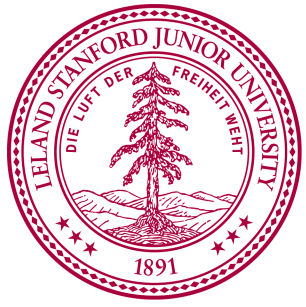
Sailing Through Data: Discoveries and Mirages

Emmanuel Candès, *Stanford University*



2018 Machine Learning Summer School, Buenos Aires, June 2018

Robustness

Robustness

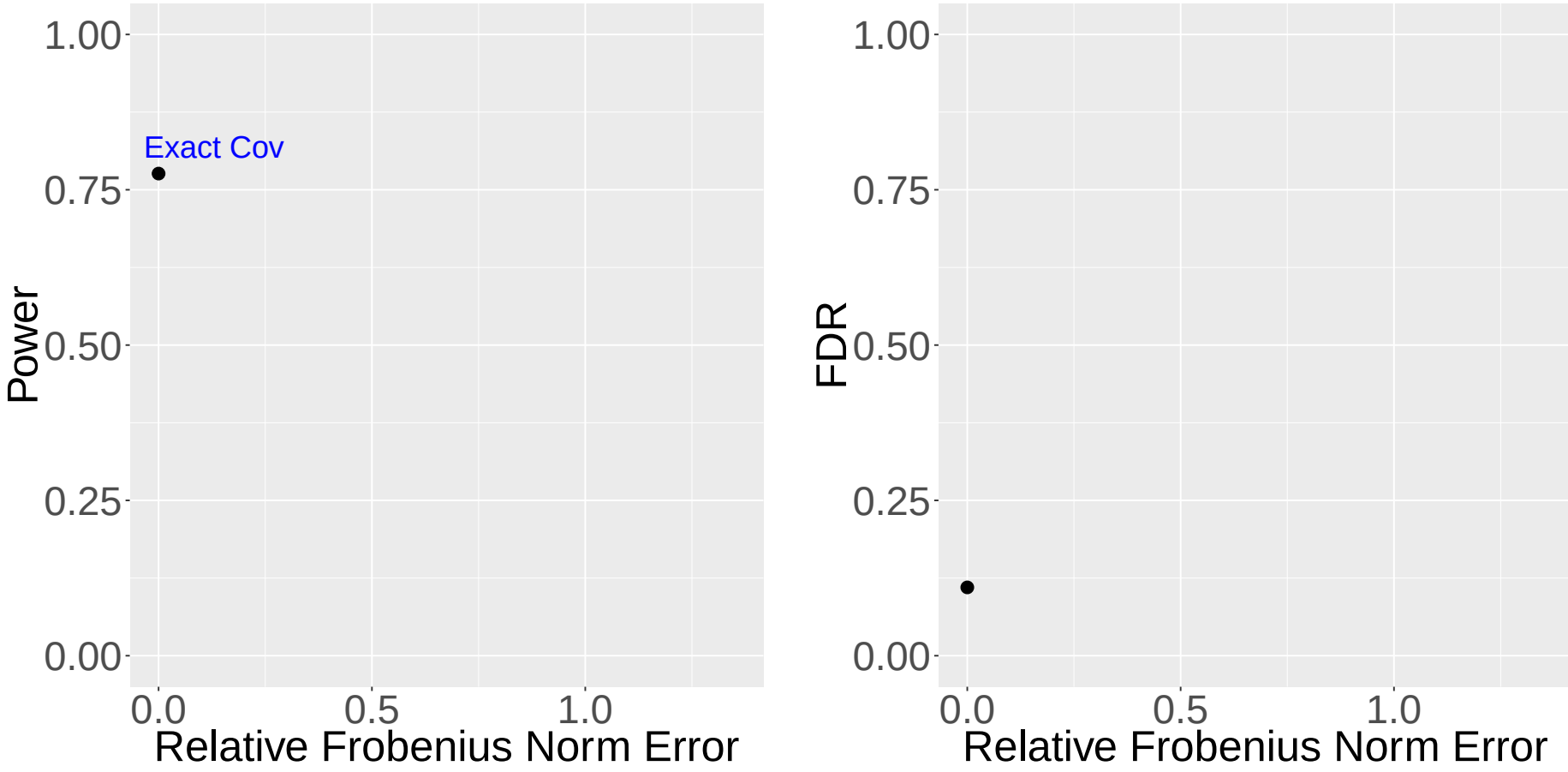


Figure: Covariates are AR(1) with autocorrelation coefficient 0.3. $n = 800$, $p = 1500$, and target FDR is 10%. $Y \mid X$ follows logistic model with 50 nonzero entries

Robustness

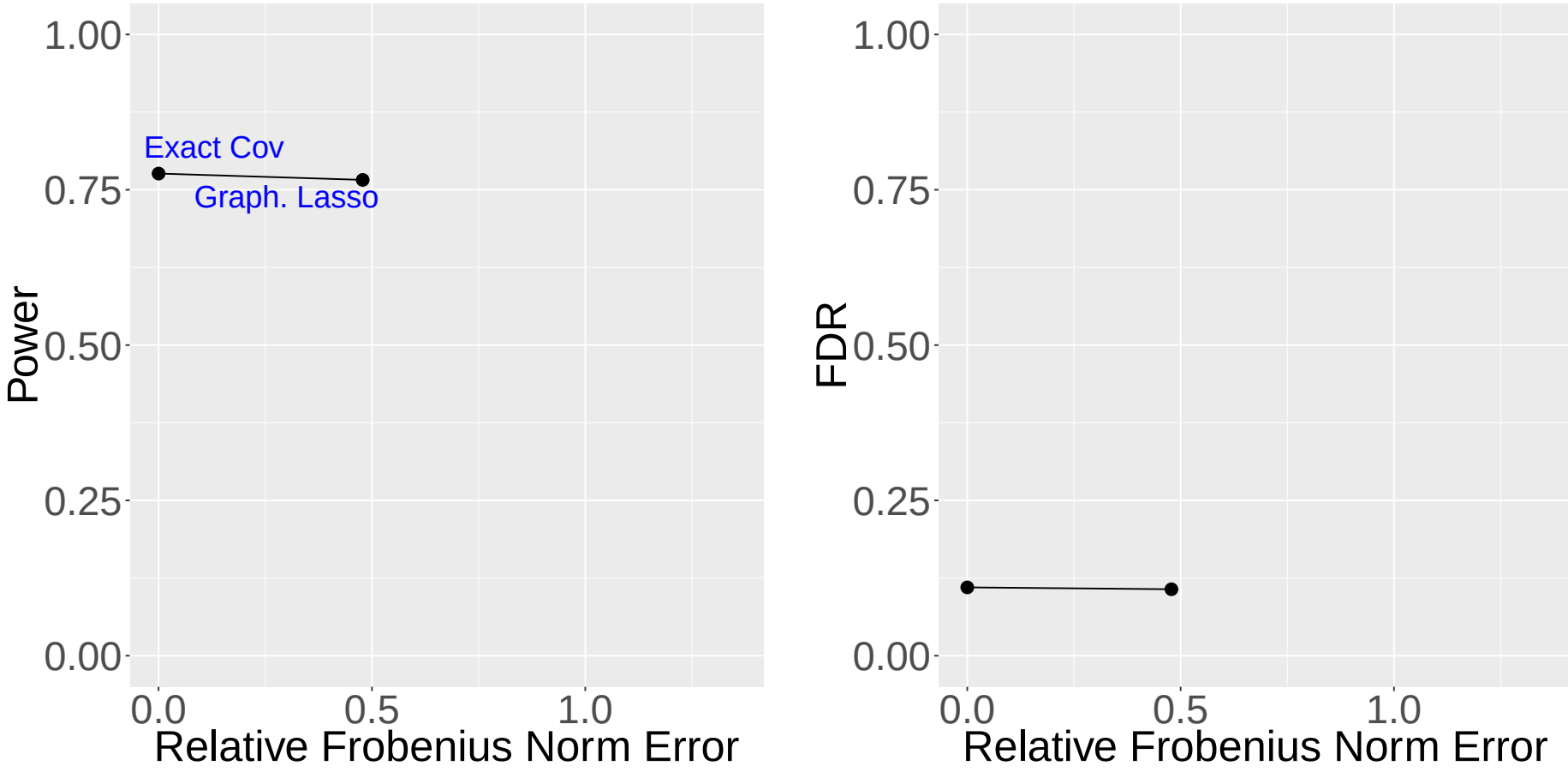


Figure: Covariates are AR(1) with autocorrelation coefficient 0.3. $n = 800$, $p = 1500$, and target FDR is 10%. $Y \mid X$ follows logistic model with 50 nonzero entries

Robustness

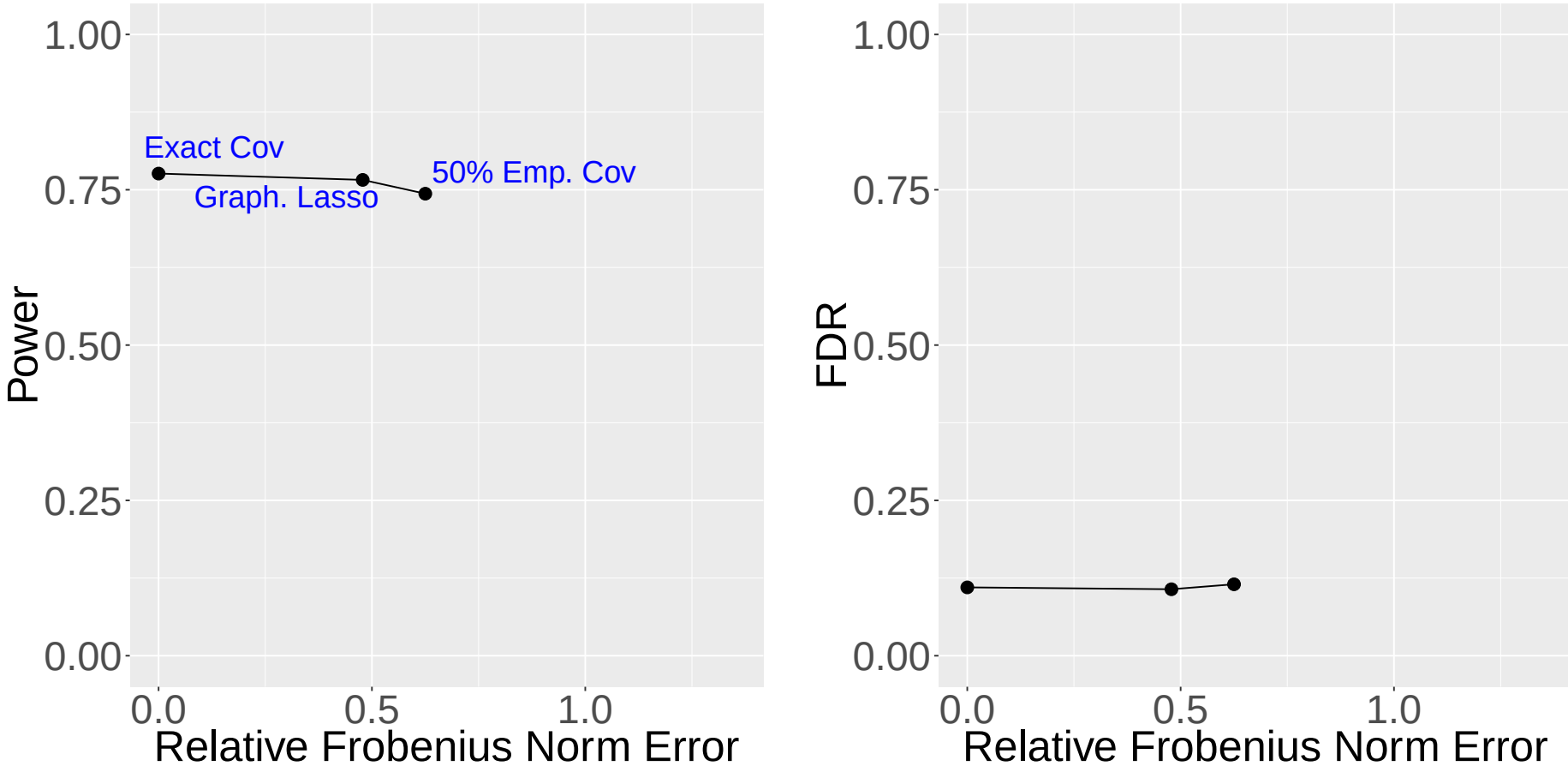


Figure: Covariates are AR(1) with autocorrelation coefficient 0.3. $n = 800$, $p = 1500$, and target FDR is 10%. $Y \mid X$ follows logistic model with 50 nonzero entries

