{SP}} : \beta = 0 \text{ in the model } \mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

Can other CRT variants be more powerful than GCM?

Testing the CI null $H_0^{\text{CI}} \cap \mathscr{R}_n$ is not the same as testing the semiparametric null

$$H_0^{\text{SP}} : \beta = 0 \text{ in the model } \mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

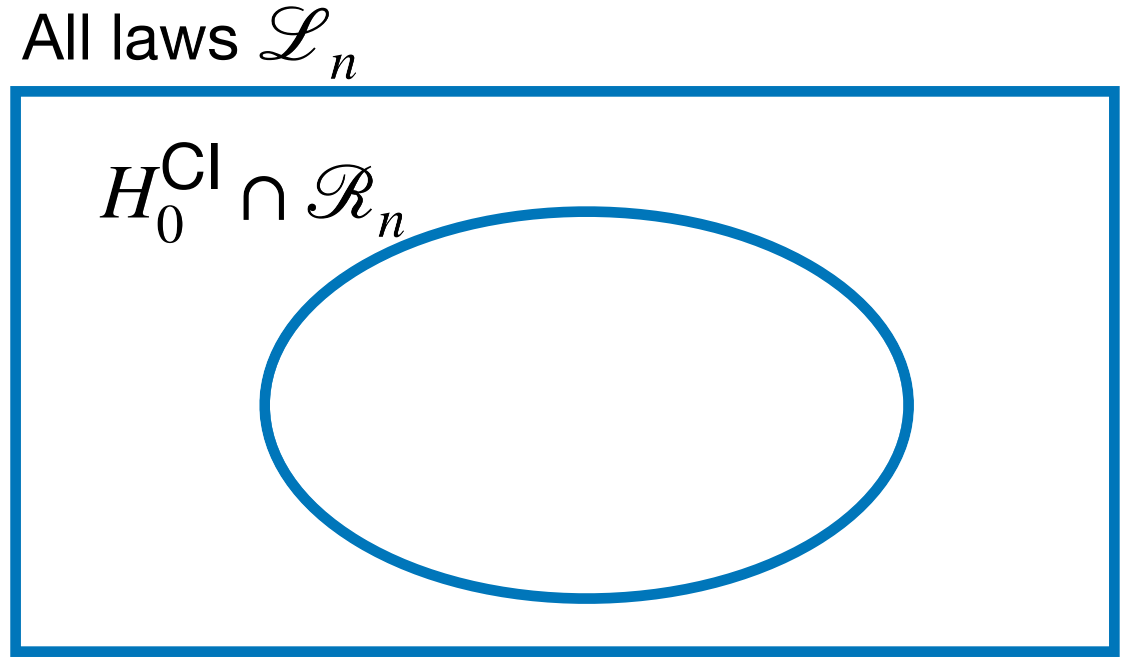
However, under certain conditions
the SP null is nested inside the CI null.

Can other CRT variants be more powerful than GCM?

Testing the CI null $H_0^{\text{CI}} \cap \mathscr{R}_n$ is not the same as testing the semiparametric null

$$H_0^{\text{SP}} : \beta = 0 \text{ in the model } \mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

However, under certain conditions
the SP null is nested inside the CI null.

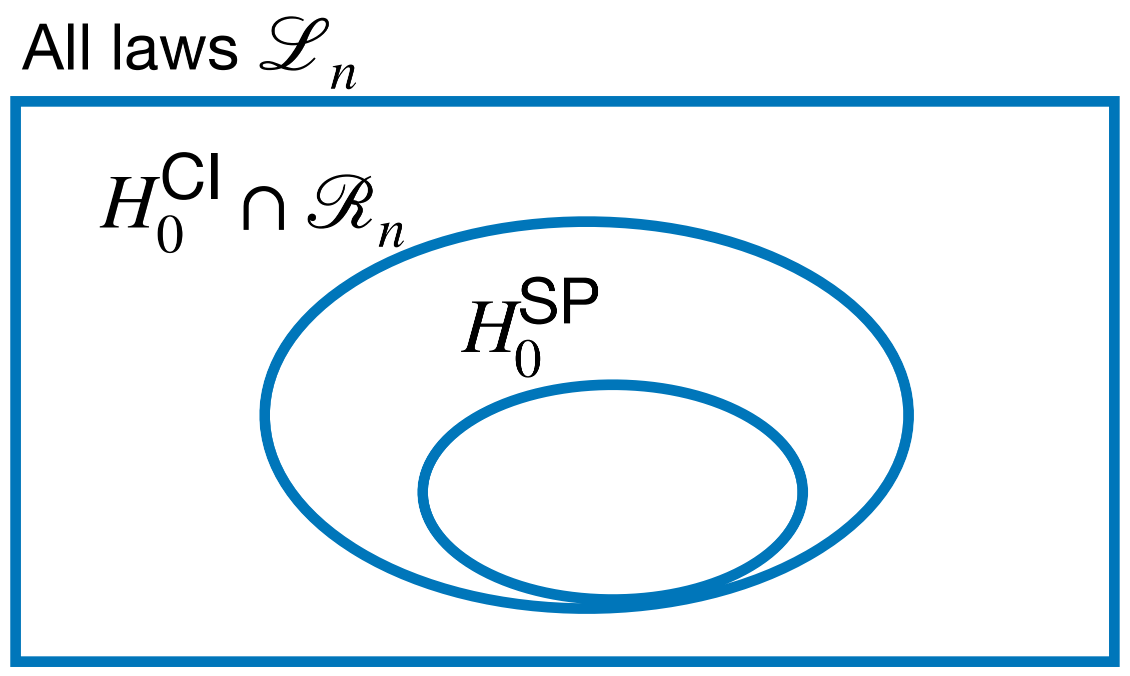


Can other CRT variants be more powerful than GCM?

Testing the CI null $H_0^{\text{CI}} \cap \mathscr{R}_n$ is not the same as testing the semiparametric null

$$H_0^{\text{SP}} : \beta = 0 \text{ in the model } \mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

However, under certain conditions
the SP null is nested inside the CI null.

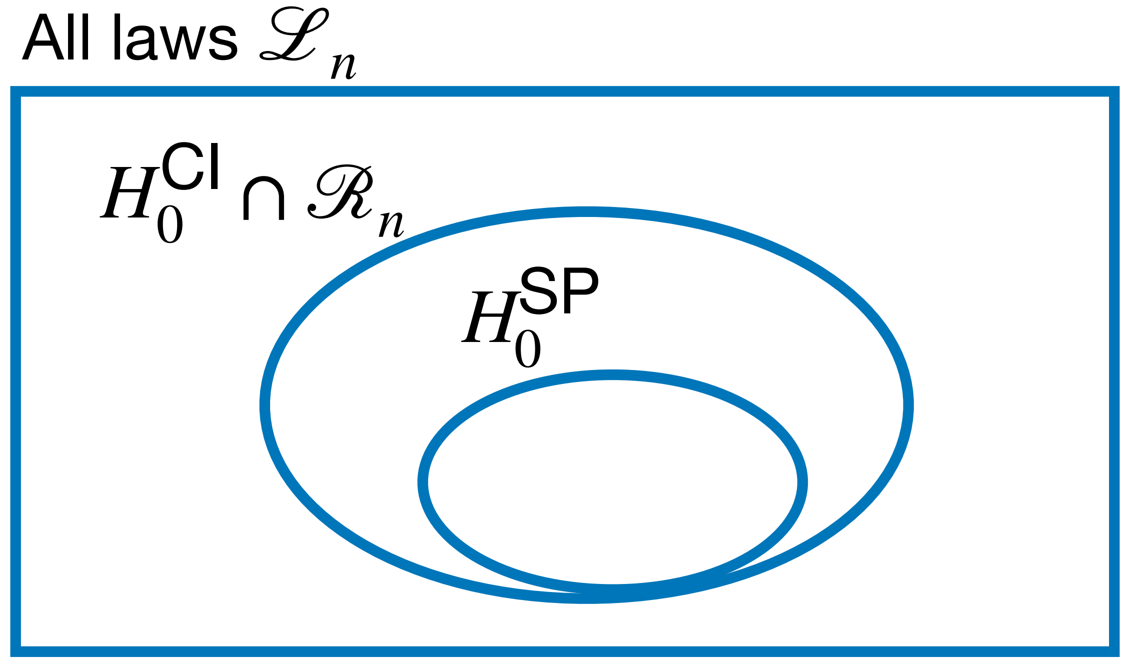


Can other CRT variants be more powerful than GCM?

Testing the CI null $H_0^{\text{CI}} \cap \mathscr{R}_n$ is not the same as testing the semiparametric null

$$H_0^{\text{SP}} : \beta = 0 \text{ in the model } \mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

However, under certain conditions
the SP null is nested inside the CI null.



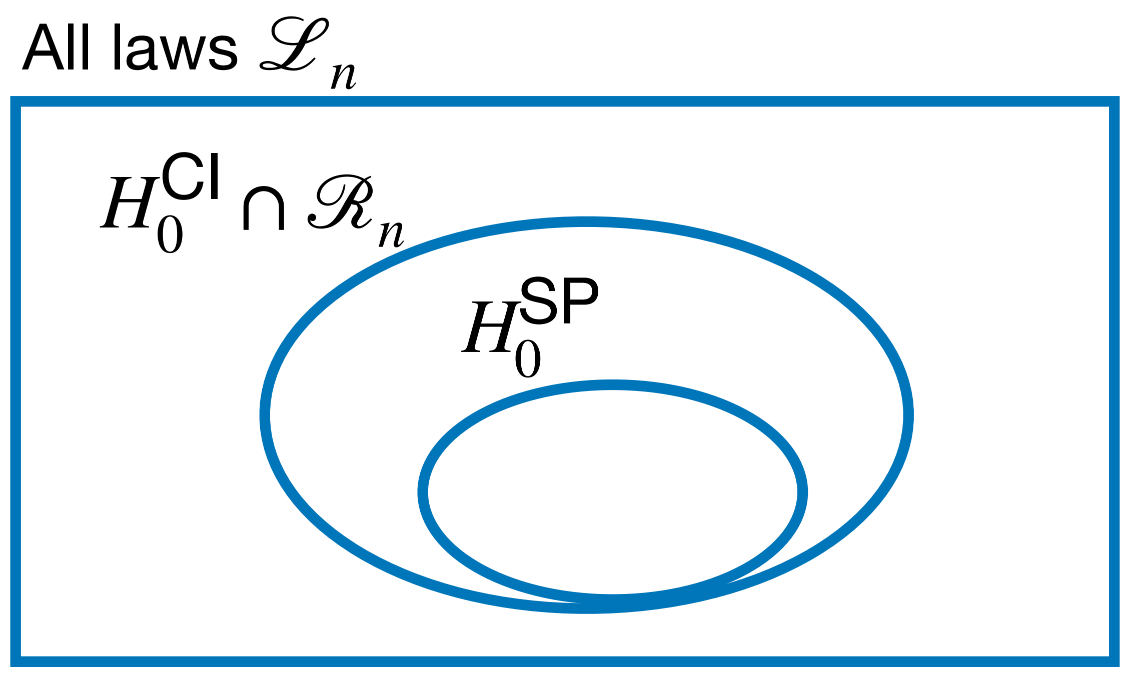
Therefore, any test controlling Type-I error on $H_0^{\text{CI}} \cap \mathscr{R}_n$ must also control Type-I error on H_0^{SP}, and so its power is bounded above by that of the GCM test.

Can other CRT variants be more powerful than GCM?

Testing the CI null $H_0^{\text{CI}} \cap \mathscr{R}_n$ is not the same as testing the semiparametric null

$$H_0^{\text{SP}} : \beta = 0 \text{ in the model } \mathbf{Y} = \mathbf{X}\beta + g(\mathbf{Z}) + \epsilon; \quad \epsilon \sim N(0,\sigma^2).$$

However, under certain conditions
the SP null is nested inside the CI null.



Therefore, any test controlling Type-I error on $H_0^{\text{CI}} \cap \mathscr{R}_n$ must also control Type-I error on H_0^{SP}, and so its power is bounded above by that of the GCM test.

Theorem (Niu et al '24; informal). Under Shah and Peters's conditions, the GCM test of $H_0^{\text{CI}} \cap \mathscr{R}_n$ is asymptotically most powerful against $H_1^{\text{SP}} : \mathbf{Y} = \mathbf{X}h/\sqrt{n} + g(\mathbf{Z}) + \epsilon$.

Can other CRT variants be more powerful than GCM?

Can other CRT variants be more powerful than GCM?

For alternatives without interactions or heteroskedasticity:

Can other CRT variants be more powerful than GCM?

For alternatives without interactions or heteroskedasticity:

- GCM is already the asymptotically most powerful test against local partially linear alternatives, so no use trying CRT with fancier test statistics.

Can other CRT variants be more powerful than GCM?

For alternatives without interactions or heteroskedasticity:

- GCM is already the asymptotically most powerful test against local partially linear alternatives, so no use trying CRT with fancier test statistics.
- Similar statement is true for *generalized* partially linear models.

Can other CRT variants be more powerful than GCM?

For alternatives without interactions or heteroskedasticity:

- GCM is already the asymptotically most powerful test against local partially linear alternatives, so no use trying CRT with fancier test statistics.
- Similar statement is true for *generalized* partially linear models.
- GCM is most powerful against local (generalized) linear model alternatives.

Can other CRT variants be more powerful than GCM?

For alternatives without interactions or heteroskedasticity:

- GCM is already the asymptotically most powerful test against local partially linear alternatives, so no use trying CRT with fancier test statistics.
- Similar statement is true for *generalized* partially linear models.
- GCM is most powerful against local (generalized) linear model alternatives.

For alternatives with interactions or heteroskedasticity:

Can other CRT variants be more powerful than GCM?

For alternatives without interactions or heteroskedasticity:

- GCM is already the asymptotically most powerful test against local partially linear alternatives, so no use trying CRT with fancier test statistics.
- Similar statement is true for *generalized* partially linear models.
- GCM is most powerful against local (generalized) linear model alternatives.

For alternatives with interactions or heteroskedasticity:

- We would not expect GCM (or dCRT) to be optimal, and alternative methods may have better power.[1]

[1]Scheidegger et al '21, Zhong et al '21, Lundborg et al. '22

Numerical simulations: Design

Numerical simulations: Design

Data-generating model:

Numerical simulations: Design

Data-generating model:

$$\mathscr{L}(\mathbf{Z}) = N(0, \Sigma(\rho)), \ \mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta, 1), \ \mathscr{L}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) = N(\mathbf{X}\theta + \mathbf{Z}^T\beta, 1),$$

Numerical simulations: Design

Data-generating model:

$$\mathscr{L}(\mathbf{Z}) = N(0,\Sigma(\rho)),\ \mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1),\ \mathscr{L}(\mathbf{Y} \mid \mathbf{X},\mathbf{Z}) = N(\mathbf{X}\theta + \mathbf{Z}^T\beta,1),$$

where $\Sigma_{j1,j2}(\rho) = \rho^{|j1-j2|}$ and $\beta_j = \begin{cases} \nu, & \text{if } j \leq s \\ 0, & \text{if } j > s \end{cases}.$

Numerical simulations: Design

Data-generating model:

$$\mathscr{L}(\mathbf{Z}) = N(0, \Sigma(\rho)),\ \mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta, 1),\ \mathscr{L}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) = N(\mathbf{X}\theta + \mathbf{Z}^T\beta, 1),$$

where $\Sigma_{j1,j2}(\rho) = \rho^{|j1-j2|}$ and $\beta_j = \begin{cases} \nu, & \text{if } j \leq s \\ 0, & \text{if } j > s \end{cases}.$

Parameters ν and θ control degree of confounding and signal strength.

Numerical simulations: Design

Data-generating model:

$$\mathscr{L}(\mathbf{Z}) = N(0, \Sigma(\rho)),\ \mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta, 1),\ \mathscr{L}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) = N(\mathbf{X}\theta + \mathbf{Z}^T\beta, 1),$$

where $\Sigma_{j1,j2}(\rho) = \rho^{|j1-j2|}$ and $\beta_j = \begin{cases} \nu, & \text{if } j \leq s \\ 0, & \text{if } j > s \end{cases}.$

Parameters ν and θ control degree of confounding and signal strength.

Methods compared:

Numerical simulations: Design

Data-generating model:

$$\mathscr{L}(\mathbf{Z}) = N(0,\Sigma(\rho)),\ \mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1),\ \mathscr{L}(\mathbf{Y} \mid \mathbf{X},\mathbf{Z}) = N(\mathbf{X}\theta + \mathbf{Z}^T\beta,1),$$

where $\Sigma_{j1,j2}(\rho) = \rho^{|j1-j2|}$ and $\beta_j = \begin{cases} \nu, & \text{if } j \leq s \\ 0, & \text{if } j > s \end{cases}.$

Parameters ν and θ control degree of confounding and signal strength.

Methods compared:

- dCRT and GCM (with lasso and post-lasso)

Numerical simulations: Design

Data-generating model:

$$\mathscr{L}(\mathbf{Z}) = N(0,\Sigma(\rho)),\ \mathscr{L}(\mathbf{X} \mid \mathbf{Z}) = N(\mathbf{Z}^T\beta,1),\ \mathscr{L}(\mathbf{Y} \mid \mathbf{X},\mathbf{Z}) = N(\mathbf{X}\theta + \mathbf{Z}^T\beta,1),$$

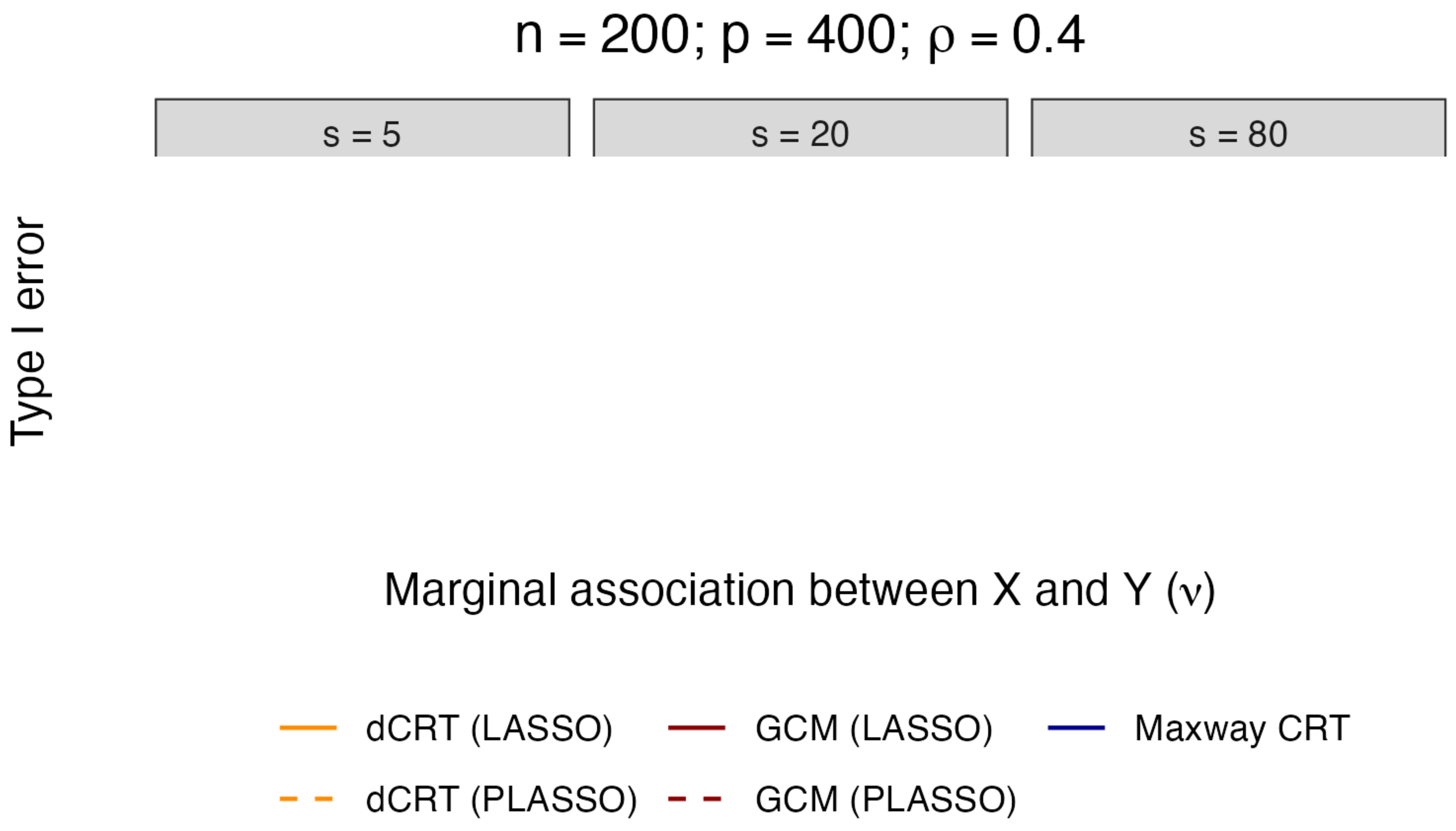
where $\Sigma_{j1,j2}(\rho) = \rho^{|j1-j2|}$ and $\beta_j = \begin{cases} \nu, & \text{if } j \leq s \\ 0, & \text{if } j > s \end{cases}.$

Parameters ν and θ control degree of confounding and signal strength.