Robustness

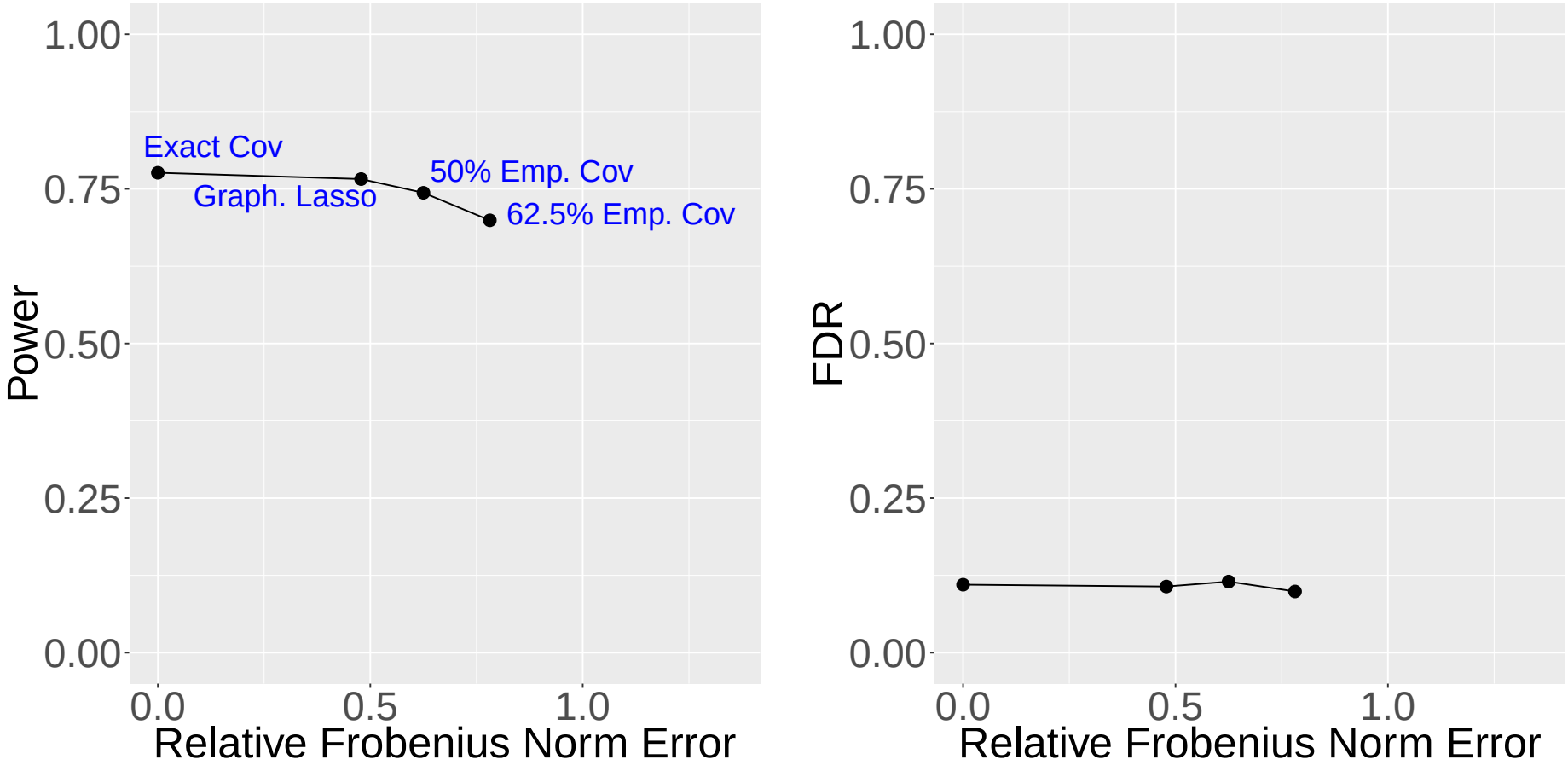


Figure: Covariates are AR(1) with autocorrelation coefficient 0.3. $n = 800$, $p = 1500$, and target FDR is 10%. $Y \mid X$ follows logistic model with 50 nonzero entries

Robustness

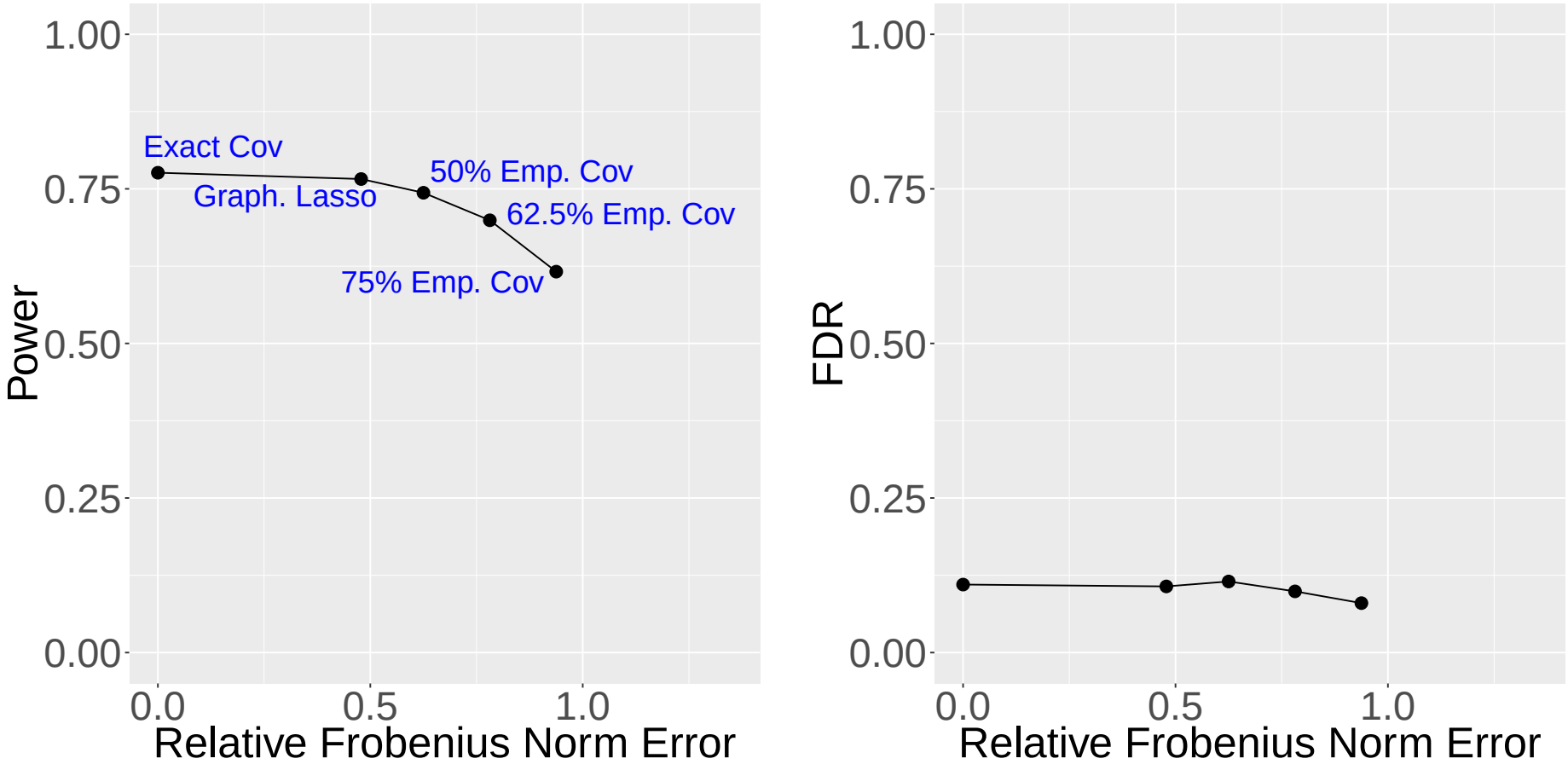


Figure: Covariates are AR(1) with autocorrelation coefficient 0.3. $n = 800$, $p = 1500$, and target FDR is 10%. $Y \mid X$ follows logistic model with 50 nonzero entries

Robustness

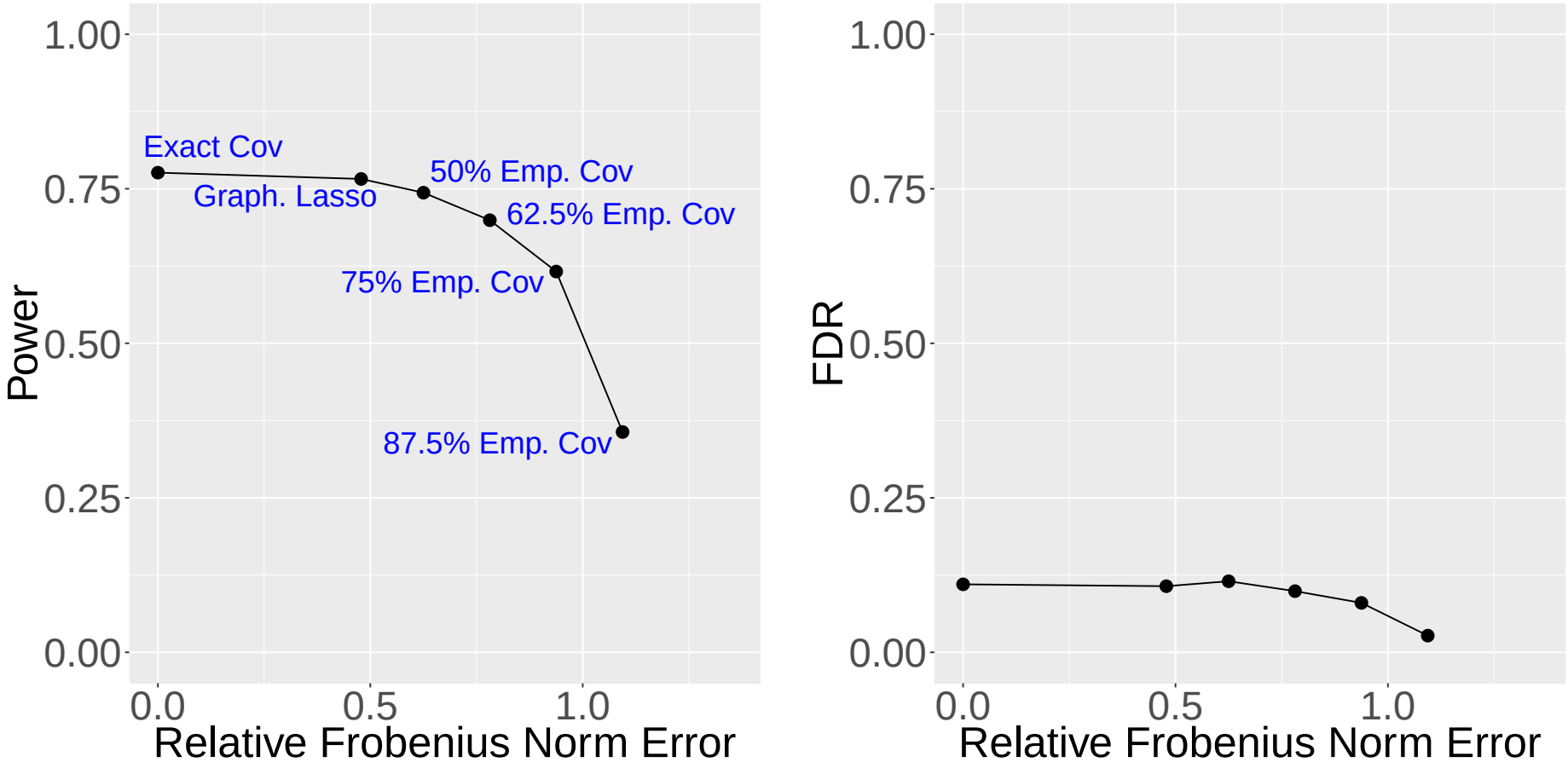


Figure: Covariates are AR(1) with autocorrelation coefficient 0.3. $n = 800$, $p = 1500$, and target FDR is 10%. $Y \mid X$ follows logistic model with 50 nonzero entries

Robustness

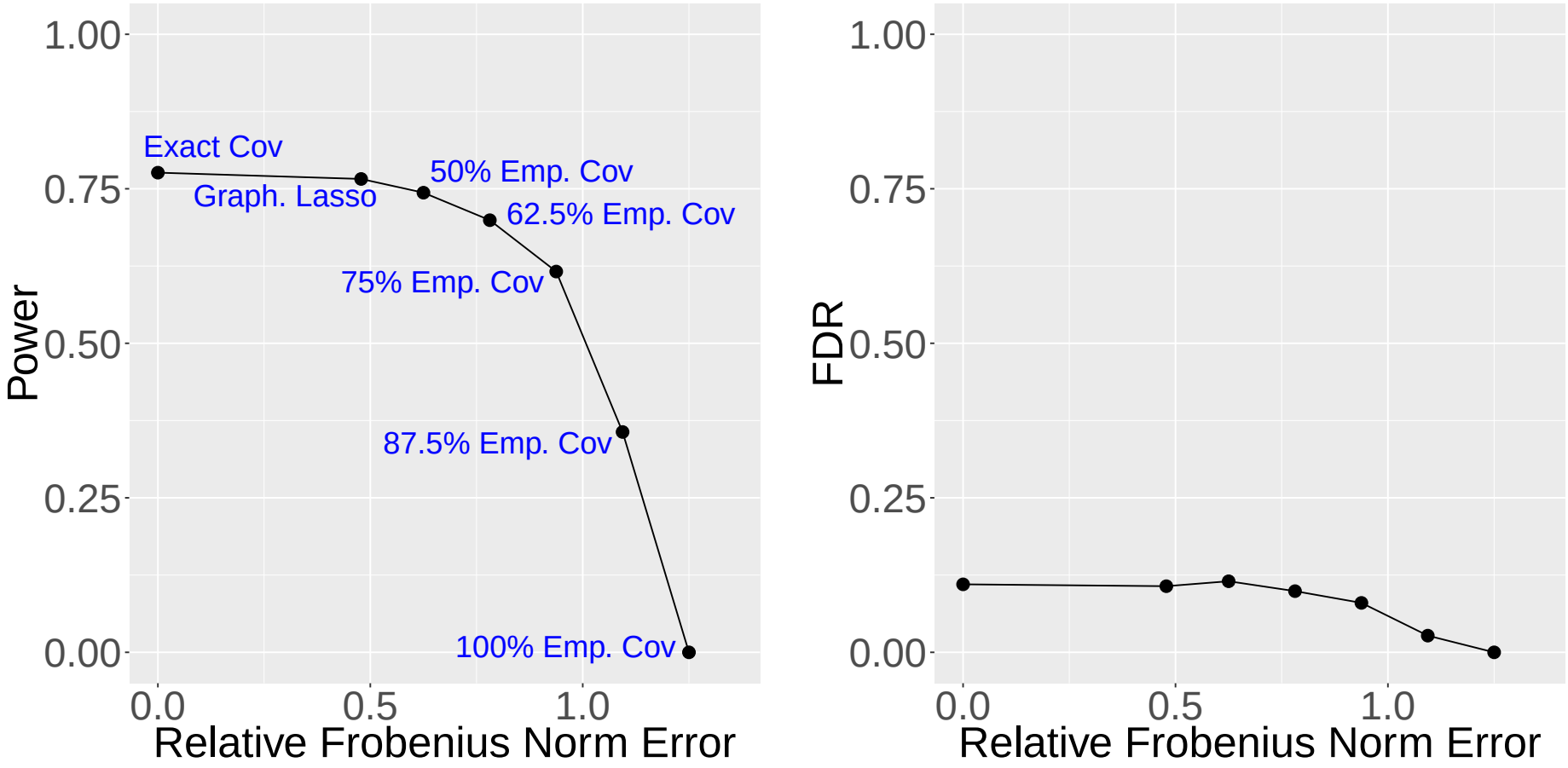


Figure: Covariates are AR(1) with autocorrelation coefficient 0.3. $n = 800$, $p = 1500$, and target FDR is 10%. $Y \mid X$ follows logistic model with 50 nonzero entries