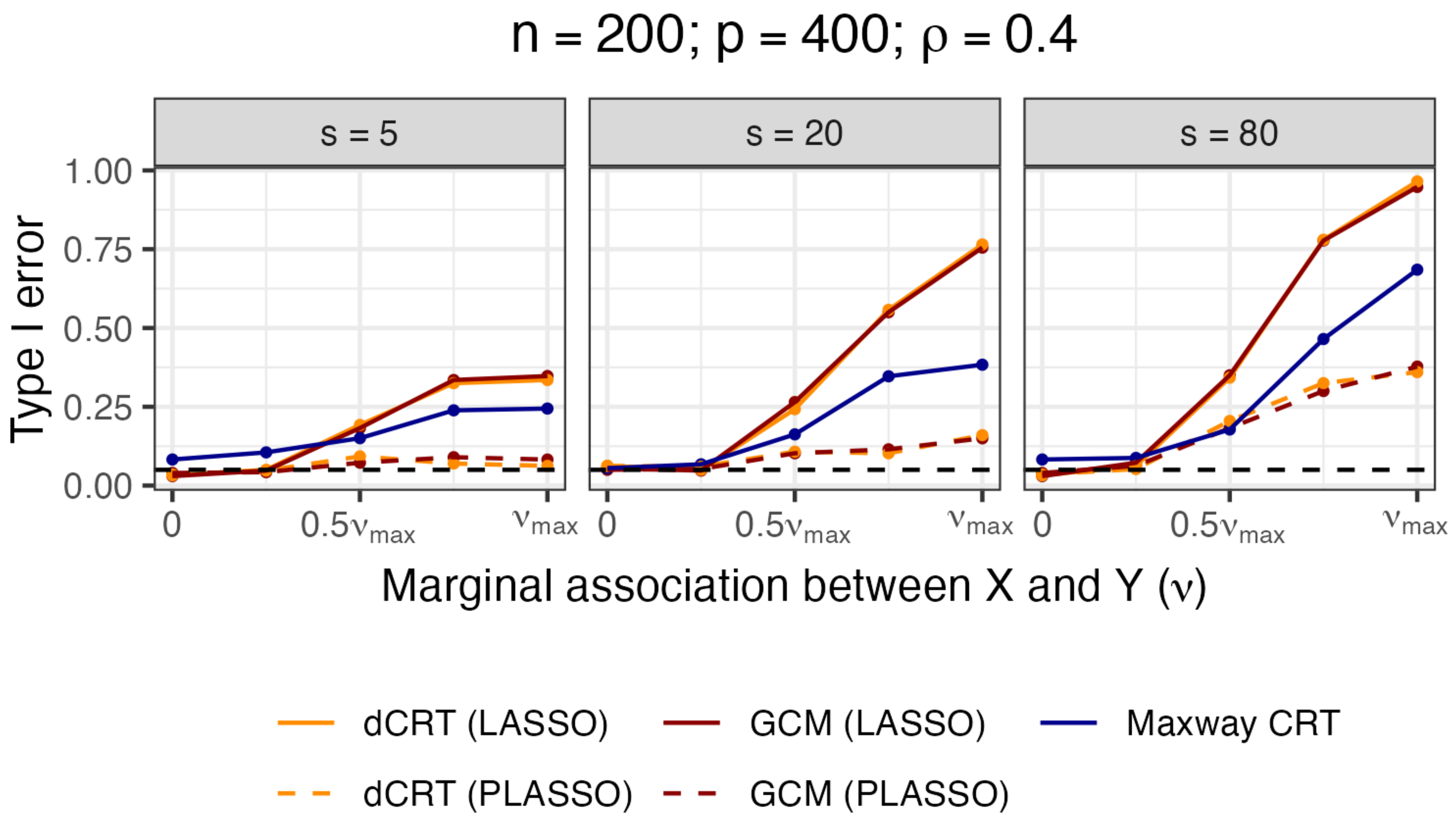
Methods compared:

- dCRT and GCM (with lasso and post-lasso)
- Maxway CRT (implemented with data splitting)

Numerical simulations: Type-I error control

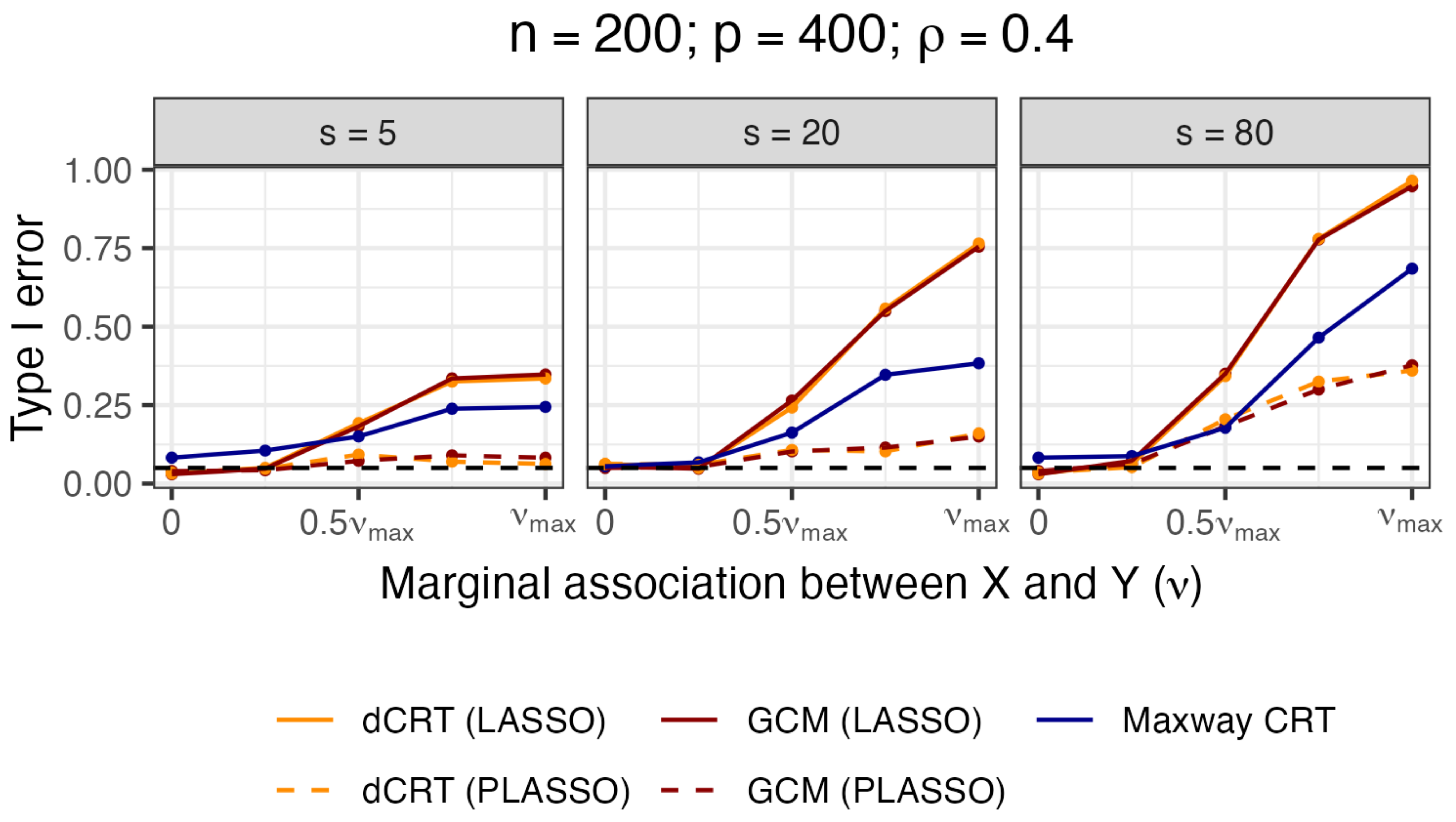


Numerical simulations: Type-I error control

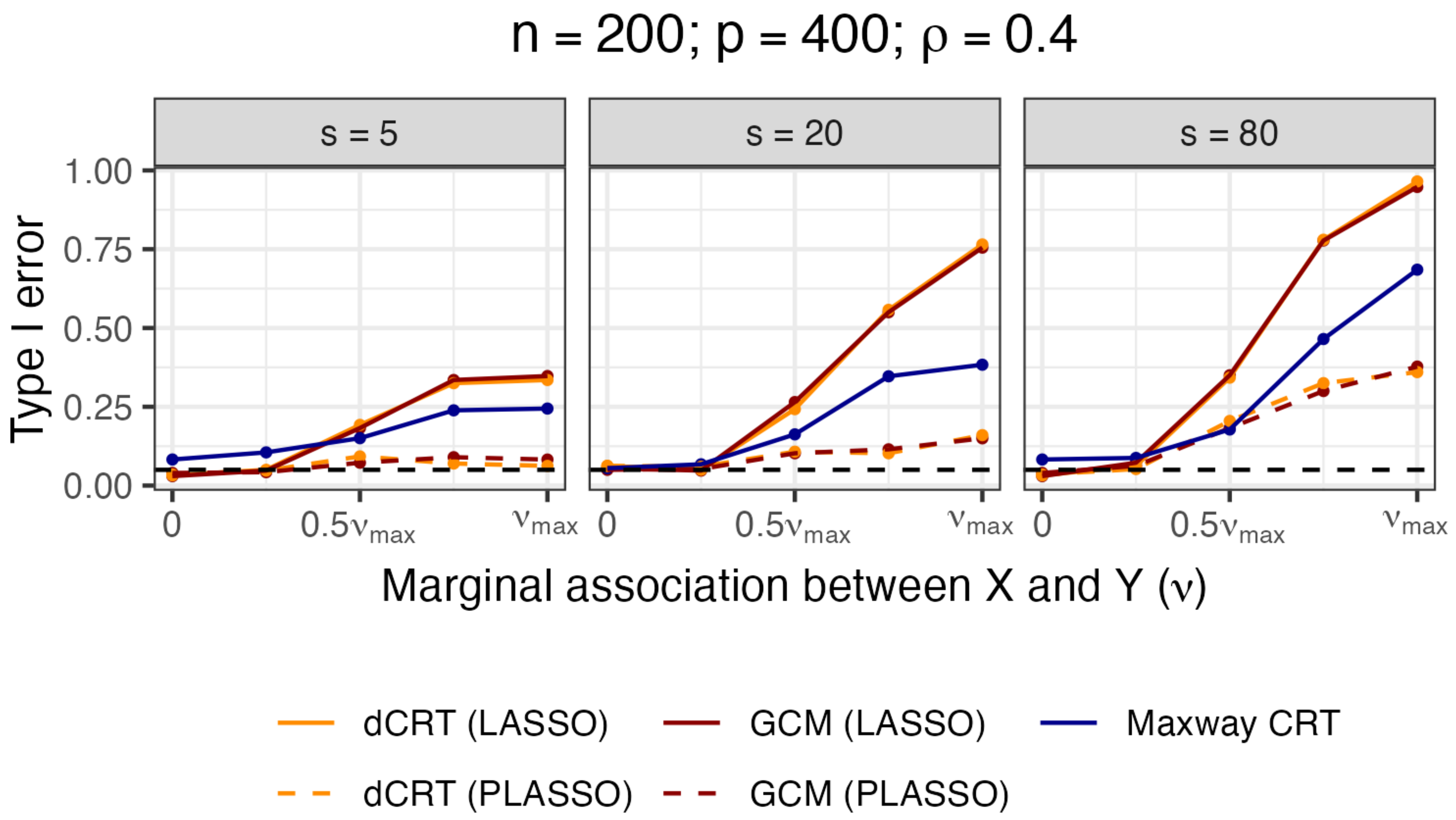


Numerical simulations: Type-I error control

Takeaways



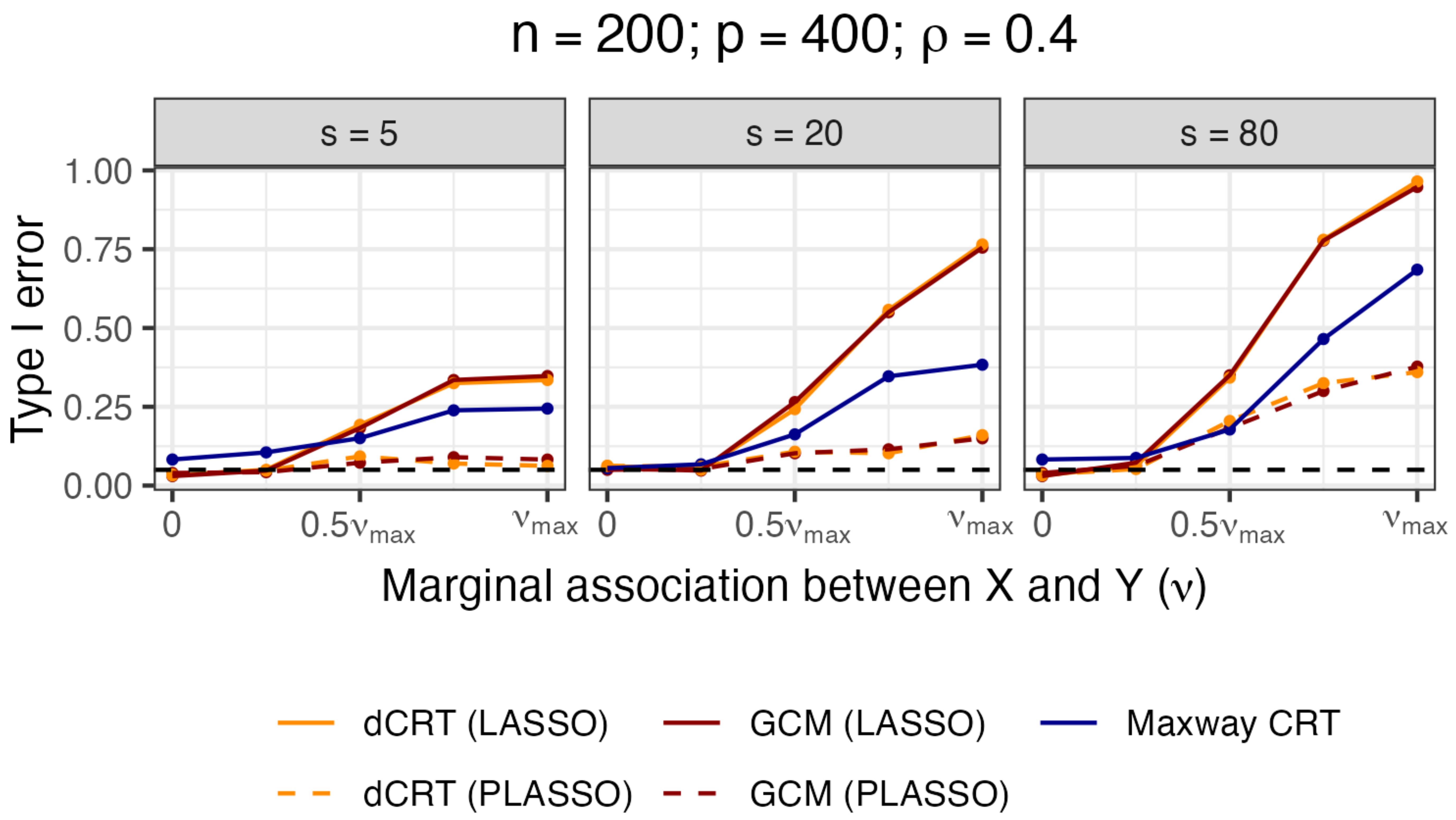
Numerical simulations: Type-I error control



Takeaways

- GCM and dCRT perform similarly, consistent with asymptotic theory.

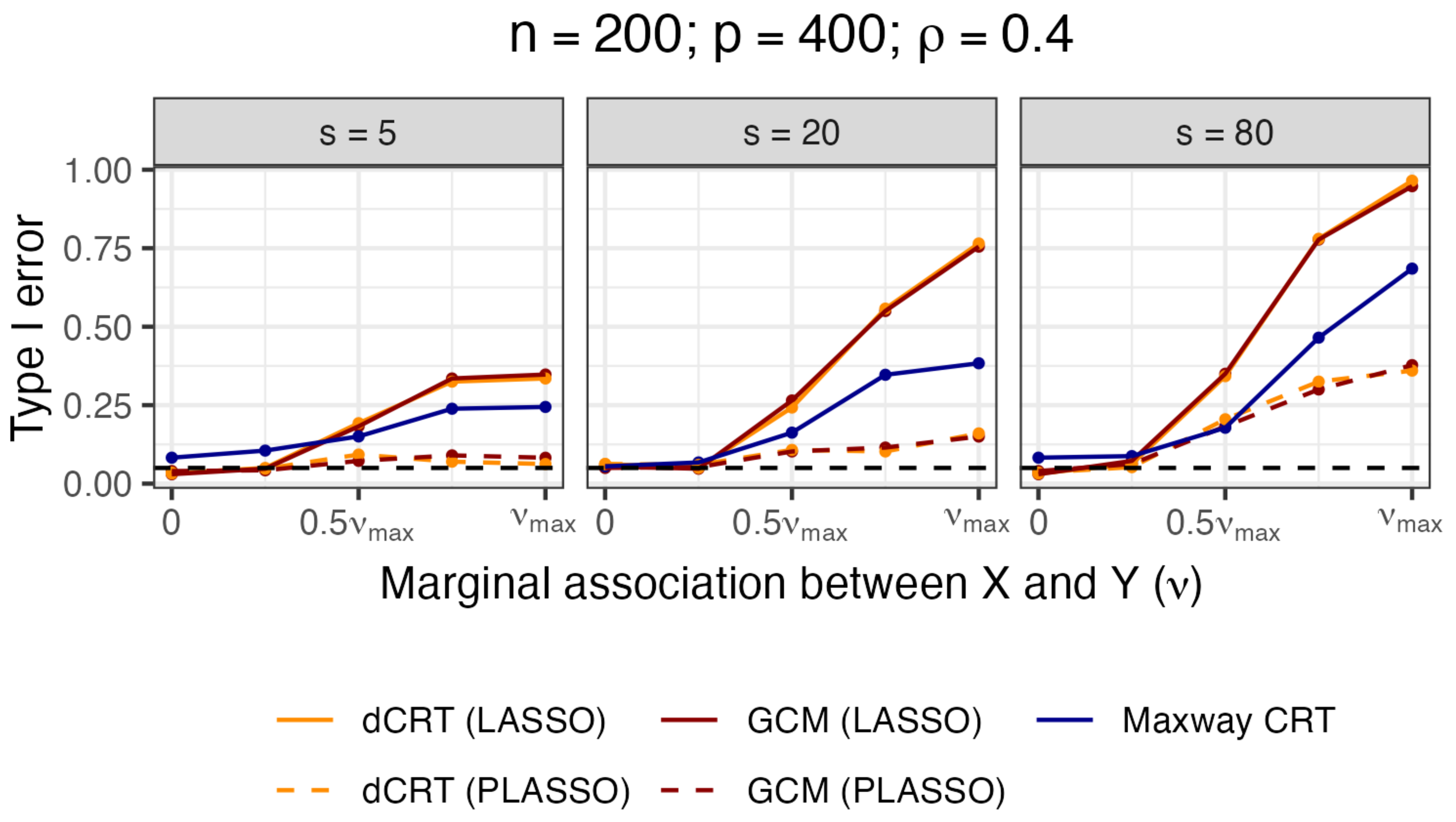
Numerical simulations: Type-I error control



Takeaways

- GCM and dCRT perform similarly, consistent with asymptotic theory.
- Lasso-based methods can have very inflated Type-I error in difficult settings.