Simulations with synthetic Markov chain

Markov chain covariates with 5 hidden states. Binomial response

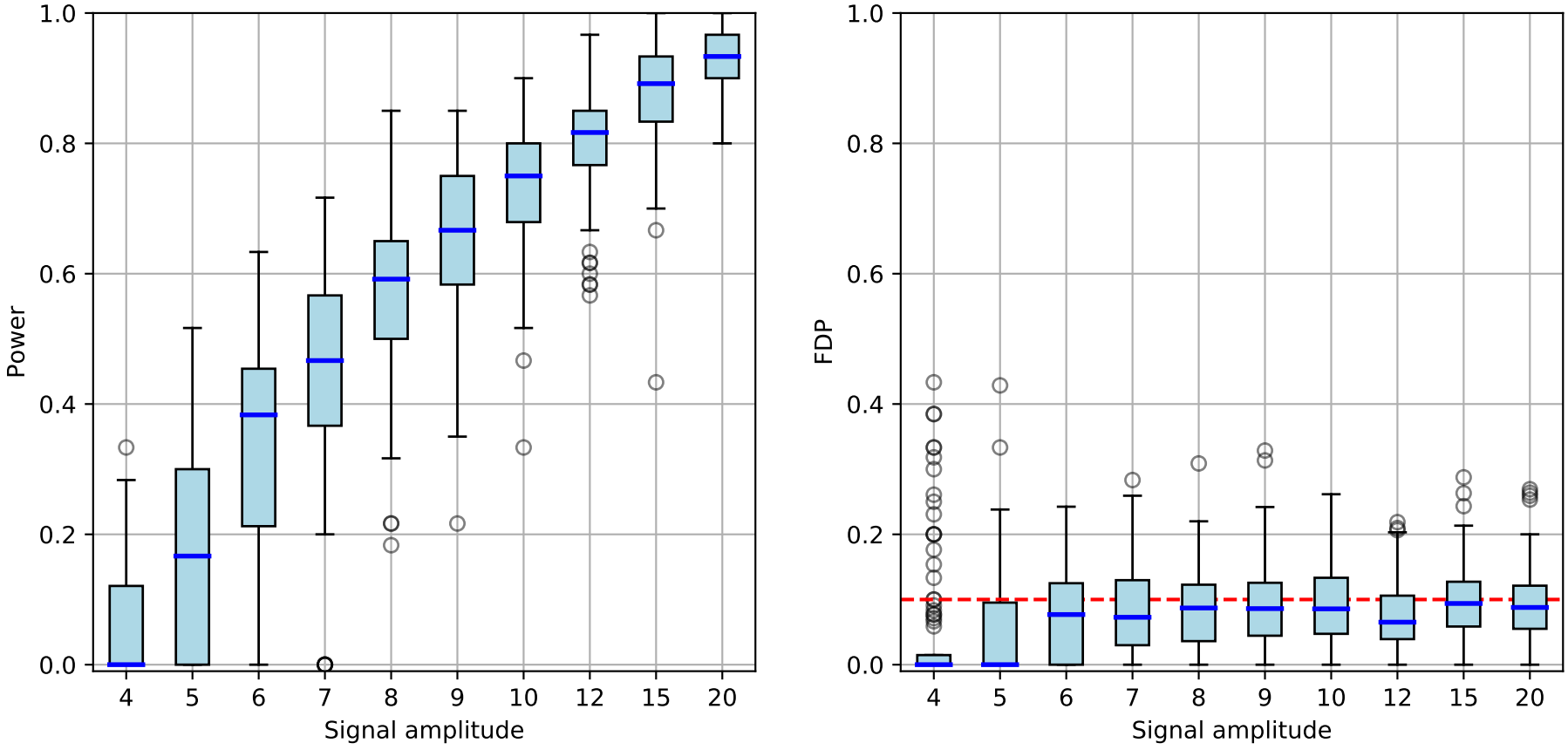


Figure: Power and FDP over 100 repetitions (true F_X)

$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$

$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Robustness

Markov chain covariates with 5 hidden states. Binomial response

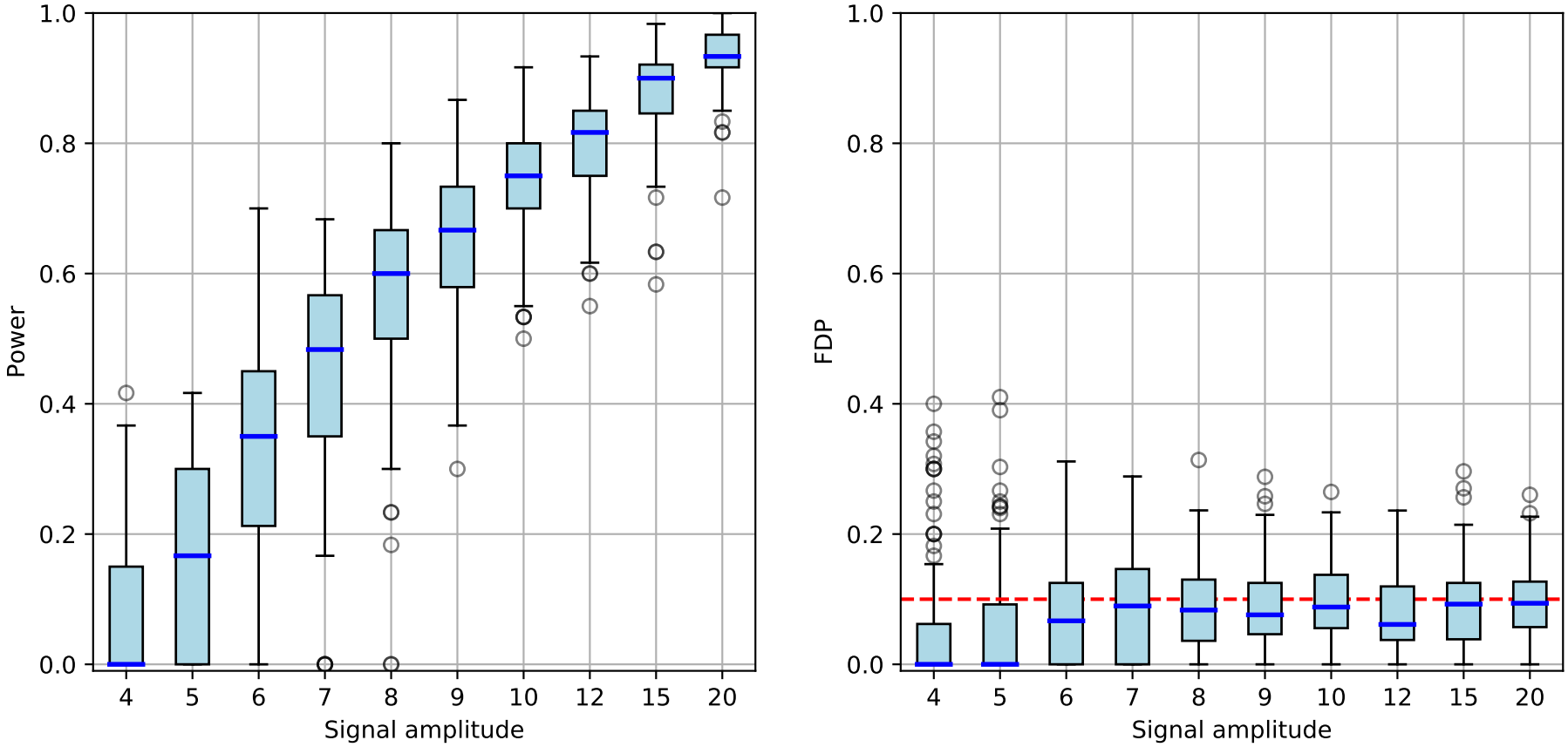


Figure: Power and FDP over 100 repetitions (estimated F_X)

$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$

$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Simulations with synthetic HMM

HMM covariates with latent "clockwise" Markov chain. Binomial response

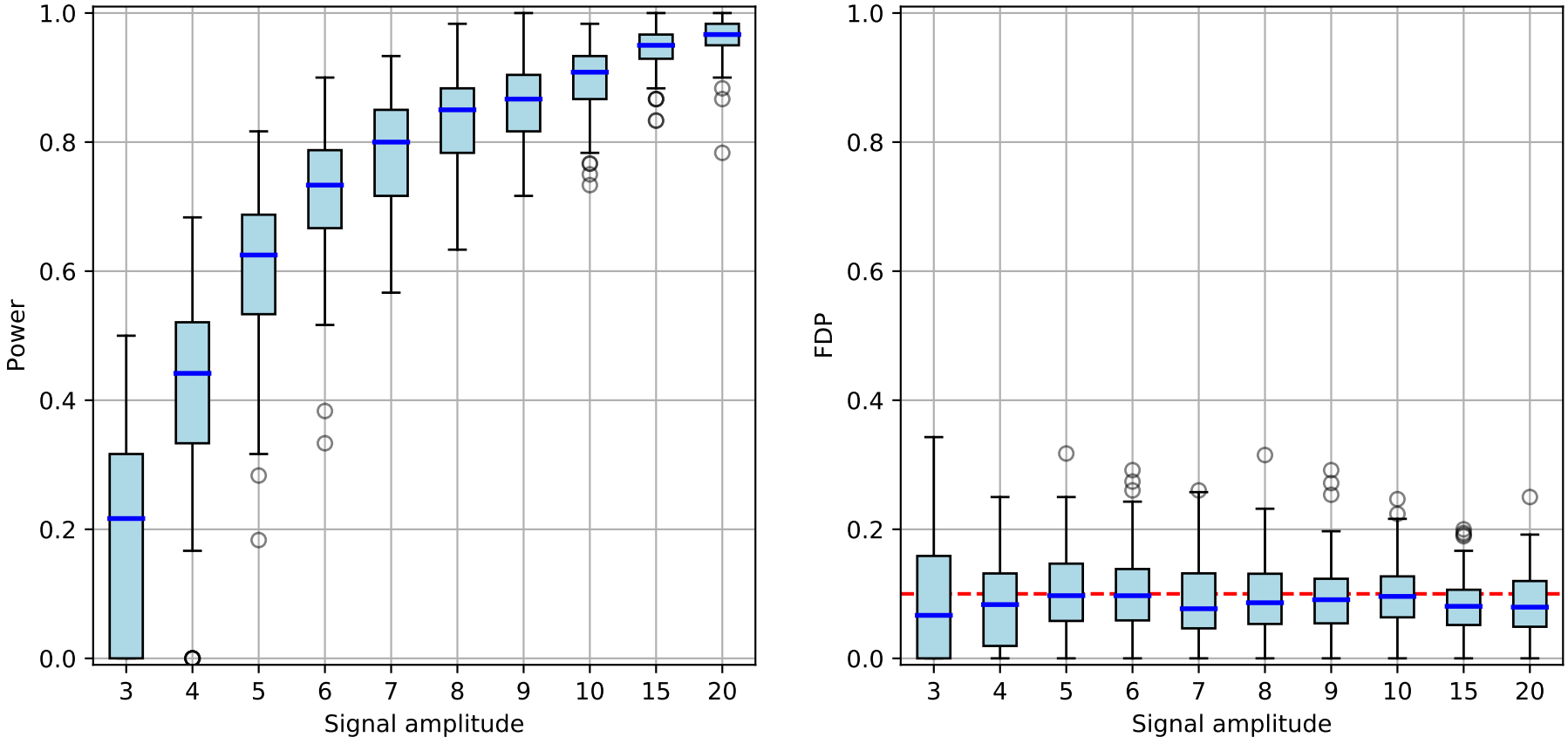


Figure: Power and FDP over 100 repetitions (true F_X)