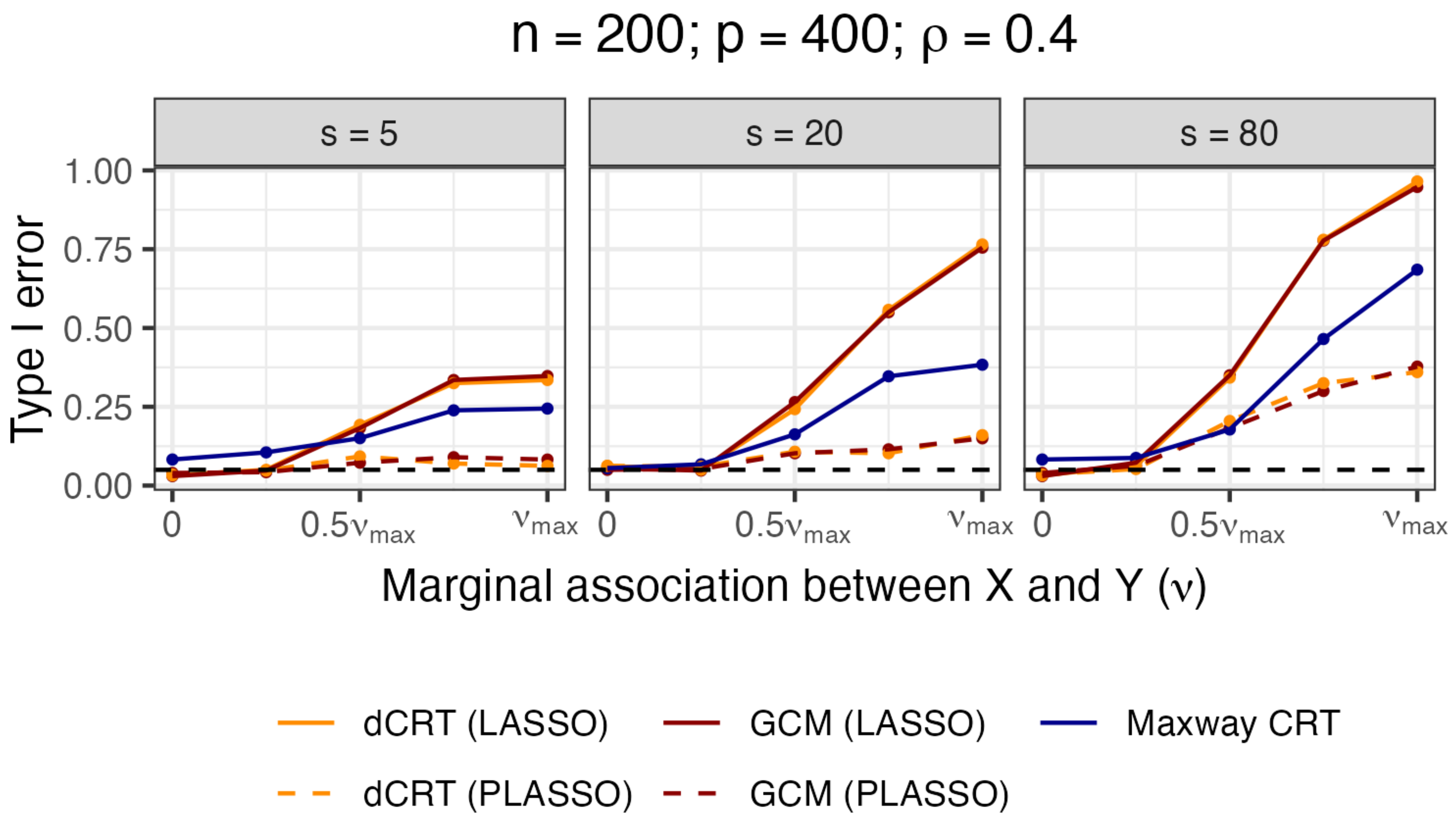
Numerical simulations: Type-I error control



Takeaways

- GCM and dCRT perform similarly, consistent with asymptotic theory.
- Lasso-based methods can have very inflated Type-I error in difficult settings.
- Maxway performs better than lasso-based dCRT and GCM, consistent with results of Li and Liu ’22.

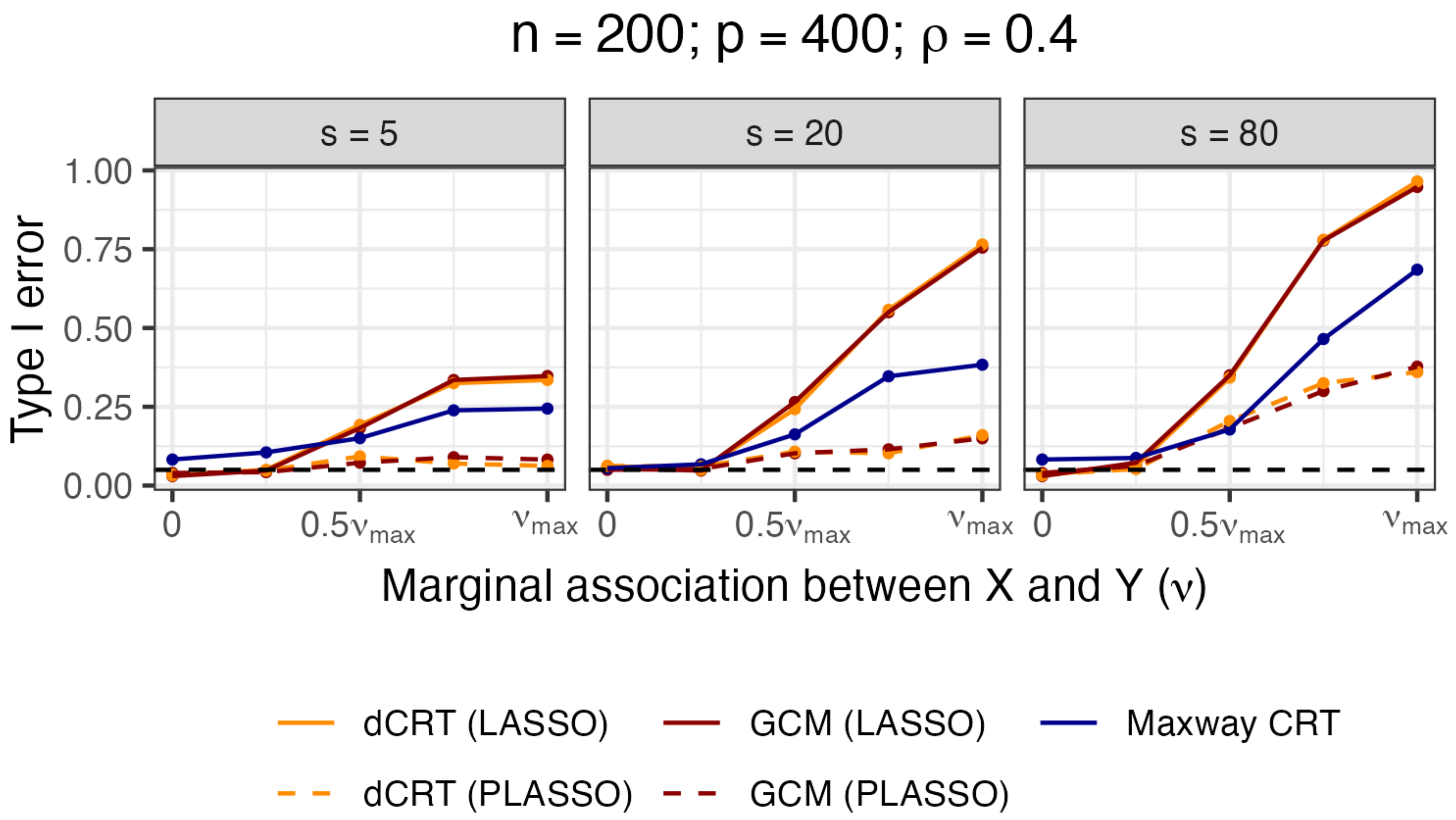
Numerical simulations: Type-I error control



Takeaways

- GCM and dCRT perform similarly, consistent with asymptotic theory.
- Lasso-based methods can have very inflated Type-I error in difficult settings.
- Maxway performs better than lasso-based dCRT and GCM, consistent with results of Li and Liu '22.
- Post-lasso-based dCRT and GCM typically outperform Maxway CRT.