$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$

$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Robustness

HMM covariates with latent "clockwise" Markov chain. Binomial response

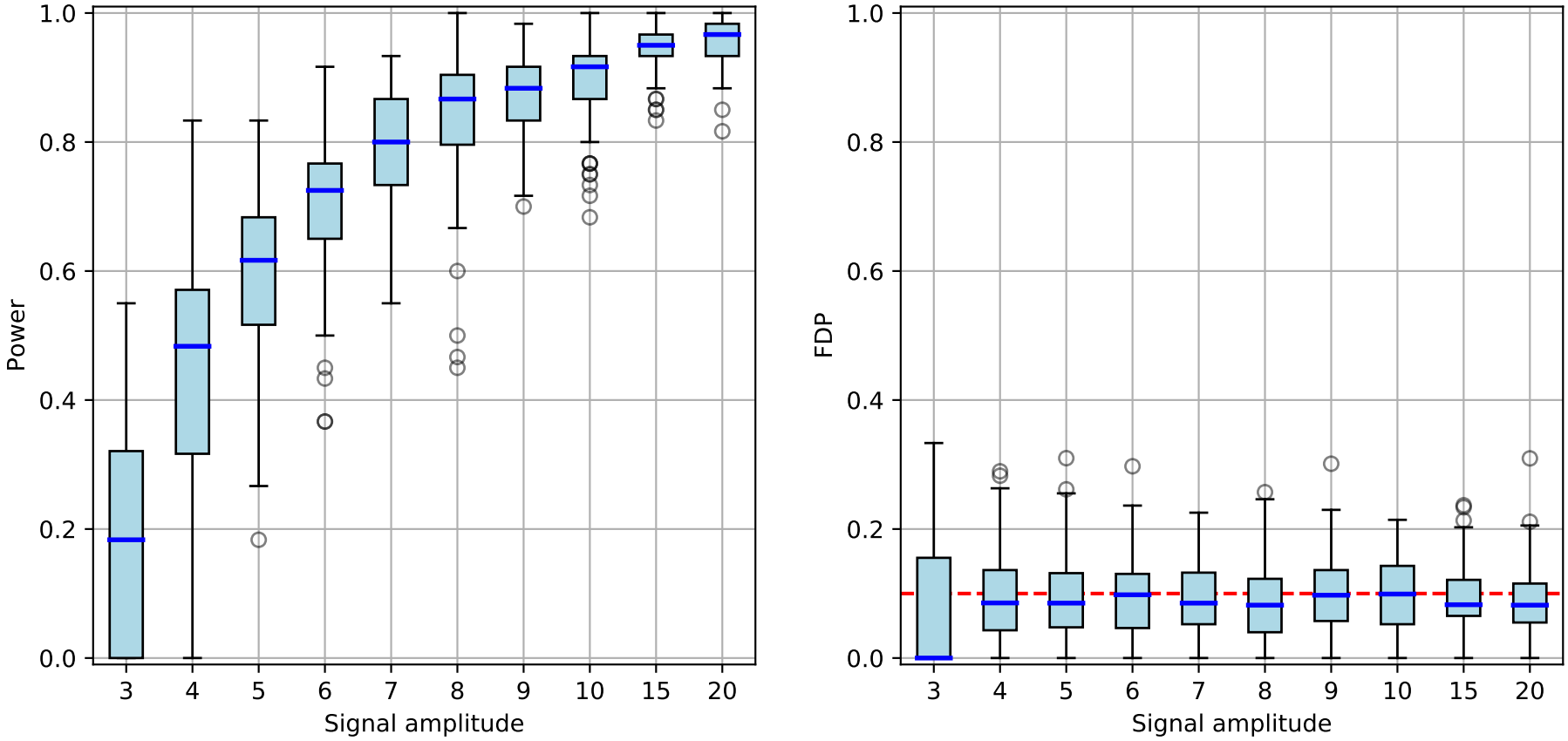


Figure: Power and FDP over 100 repetitions (estimated F_X)

$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$

$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Out-of-sample parameter estimation

Inhomogeneous Markov chain covariates with 5 hidden states. Binomial response

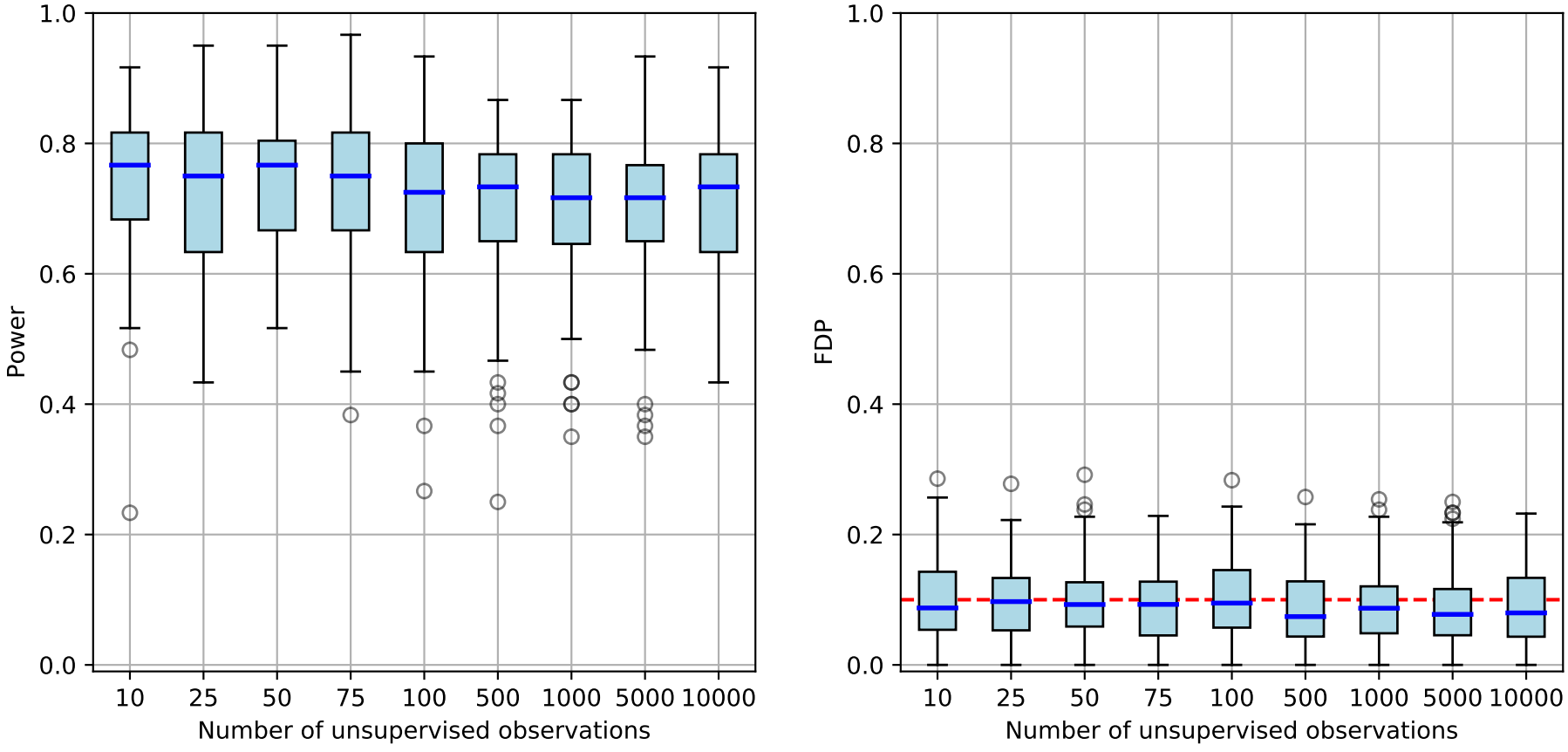


Figure: Power and FDP over 100 repetitions (estimated F_X from independent dataset)

$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$

$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Model-X knockoff variables (robust version)

i.i.d. samples from P_{XY}

- Distr. P_X of X only 'approx' known
- Distr. $P_{Y|X}$ of $Y \mid X$ completely unknown

Model-X knockoff variables (robust version)

i.i.d. samples from P_{XY}

- Distr. P_X of X only 'approx' known
- Distr. $P_{Y|X}$ of $Y \,|\, X$ completely unknown

Knockoffs wrt. to user input Q_X (Barber, C. and Samworth, '18)

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

Model-X knockoff variables (robust version)

i.i.d. samples from P_{XY}

- Distr. P_X of X only 'approx' known
- Distr. $P_{Y|X}$ of $Y \,|\, X$ completely unknown

Knockoffs wrt. to user input Q_X (Barber, C. and Samworth, '18)

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

(1) Pairwise exchangeability wrt Q_X: If $X \sim Q_X$

$$(X, \tilde{X})_{\mathsf{swap}(S)} \stackrel{d}{=} (X, \tilde{X})$$

e.g.

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\mathsf{swap}(\{2,3\})} \stackrel{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$

Model-X knockoff variables (robust version)

i.i.d. samples from P_{XY}

- Distr. P_X of X only 'approx' known
- Distr. $P_{Y|X}$ of $Y \,|\, X$ completely unknown

Knockoffs wrt. to user input Q_X (Barber, C. and Samworth, '18)

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

(1) Pairwise exchangeability wrt Q_X: If $X \sim Q_X$

$$(X, \tilde{X})_{\mathsf{swap}(S)} \stackrel{d}{=} (X, \tilde{X})$$

e.g.

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\mathsf{swap}(\{2,3\})} \stackrel{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$

(2) Ignore Y when constructing knockoffs: $\tilde{X} \perp\!\!\!\perp Y \,|\, X$

Model-X knockoff variables (robust version)

i.i.d. samples from P_{XY}

- Distr. P_X of X only 'approx' known
- Distr. $P_{Y|X}$ of $Y \,|\, X$ completely unknown

Knockoffs wrt. to user input Q_X (Barber, C. and Samworth, '18)

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

(1) Pairwise exchangeability wrt Q_X: If $X \sim Q_X$

$$(X, \tilde{X})_{\text{swap}(S)} \stackrel{d}{=} (X, \tilde{X})$$

e.g.

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\text{swap}(\{2,3\})} \stackrel{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$

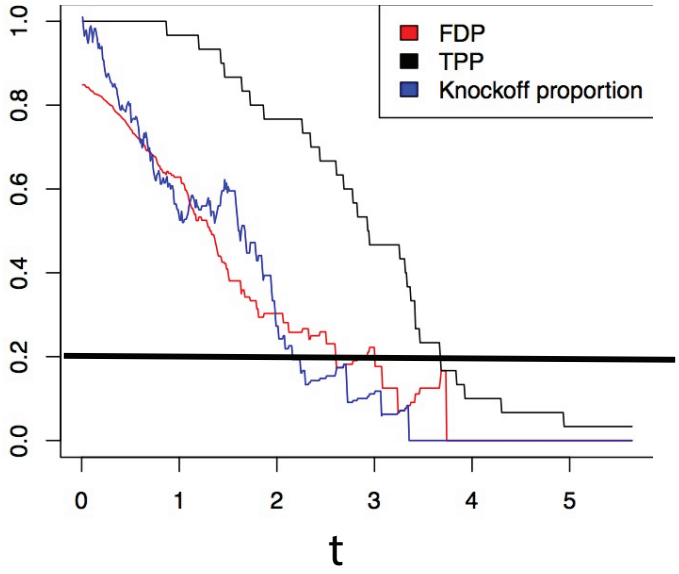
(2) Ignore Y when constructing knockoffs: $\tilde{X} \perp\!\!\!\perp Y \,|\, X$

Only require conditionals $Q(X_j|X_{-j})$ which do not have to be compatible

FDR control

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$

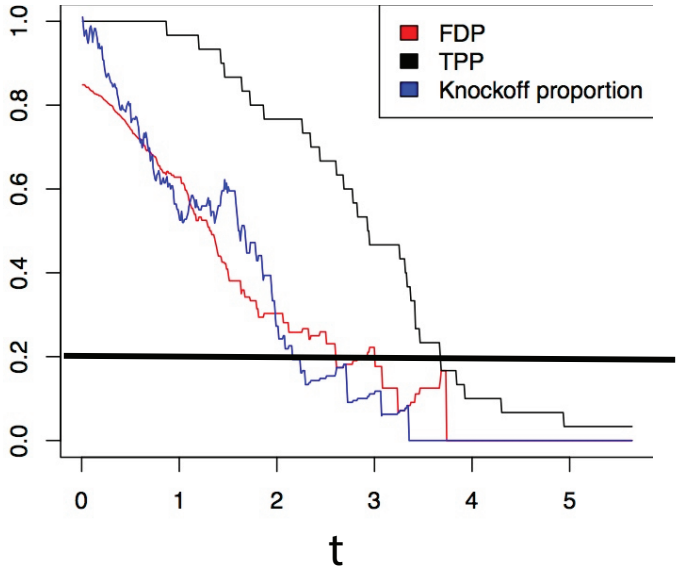
$$\tau = \min\left\{t : \underbrace{\frac{1+|\{j : W_j \leq -t\}|}{1 \vee |\{j : W_j \geq t\}|}}_{\widehat{\mathrm{FDP}}(t)} \leq q\right\}$$



FDR control

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$

$$\tau = \min\left\{t : \underbrace{\frac{1+|\{j : W_j \leq -t\}|}{1 \vee |\{j : W_j \geq t\}|}}_{\widehat{\mathrm{FDP}}(t)} \leq q\right\}$$

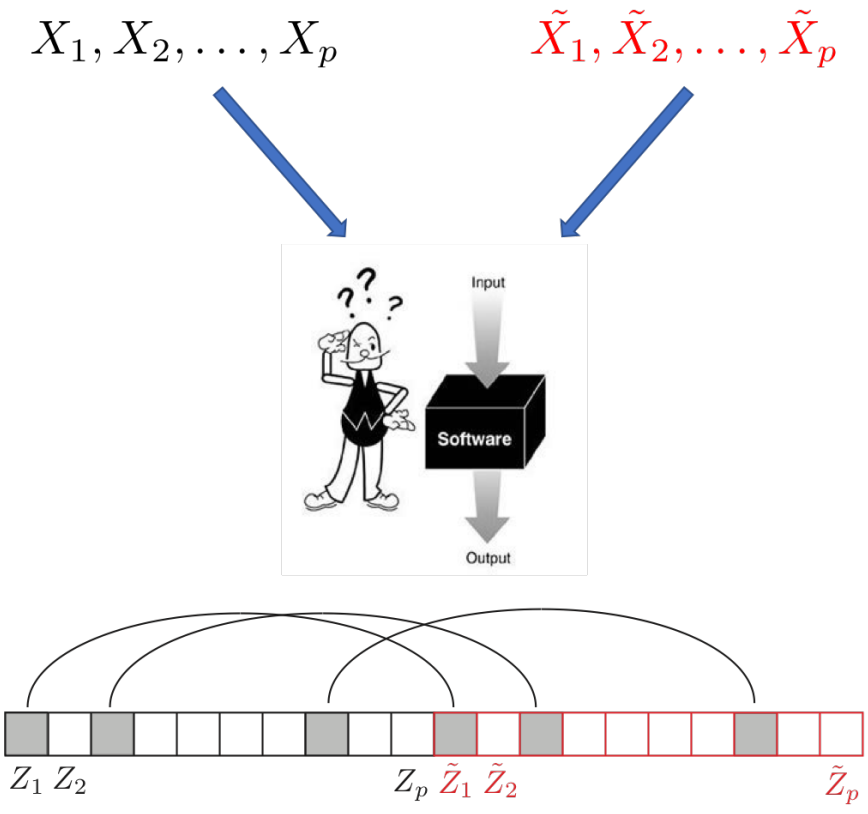


Theorem (Barber and C. ('15))

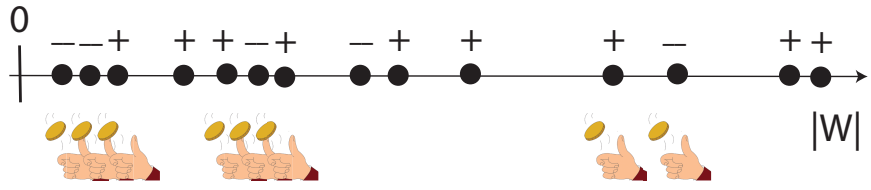
If user-input Q_X is correct ($Q_X = P_X$), then for knockoff+

$$\mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections}}\right] \leq q$$

Robustness of knockoffs?