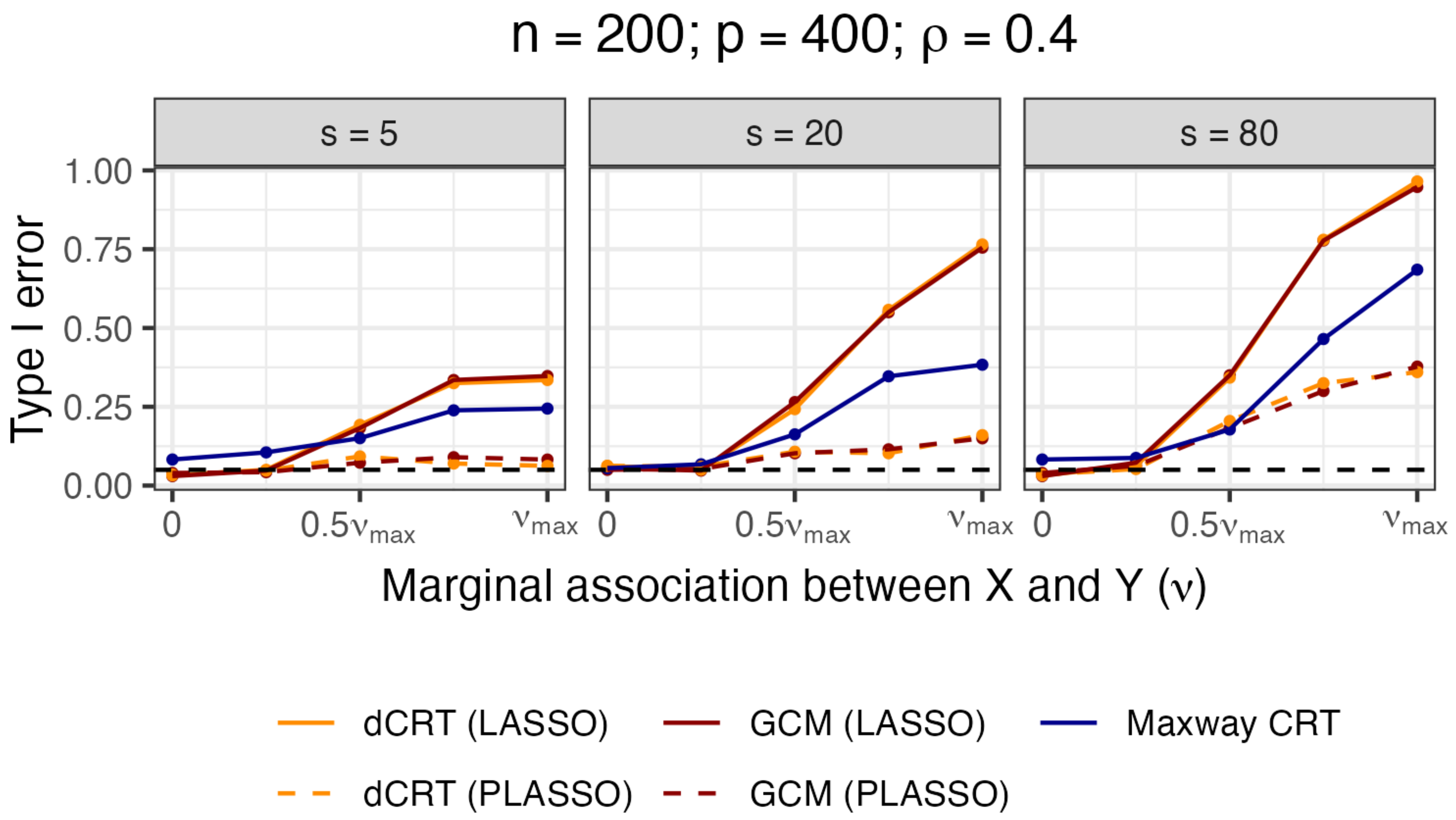
Numerical simulations: Type-I error control



Takeaways

- GCM and dCRT perform similarly, consistent with asymptotic theory.
- Lasso-based methods can have very inflated Type-I error in difficult settings.
- Maxway performs better than lasso-based dCRT and GCM, consistent with results of Li and Liu ’22.
- Post-lasso-based dCRT and GCM typically outperform Maxway CRT.
- Best-performing post-lasso methods break down as problem becomes too difficult, i.e. S & P’s conditions violated.

Numerical simulations: Type-I error control



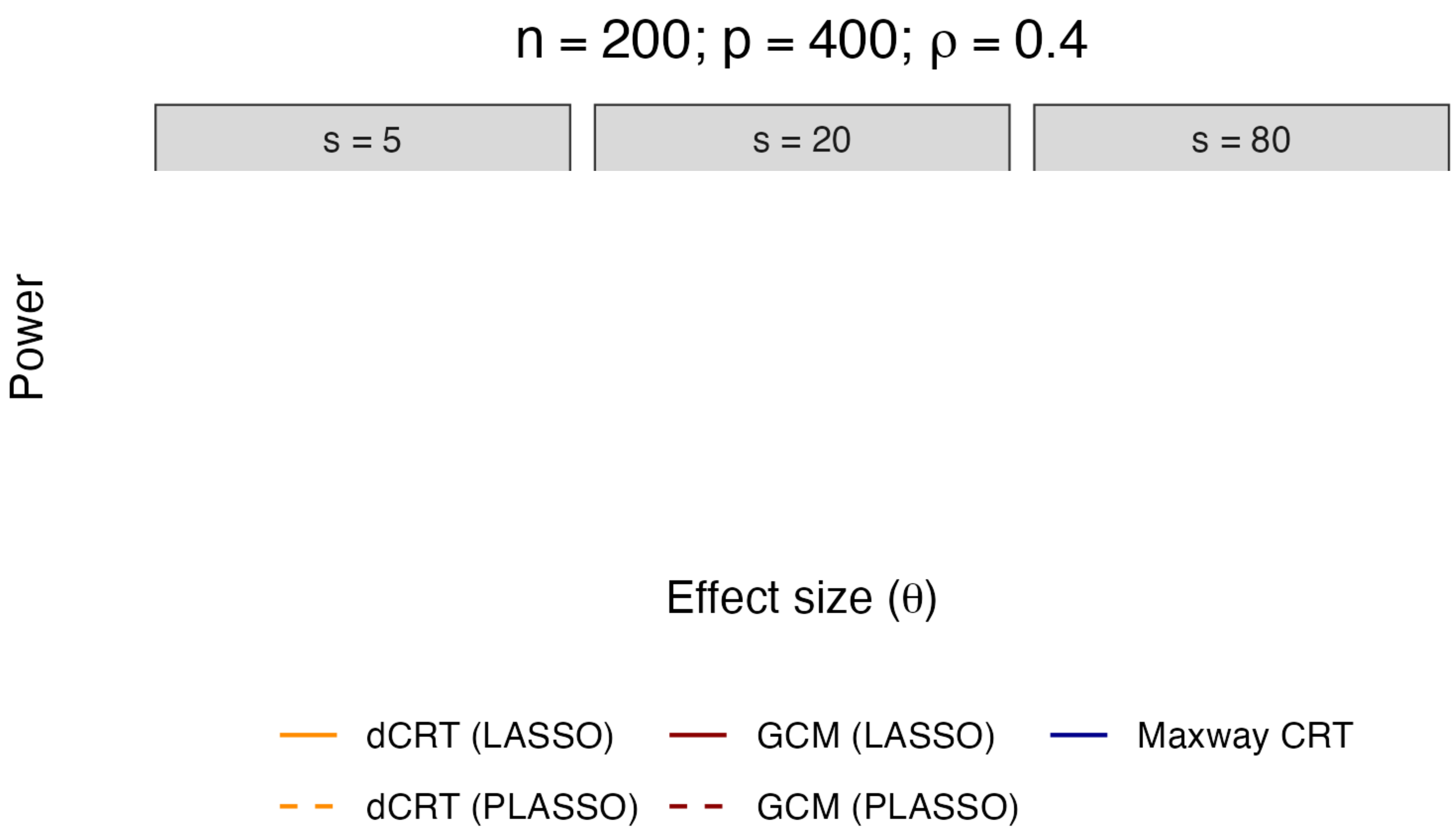
Takeaways

- GCM and dCRT perform similarly, consistent with asymptotic theory.
- Lasso-based methods can have very inflated Type-I error in difficult settings.
- Maxway performs better than lasso-based dCRT and GCM, consistent with results of Li and Liu ’22.
- Post-lasso-based dCRT and GCM typically outperform Maxway CRT.
- Best-performing post-lasso methods break down as problem becomes too difficult, i.e. S & P’s conditions violated.

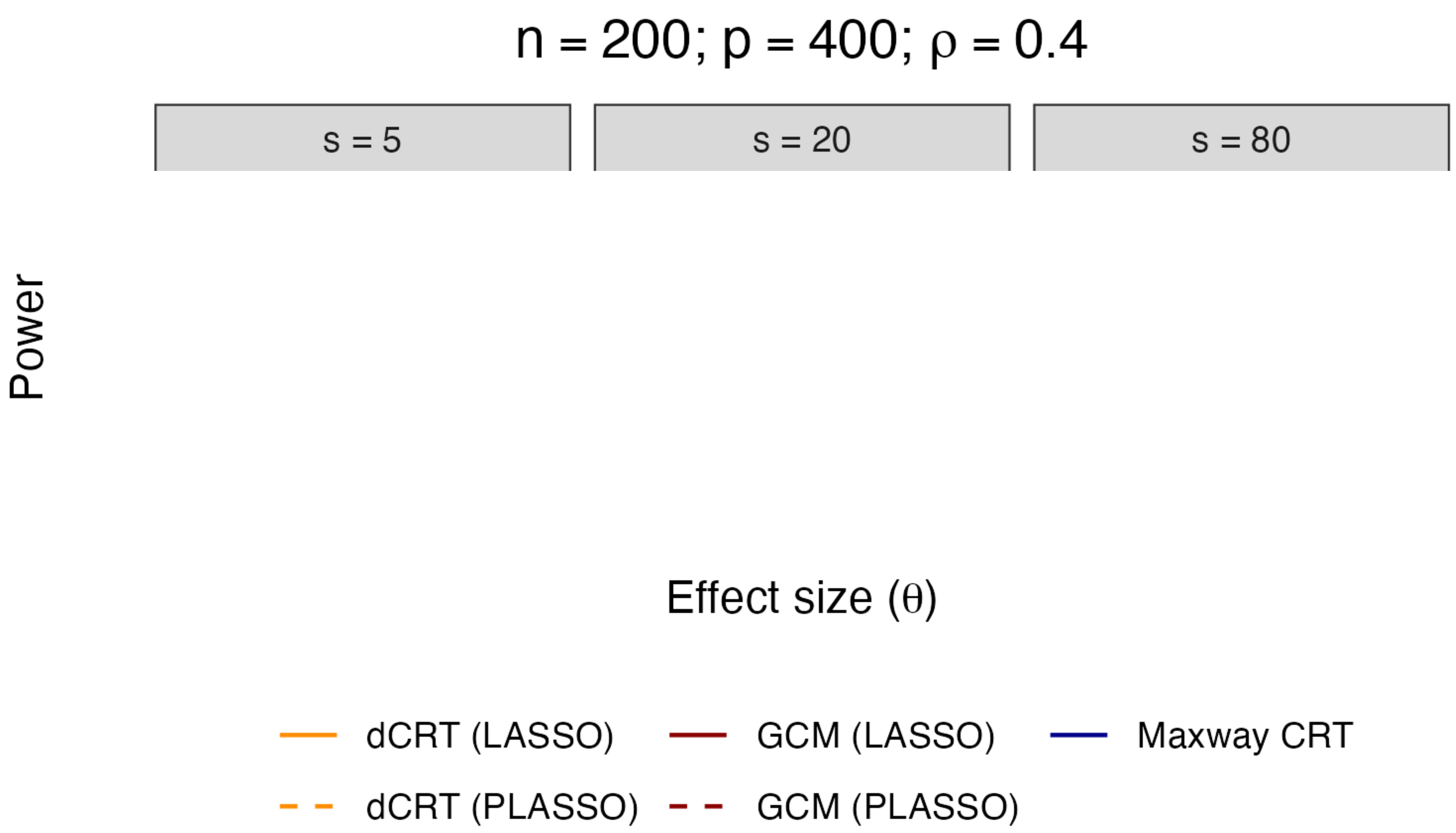
Remark

- We expect, for smaller samples sizes or more discrete data, that dCRT can have better Type-I error control than GCM.