Does exchangeability hold approx. when $Q_X \neq P_X$?



If $P_X = Q_X$, coins are

- unbiased
- independent

Problem: if $P_X \neq Q_X$, coins may be

- (slighltly) biased
- (slightly) dependent

KL divergence condition

- The KL condition

$$\widehat{\mathsf{KL}}_j := \sum_i \log\left(\frac{P_j(X_{ij}|X_{i,-j})\,Q_j(\widetilde{X}_{ij}|X_{i,-j})}{Q_j(X_{ij}|X_{i,-j})\,P_j(\widetilde{X}_{ij}|X_{i,-j})}\right) \leq \epsilon$$

- $\mathbb{E}[\widehat{\mathsf{KL}}_j] =$ KL divergence between distributions of

$$(X_j, \widetilde{X}_j, X_{-j}, \widetilde{X}_{-j}) \quad \& \quad (\widetilde{X}_j, X_j, X_{-j}, \widetilde{X}_{-j})$$

From KL condition to FDR control

Theorem (Barber, C. and Samworth (2018))

For any $\epsilon \geq 0$

$$\mathbb{E}\left[\frac{\#\text{ false positives } j \text{ with } \widehat{\mathsf{KL}}_j \leq \epsilon}{\#\text{ selections}}\right] \leq q\, \exp(\epsilon)$$

From KL condition to FDR control

Theorem (Barber, C. and Samworth (2018))

For any $\epsilon \geq 0$

$$\mathbb{E}\left[\frac{\#\text{ false positives } j \text{ with } \widehat{\mathsf{KL}}_j \leq \epsilon}{\#\text{ selections}}\right] \leq q\,\exp(\epsilon)$$

Corollary

$$\mathsf{FDR} \leq \min_{\epsilon \geq 0}\left\{q\,\exp(\epsilon) + \mathbb{P}\left(\max_{\text{null } j} \widehat{\mathsf{KL}}_j > \epsilon\right)\right\}$$

Information theoretically optimal

New directions

ML inspired knockoffs

Joint with S. Bates, Y. Romano, M. Sesia and J. Zhou

- Knockoffs for graphical models
- Knockoffs via restricted Boltzmann machines
- Knockoffs via variational auto-encoders?
- Knockoffs via generative adversarial networks?

Improving power?

Joint with Z. Ren and M. Sesia

Derandomization

Combine information from mutiple knockoffs: who's consistently showing up?

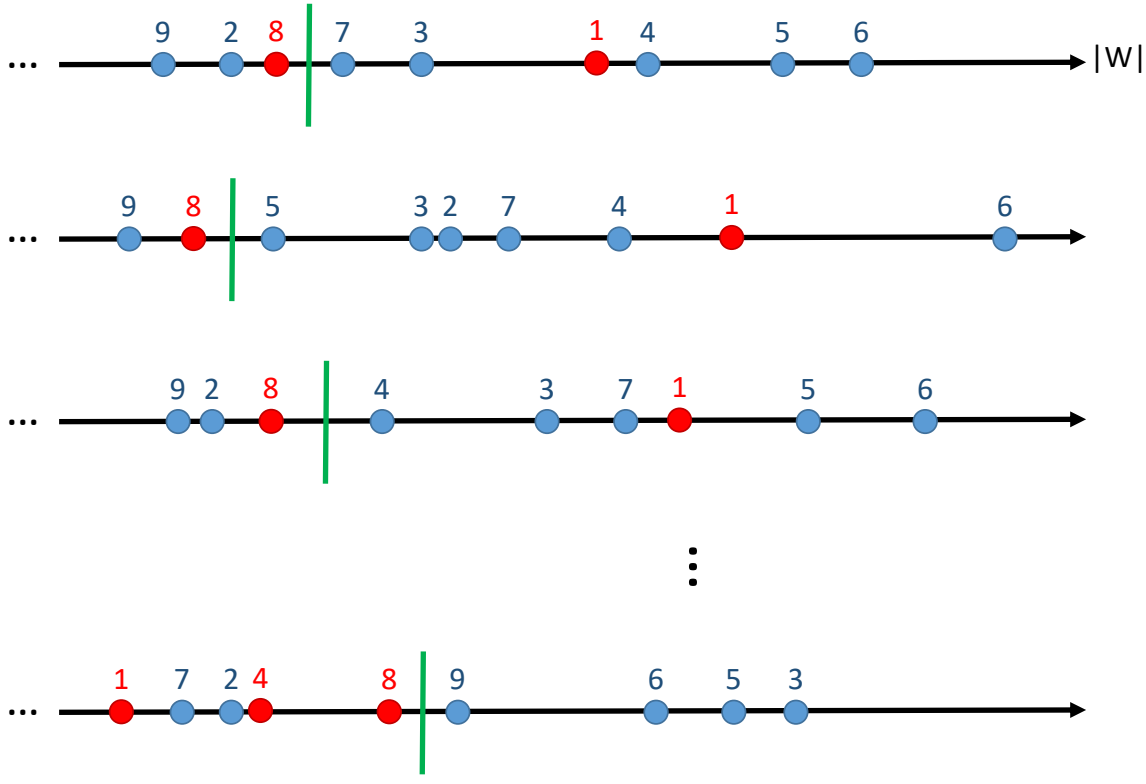
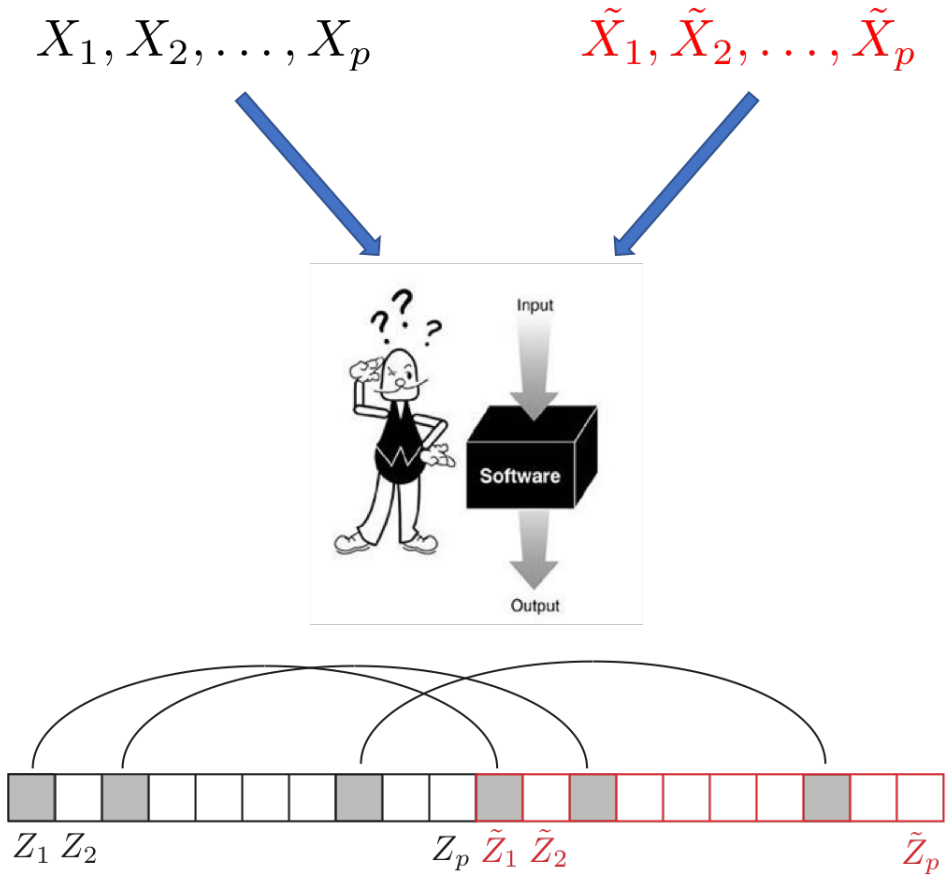


Figure: Cartoon representation of W's from different sample realizations of knockoffs

Knockoffs for Fixed Features

Joint with Barber



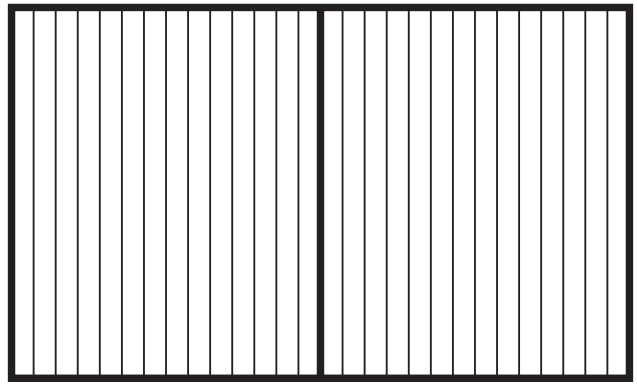
Linear model

$$\underset{n\times 1}{y} = \overbrace{\underset{n\times p}{X}\underset{p\times 1}{\beta}}^{\sum_j \beta_j X_j} + \underset{n\times 1}{z} \qquad y \sim \mathcal{N}(X\beta, \sigma^2 I)$$

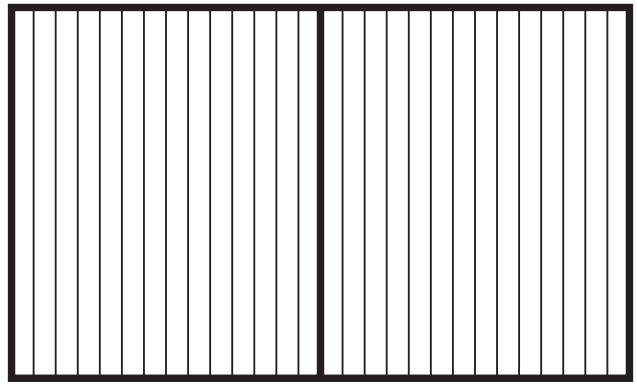
- Fixed design X
- Noise level σ unknown
- Multiple testing: $H_j : \beta_j = 0$ (is jth variable in the model?)
- Identifiability $\implies$ $p \leq n$

Inference (FDR control) will hold conditionally on X

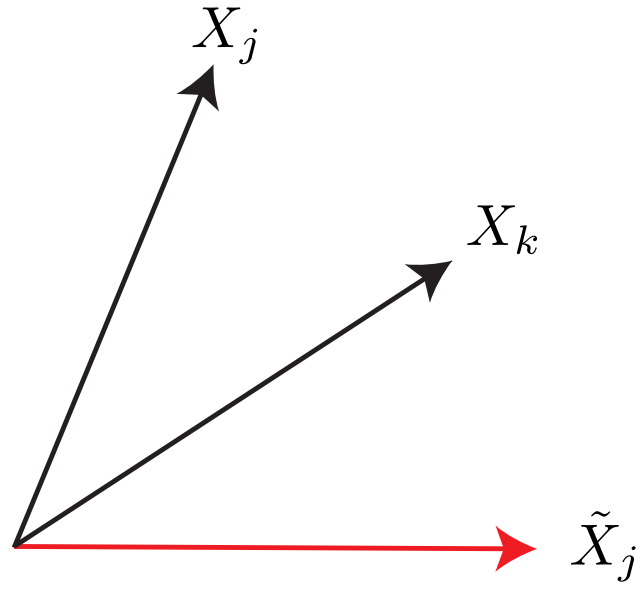
Knockoff features (fixed X)



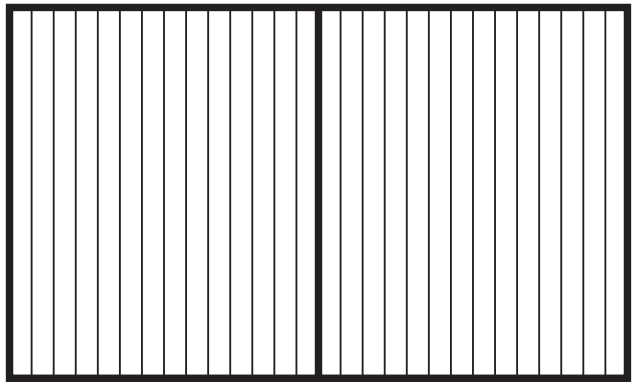
Knockoff features (fixed X)



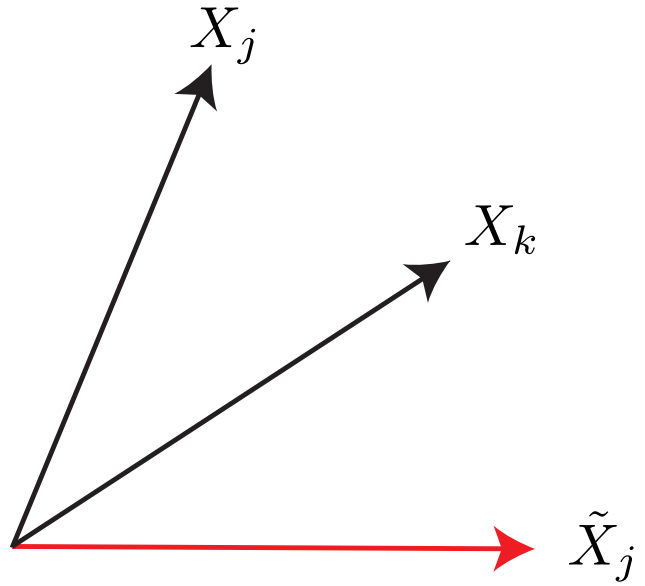
$$\tilde{X}_j' \tilde{X}_k = X_j' X_k \qquad \text{for all } j, k$$
$$\tilde{X}_j' X_k = X_j' X_k \qquad \text{for all } j \neq k$$



Knockoff features (fixed X)



$$\tilde{X}_j'\tilde{X}_k = X_j'X_k \qquad \text{for all } j,k$$
$$\tilde{X}_j'X_k = X_j'X_k \qquad \text{for all } j \neq k$$



- No need for new data or experiment
- No knowledge of response y

Knockoff construction ($n \geq 2p$)

Problem: given $X \in \mathbb{R}^{n \times p}$, find $\tilde{X} \in \mathbb{R}^{n \times p}$ s.t.

$$[X \;\; \tilde{X}]' [X \;\; \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} := G$$

Knockoff construction ($n \geq 2p$)

Problem: given $X \in \mathbb{R}^{n \times p}$, find $\tilde{X} \in \mathbb{R}^{n \times p}$ s.t.

$$[X \;\; \tilde{X}]' [X \;\; \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} := G \succeq 0$$

Knockoff construction ($n \geq 2p$)

Problem: given $X \in \mathbb{R}^{n \times p}$, find $\tilde{X} \in \mathbb{R}^{n \times p}$ s.t.

$$[X \;\; \tilde{X}]' [X \;\; \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} := G \succeq 0$$

$$G \succeq 0 \quad \Longleftrightarrow \quad \begin{array}{l} \mathrm{diag}\{s\} \succeq 0 \\ 2\Sigma - \mathrm{diag}\{s\} \succeq 0 \end{array}$$

Knockoff construction ($n \geq 2p$)

Problem: given $X \in \mathbb{R}^{n \times p}$, find $\tilde{X} \in \mathbb{R}^{n \times p}$ s.t.

$$[X \;\; \tilde{X}]' [X \;\; \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} := G \succeq 0$$

$$G \succeq 0 \quad \Longleftrightarrow \quad \begin{array}{l} \mathrm{diag}\{s\} \succeq 0 \\ 2\Sigma - \mathrm{diag}\{s\} \succeq 0 \end{array}$$

Solution

$$\tilde{X} = X(I - \Sigma^{-1}\,\mathrm{diag}\{s\}) + \tilde{U}C$$

- $\tilde{U} \in \mathbb{R}^{n \times p}$ with col. space orthogonal to that of X
- $C'C$ Cholevsky factorization of $2\,\mathrm{diag}\{s\} - \mathrm{diag}\{s\}\Sigma^{-1}\,\mathrm{diag}\{s\} \succeq 0$

Knockoff construction ($n \geq 2p$)

$$\tilde{X}_j' X_j = 1 - s_j \qquad \text{(Standardized columns)}$$

Equi-correlated knockoffs

$$s_j = 2\lambda_{\min}(\Sigma) \wedge 1$$

Under equivariance, minimizes the value of $|\langle X_j, \tilde{X}_j \rangle|$

Knockoff construction ($n \geq 2p$)

$$\tilde{X}_j' X_j = 1 - s_j \qquad \text{(Standardized columns)}$$

Equi-correlated knockoffs

$$s_j = 2\lambda_{\min}(\Sigma) \wedge 1$$

Under equivariance, minimizes the value of $|\langle X_j, \tilde{X}_j \rangle|$

SDP knockoffs

$$\begin{array}{ll} \text{minimize} & \sum_j |1 - s_j| \\ \text{subject to} & s_j \geq 0 \\ & \mathrm{diag}\{s\} \preceq 2\Sigma \end{array}$$

Highly structured semidefinite program (SDP)

Knockoff construction ($n \geq 2p$)

$$\tilde{X}_j' X_j = 1 - s_j \qquad \text{(Standardized columns)}$$

Equi-correlated knockoffs

$$s_j = 2\lambda_{\min}(\Sigma) \wedge 1$$

Under equivariance, minimizes the value of $|\langle X_j, \tilde{X}_j \rangle|$

SDP knockoffs

$$\begin{array}{ll} \text{minimize} & \sum_j |1 - s_j| \\ \text{subject to} & s_j \geq 0 \\ & \mathrm{diag}\{s\} \preceq 2\Sigma \end{array}$$

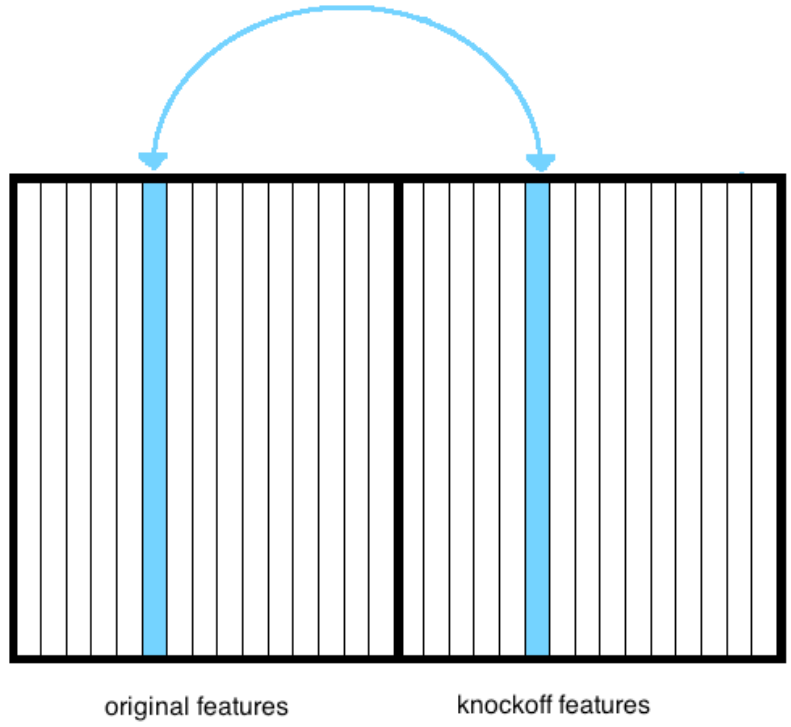
Highly structured semidefinite program (SDP)

Other possibilities ...

Why?

For null feature X_j

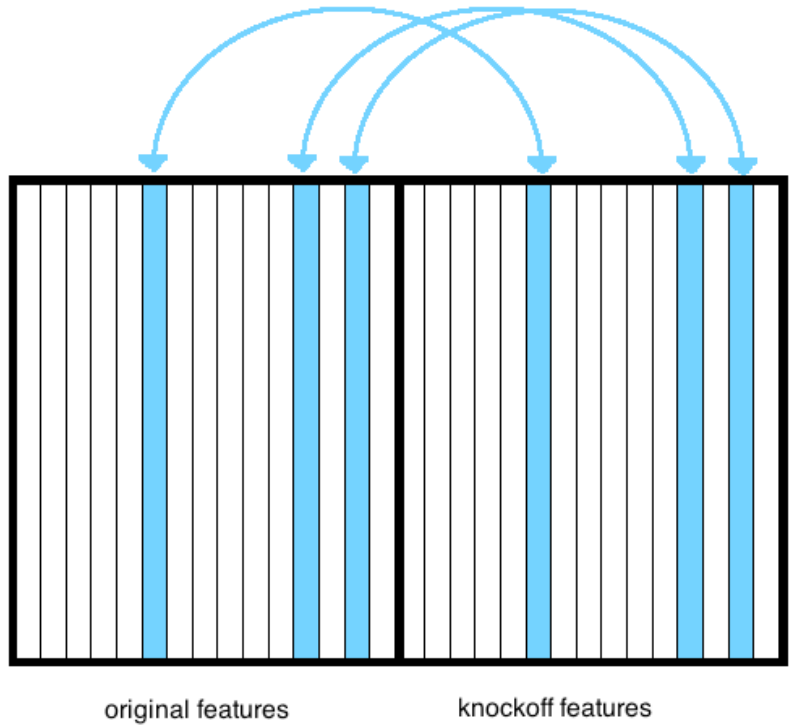
$$X_j'y = X_j'X\beta + X_j'z \stackrel{d}{=} \tilde{X}_j'X\beta + \tilde{X}_j'z = \tilde{X}_j'y$$



Why?

For null feature X_j

$$X_j'y = X_j'X\beta + X_j'z \overset{d}{=} \tilde{X}_j'X\beta + \tilde{X}_j'z = \tilde{X}_j'y$$

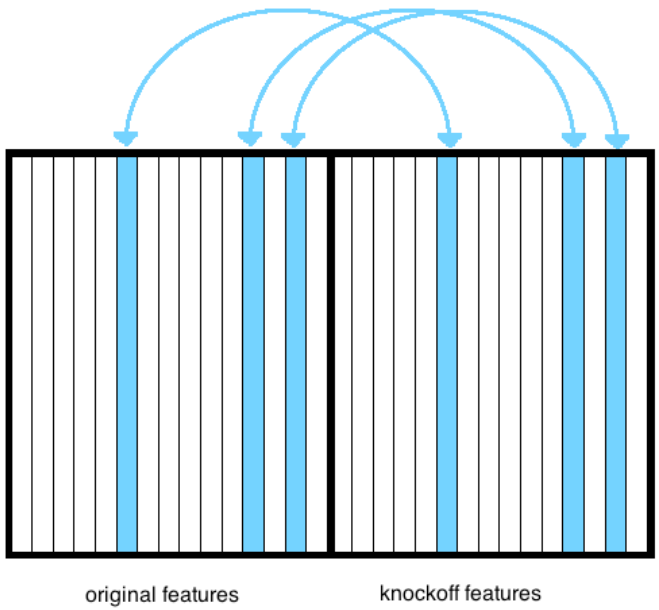


Why?

For any subset of nulls $\mathcal{T}$

$$[X\ \tilde{X}]'_{\text{swap}(\mathcal{T})}\ y \overset{d}{=} [X\ \tilde{X}]'\ y$$

$$[X\ \tilde{X}]'_{\text{swap}(\mathcal{T})} =$$



Exchangeability of feature importance statistics

- *Sufficiency:*

$$(Z, \tilde{Z}) = z\left(\left[X \;\; \tilde{X}\right]' \left[X \;\; \tilde{X}\right], \left[X \;\; \tilde{X}\right]' y\right)$$

- *Knockoff-agnostic:* swapping originals and knockoffs $\Longrightarrow$ swaps Z's

$$z(\left[X \;\; \tilde{X}\right]_{\mathsf{swap}(\mathcal{T})}, y) = (Z, \tilde{Z})_{\mathsf{swap}(\mathcal{T})}$$

Exchangeability of feature importance statistics

- *Sufficiency:*

$$(Z, \tilde{Z}) = z\left(\left[X \;\; \tilde{X}\right]' \left[X \;\; \tilde{X}\right], \left[X \;\; \tilde{X}\right]' y\right)$$

- *Knockoff-agnostic:* swapping originals and knockoffs $\Longrightarrow$ swaps Z's

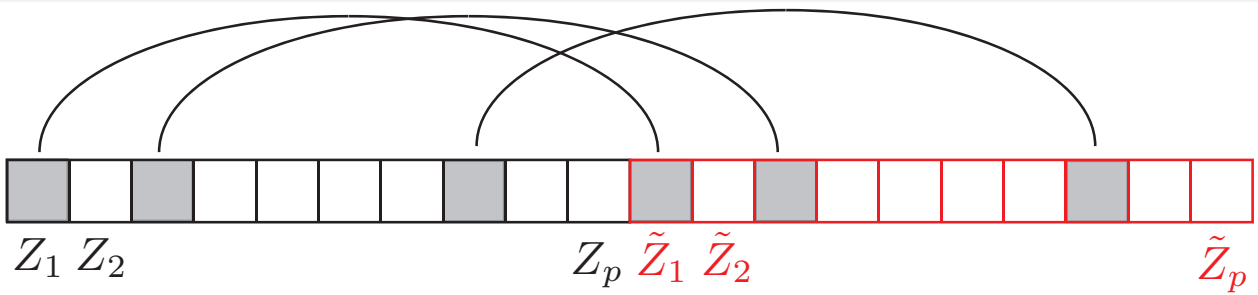
$$z(\left[X \;\; \tilde{X}\right]_{\text{swap}(\mathcal{T})}, y) = (Z, \tilde{Z})_{\text{swap}(\mathcal{T})}$$

Theorem (Barber and C. (15))

For any subset $\mathcal{T}$ of nulls

$$(Z, Z)_{\text{swap}(\mathcal{T})} \stackrel{d}{=} (Z, \tilde{Z})$$

$\Longrightarrow$ *FDR control (conditional on X)*



Telling the effect direction

[...] in classical statistics, the significance of comparisons (e. g., $\theta_1 - \theta_2$) is calibrated using Type I error rate, relying on the assumption that the true difference is zero, which makes no sense in many applications. [...] a more relevant framework in which a true comparison can be positive or negative, and, based on the data, you can state "$\theta_1 > \theta_2$ with confidence", "$\theta_2 > \theta_1$ with confidence", or "no claim with confidence".