Numerical simulations: Power

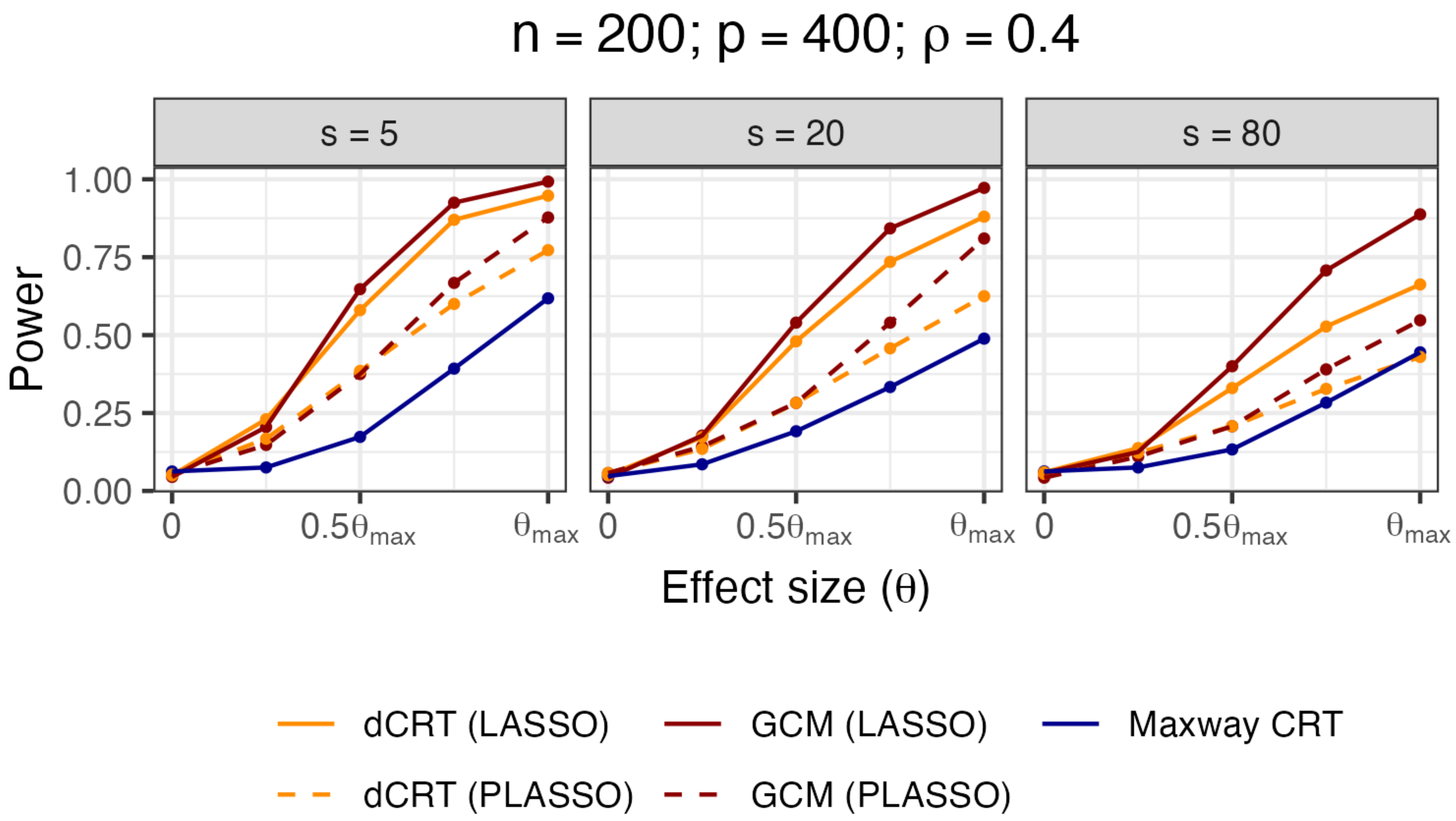


Numerical simulations: Power



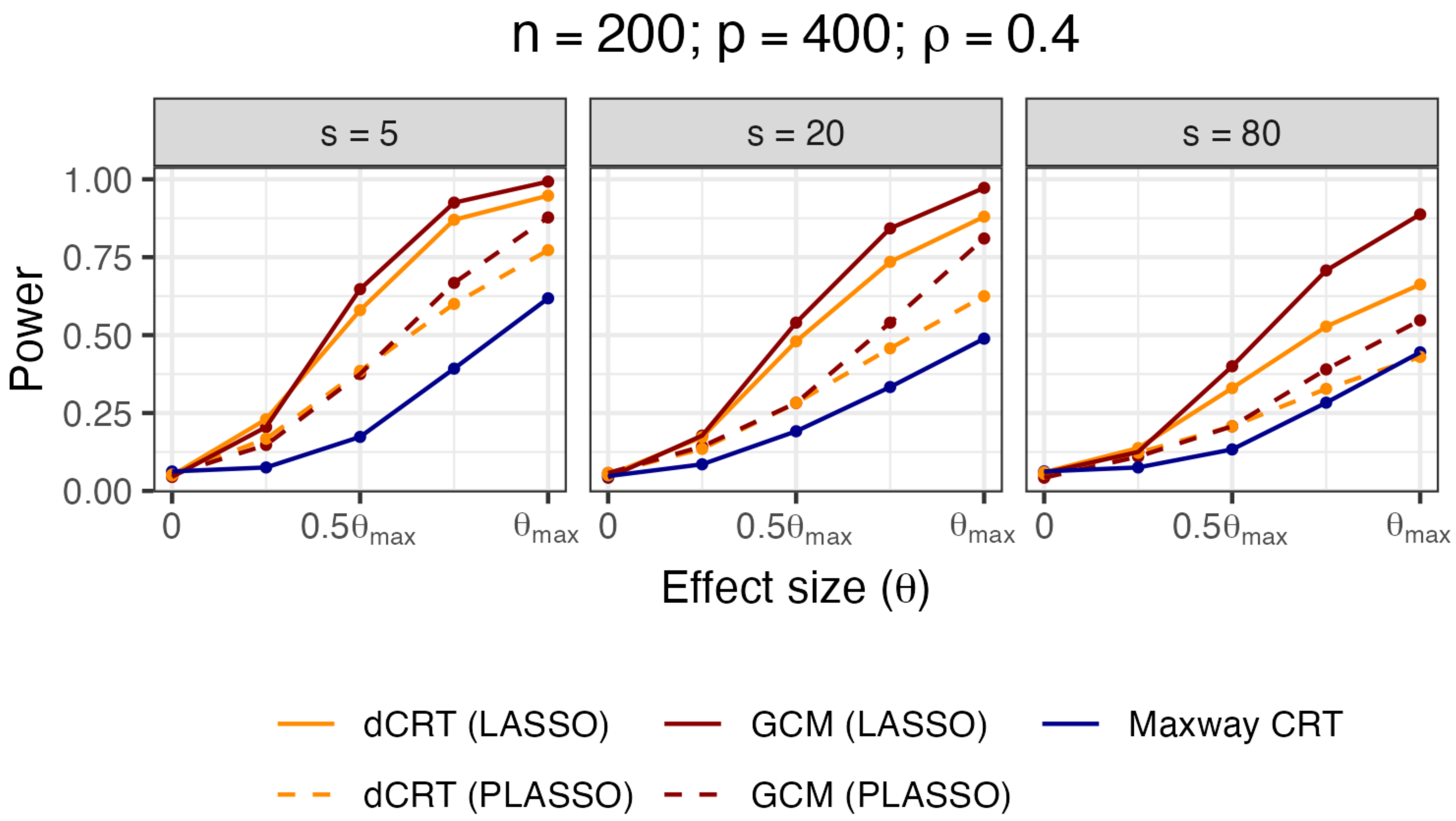
Note: All methods subjected to “oracle calibration” for fair power comparison.

Numerical simulations: Power



Note: All methods subjected to “oracle calibration” for fair power comparison.