A. Gelman & F. Tuerlinckx

Directional FDR

Are any effects exactly zero?

$$\underbrace{\text{FDR}_{\text{dir}}}_{\text{Directional false discovery rate}} = \mathbb{E}\underbrace{\left[\frac{\#\text{ selections with wrong effect direction}}{\#\text{ selections}}\right]}_{\text{Directional false discovery proportion}}$$

- Directional FDR (Benjamini & Yekutieli, '05)
- Sign errors (Type-S) (Gelman & Tuerlinckx, '00)

Important for misspecified models — exact sparsity unlikely

Directional FDR control

$$(X_j - \tilde{X}_j)'y \overset{\text{ind}}{\sim} \mathcal{N}(s_j \cdot \beta_j,\, 2\sigma^2 \cdot s_j) \qquad s_j \geq 0$$

$$\text{Sign estimate} \quad \rightsquigarrow \quad \text{sgn}((X_j - \tilde{X}_j)'y)$$

Directional FDR control

$$(X_j - \tilde{X}_j)'y \overset{\text{ind}}{\sim} \mathcal{N}(s_j \cdot \beta_j,\, 2\sigma^2 \cdot s_j) \qquad s_j \geq 0$$

$$\text{Sign estimate} \quad \rightsquigarrow \quad \text{sgn}((X_j - \tilde{X}_j)'y)$$

Theorem (Barber and C., '16)

Exact same knockoff selection + sign estimate

$$FDR \leq FDR_{dir} \leq q$$

Directional FDR control

$$(X_j - \tilde{X}_j)'y \overset{\text{ind}}{\sim} \mathcal{N}(s_j \cdot \beta_j,\, 2\sigma^2 \cdot s_j) \qquad s_j \geq 0$$

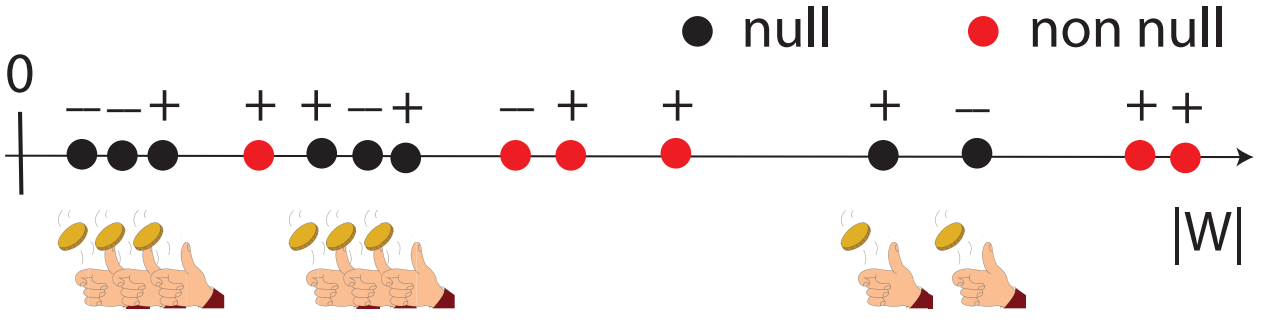
$$\text{Sign estimate} \quad \rightsquigarrow \quad \text{sgn}((X_j - \tilde{X}_j)'y)$$

Theorem (Barber and C., '16)

Exact same knockoff selection + sign estimate

$$FDR \leq FDR_{dir} \leq q$$

Null coin flips are unbiased



Directional FDR control

$$(X_j - \tilde{X}_j)'y \overset{\text{ind}}{\sim} \mathcal{N}(s_j \cdot \beta_j,\, 2\sigma^2 \cdot s_j) \qquad s_j \geq 0$$

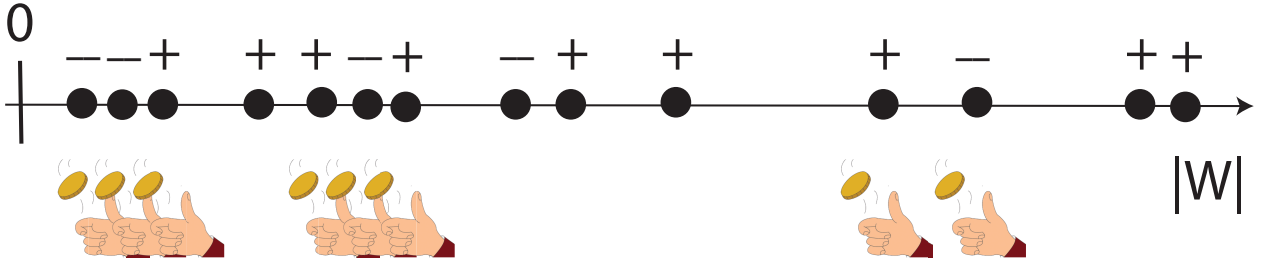
$$\text{Sign estimate} \quad \rightsquigarrow \quad \text{sgn}((X_j - \tilde{X}_j)'y)$$

Theorem (Barber and C. (16))

Exact same knockoff selection + sign estimate

$$FDR \leq FDR_{dir} \leq q$$

Great subtlety: coin flips are now biased



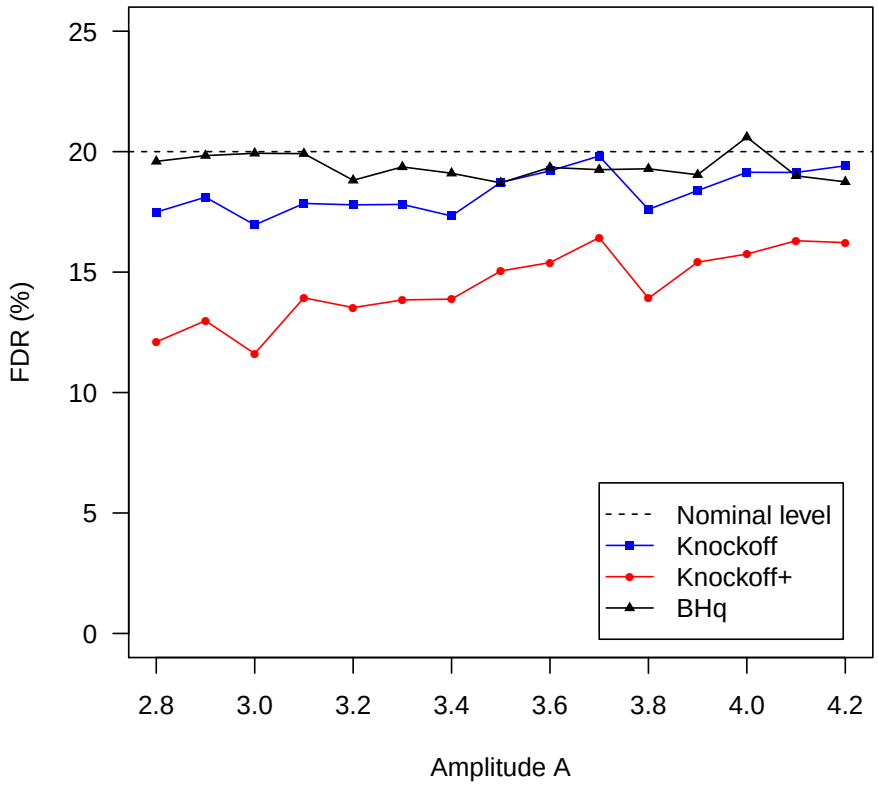
Empirical results

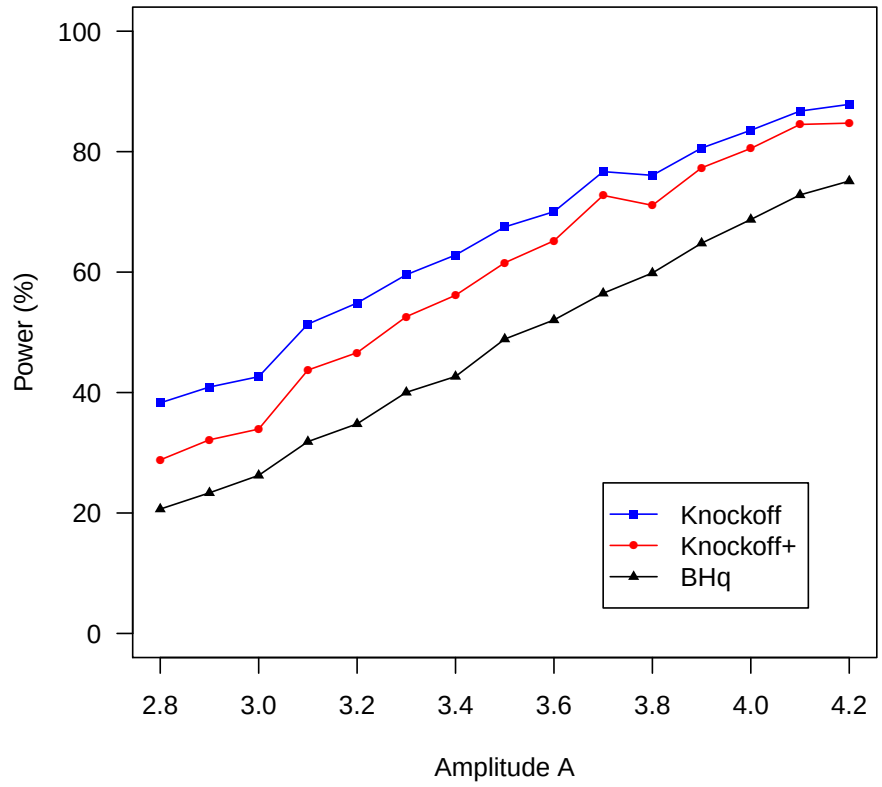
- Features $\mathcal{N}(0, I_n)$, $n = 3000$, $p = 1000$
- $k = 30$ variables with regression coefficients of magnitude 3.5

Method	FDR (%) (nominal level $q = 20\%$)	Power (%)	Theor. FDR control?
Knockoff+ (equivariant)	**14.40**	**60.99**	**Yes**
Knockoff (equivariant)	17.82	66.73	No
Knockoff+ (SDP)	**15.05**	**61.54**	**Yes**
Knockoff (SDP)	18.72	67.50	No
BHq	18.70	48.88	No
BHq + log-factor correction	**2.20**	**19.09**	**Yes**
BHq with whitened noise	**18.79**	**2.33**	**Yes**

Effect of signal amplitude

Same setup with $k = 30$ $(q = 0.2)$

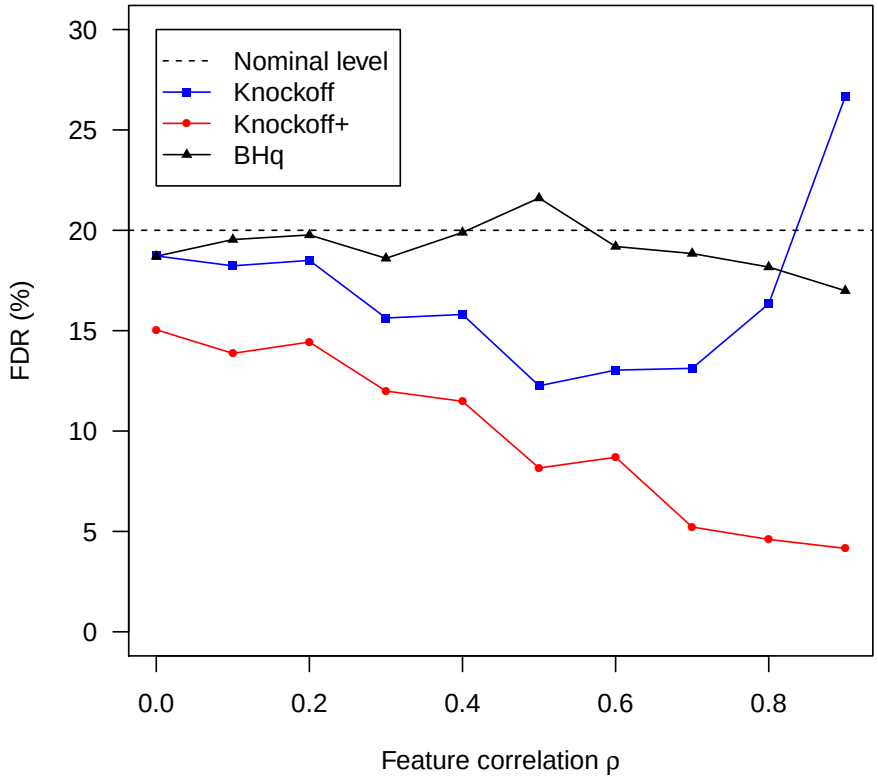


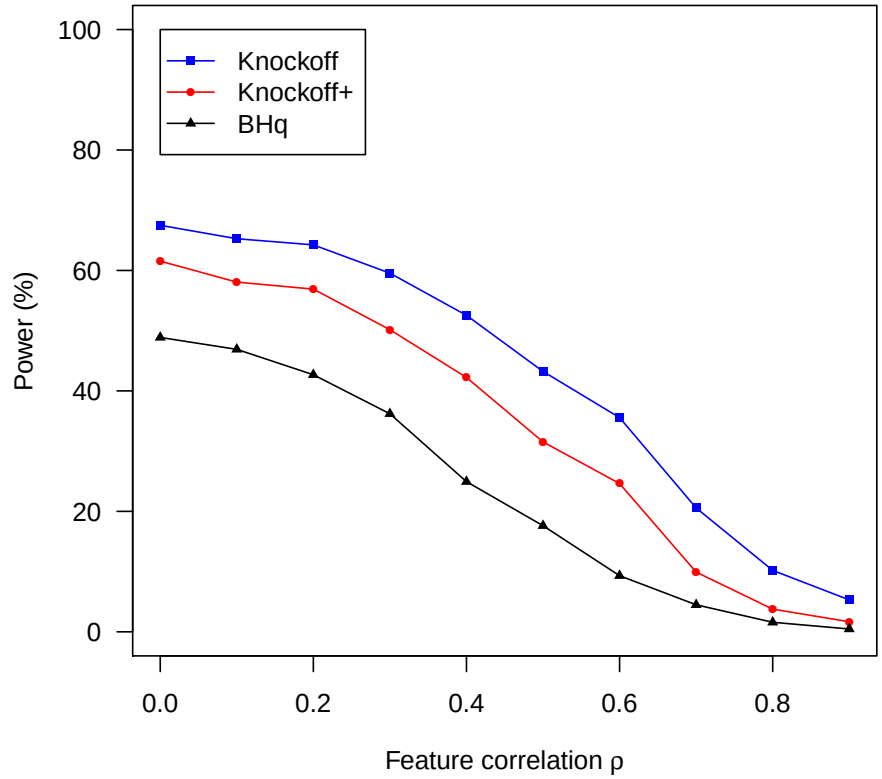


Effect of feature correlation

$$\text{Features} \sim \mathcal{N}(0, \Theta) \qquad \Theta_{jk} = \rho^{|j-k|}$$