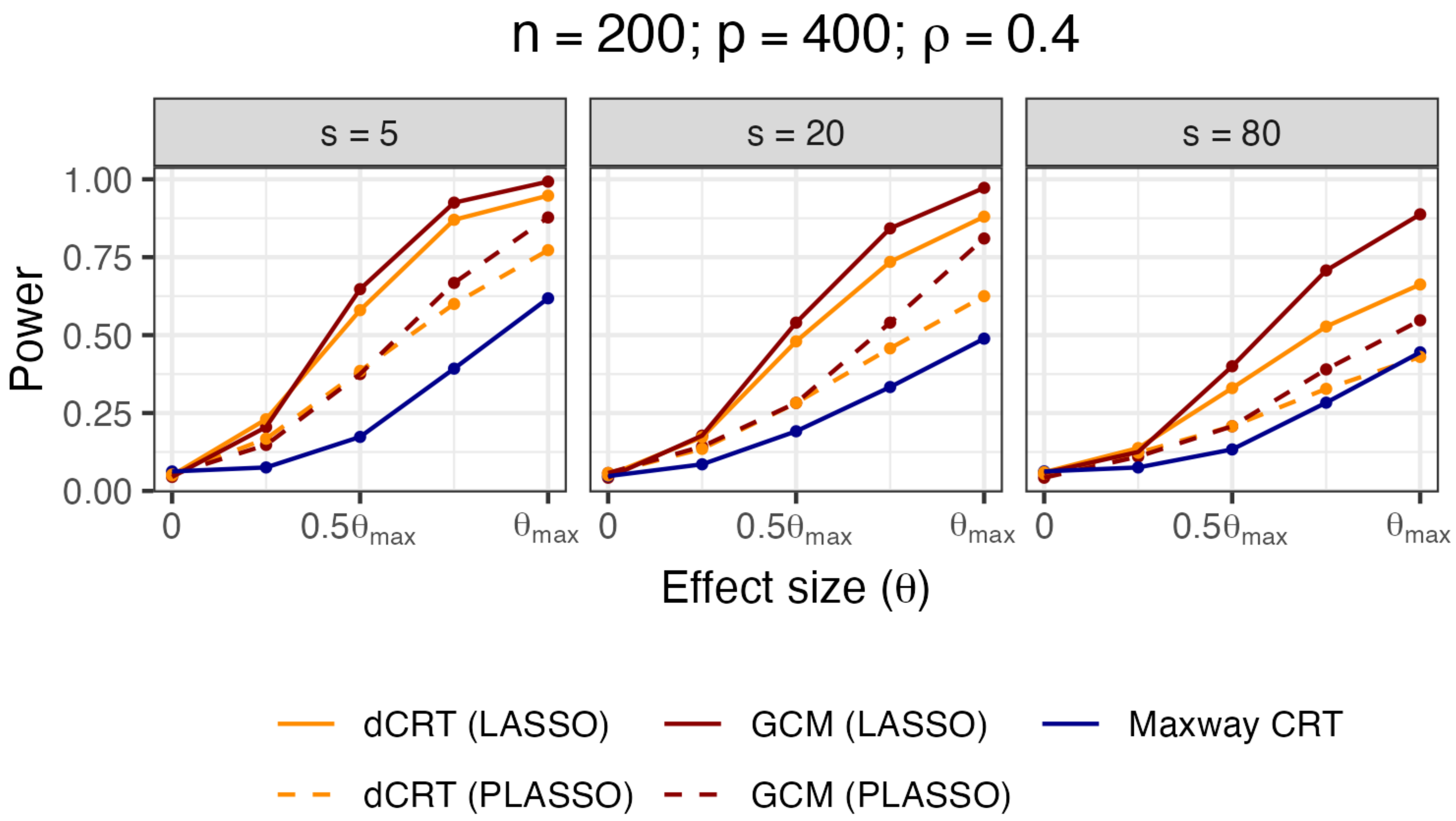
Numerical simulations: Power



Note: All methods subjected to “oracle calibration” for fair power comparison.

Takeaways

Numerical simulations: Power

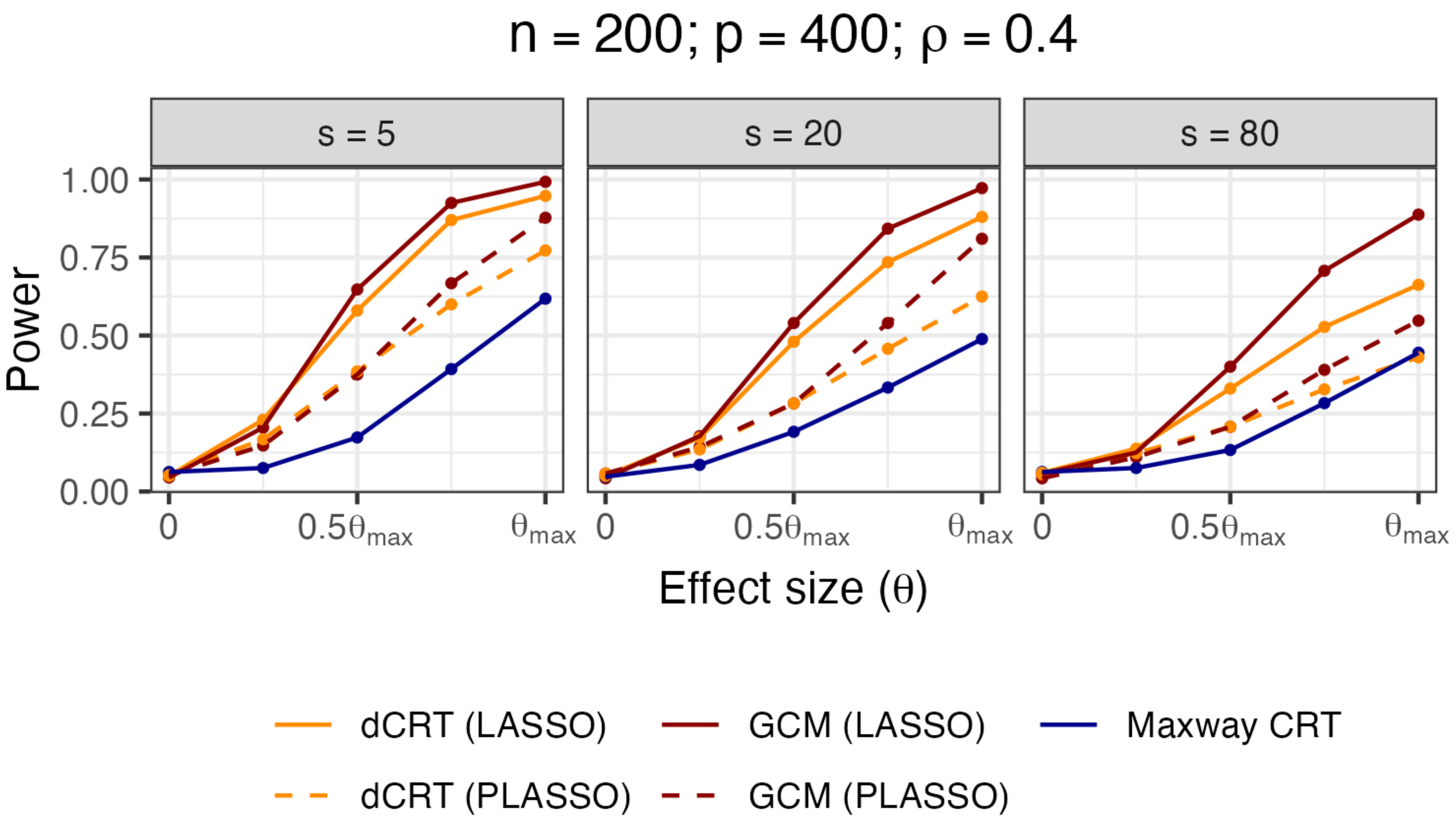


Note: All methods subjected to “oracle calibration” for fair power comparison.

Takeaways

- GCM tends to outperform dCRT.

Numerical simulations: Power

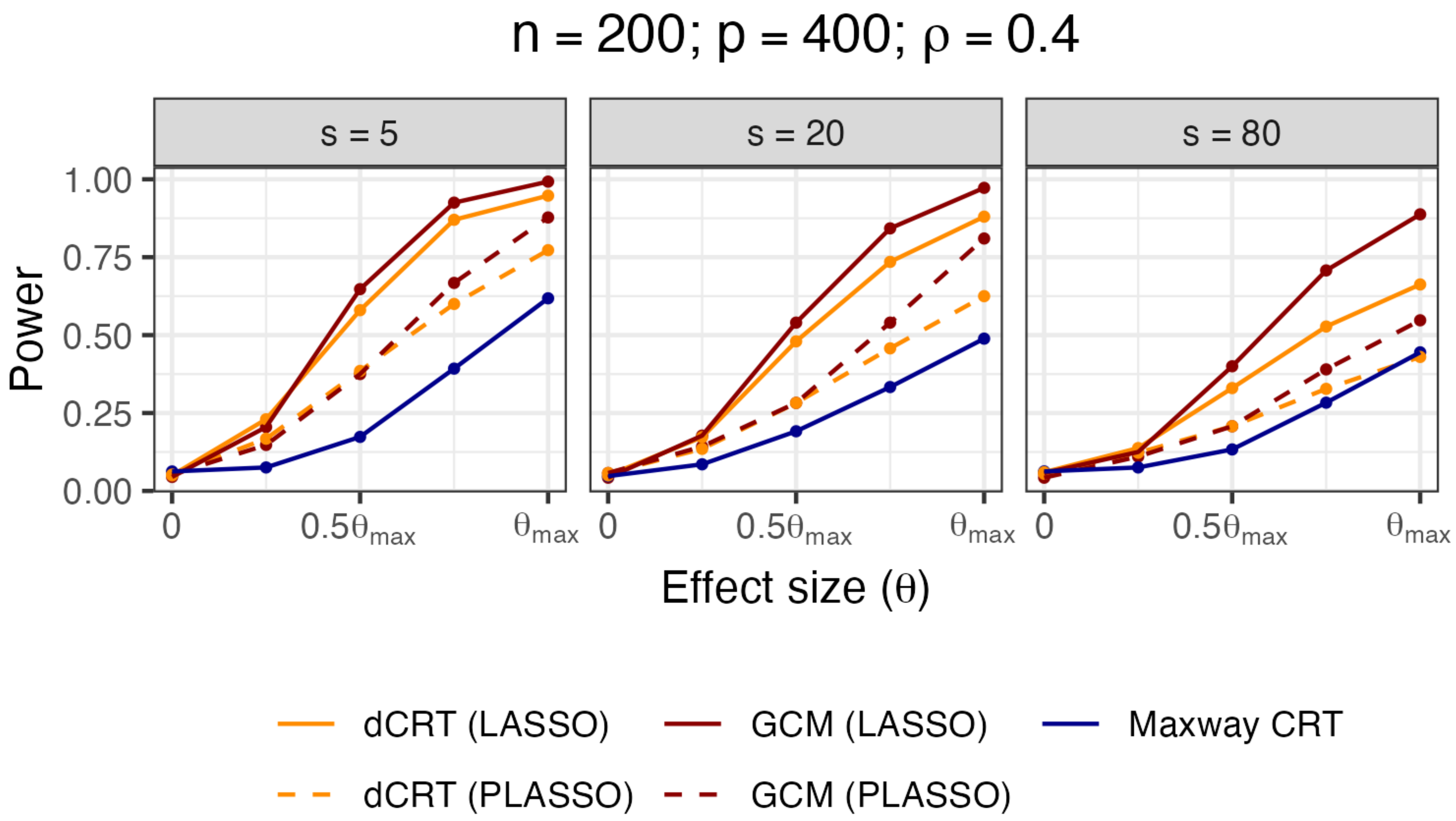


Note: All methods subjected to “oracle calibration” for fair power comparison.

Takeaways

- GCM tends to outperform dCRT.
- Lasso outperforms post-lasso, suggesting bias-variance trade-off.

Numerical simulations: Power



Note: All methods subjected to “oracle calibration” for fair power comparison.

Takeaways

- GCM tends to outperform dCRT.
- Lasso outperforms post-lasso, suggesting bias-variance trade-off.
- Maxway CRT has lowest power, due to data splitting. Better performance in separate semi-supervised simulation.