$n = 3000$, $p = 1000$, and $k = 30$ and amplitude $= 3.5$





Fixed Design Knockoff Data Analysis

HIV drug resistance

Drug type	# drugs	Sample size	# protease or RT positions genotyped	# mutations appearing ≥ 3 times in sample
PI	6	848	99	209
NRTI	6	639	240	294
NNRTI	3	747	240	319

- response y: log-fold-increase of lab-tested drug resistance
- covariate X_j: presence or absence of mutation $\#j$

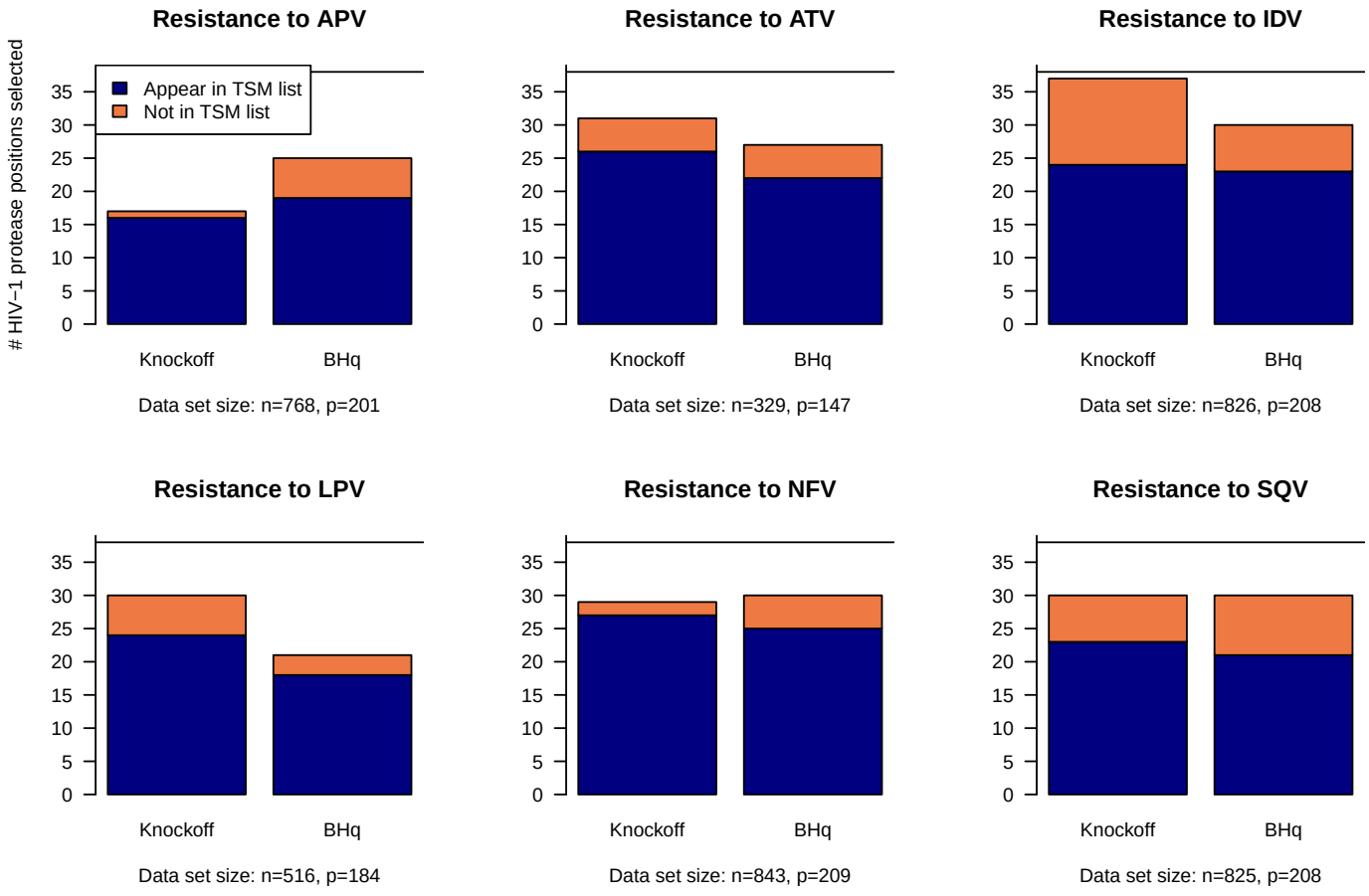
Data from R. Shafer (Stanford) available at:

`http://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006/`

HIV data

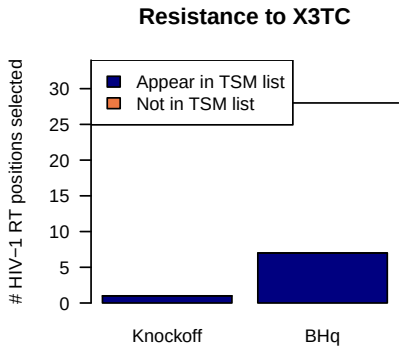
TSM list: mutations associated with the PI class of drugs in general, and is not specialized to the individual drugs in the class

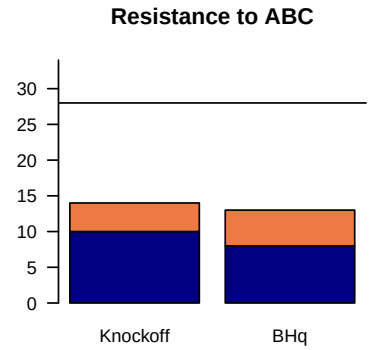
Results for PI type drugs

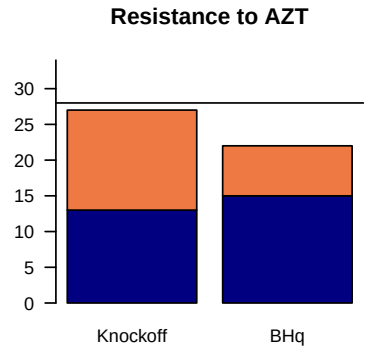


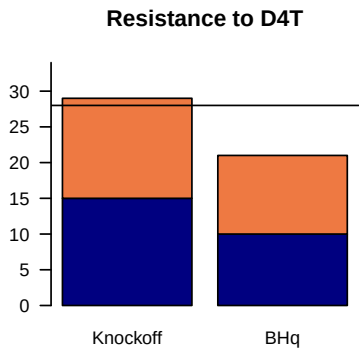
HIV data

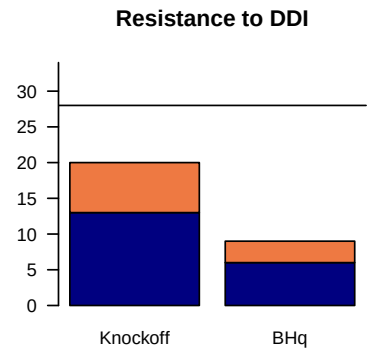
Results for
NRTI type drugs

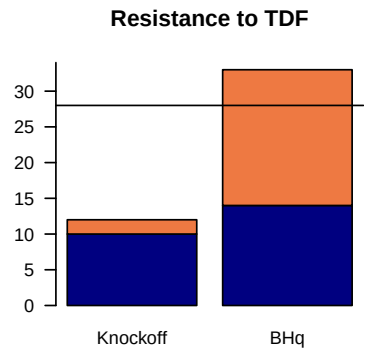




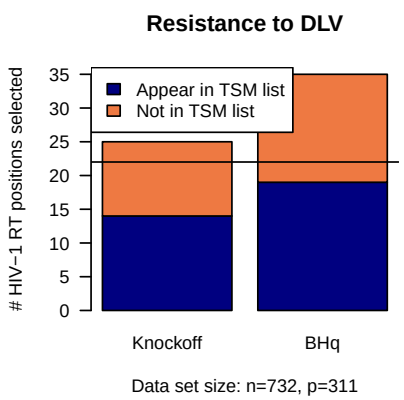


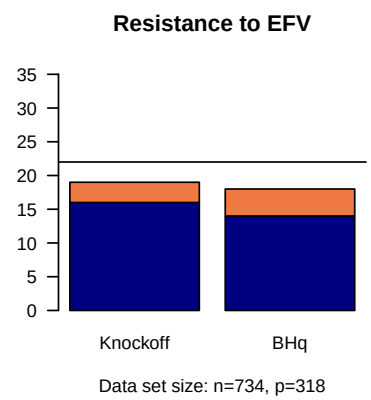


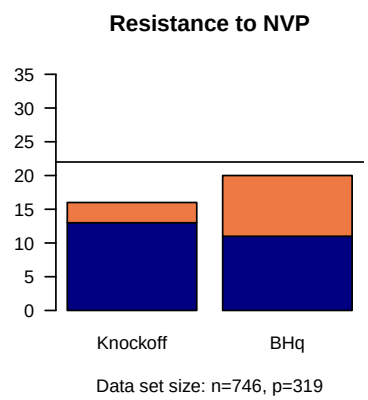




Results for
NNRTI type drugs

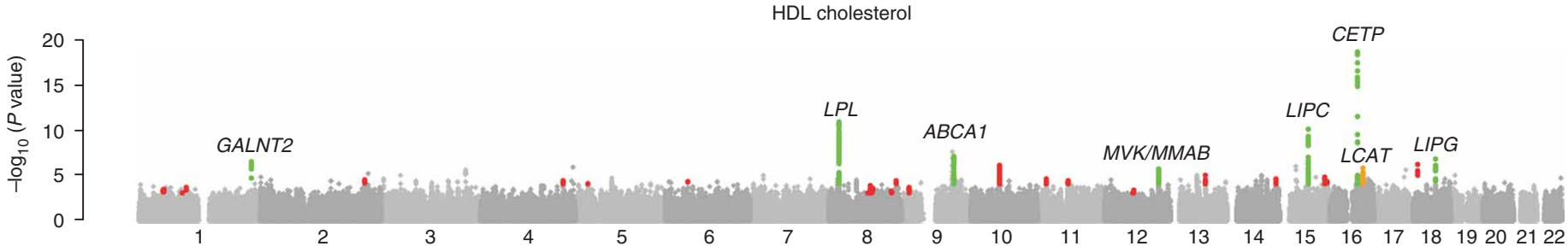






High-dimensional setting

- $n \approx 5,000$ subjects
- $p \approx 330,000$ SNPs/vars to test

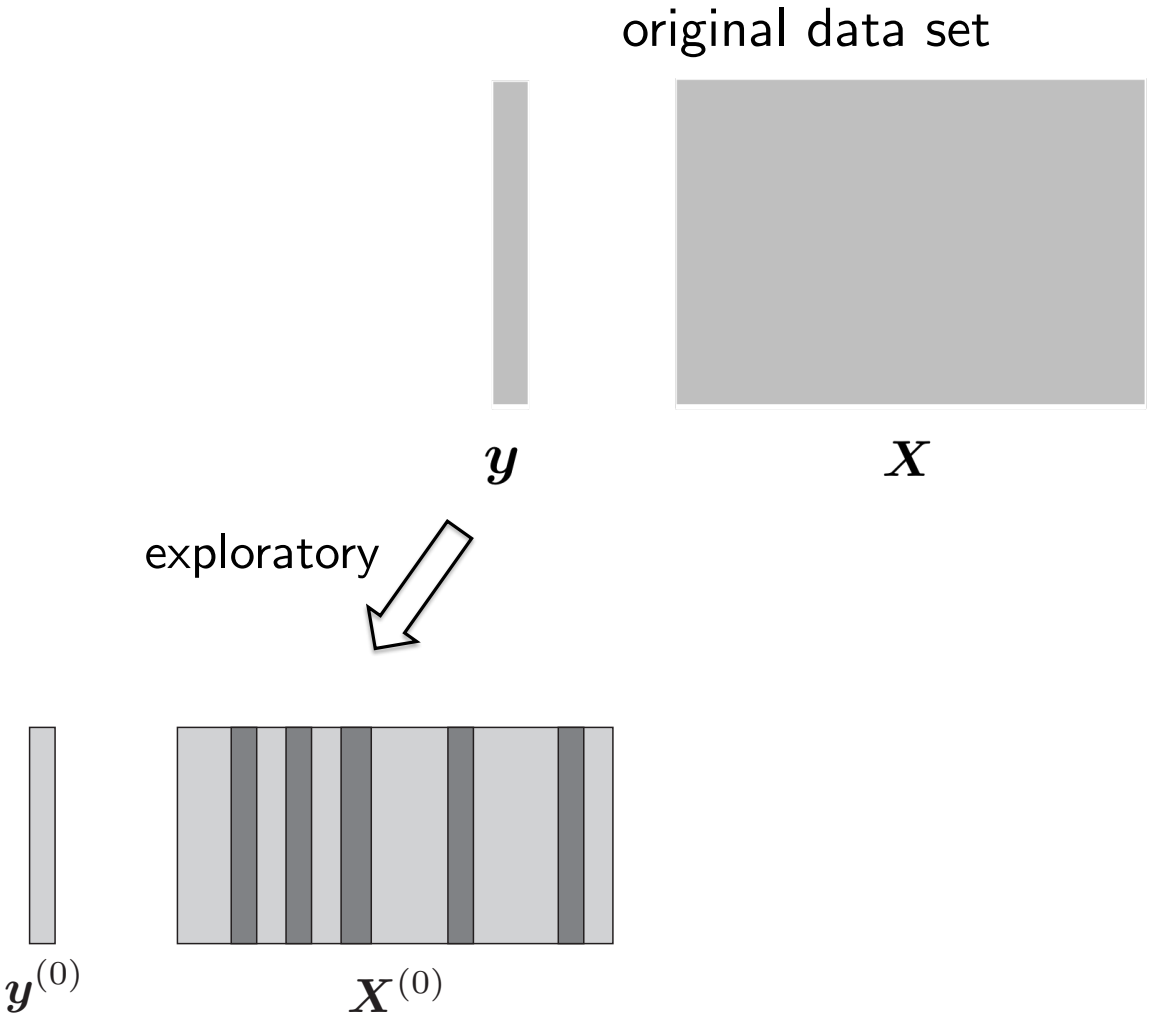


$p > n \longrightarrow$ cannot construct knockoffs as before

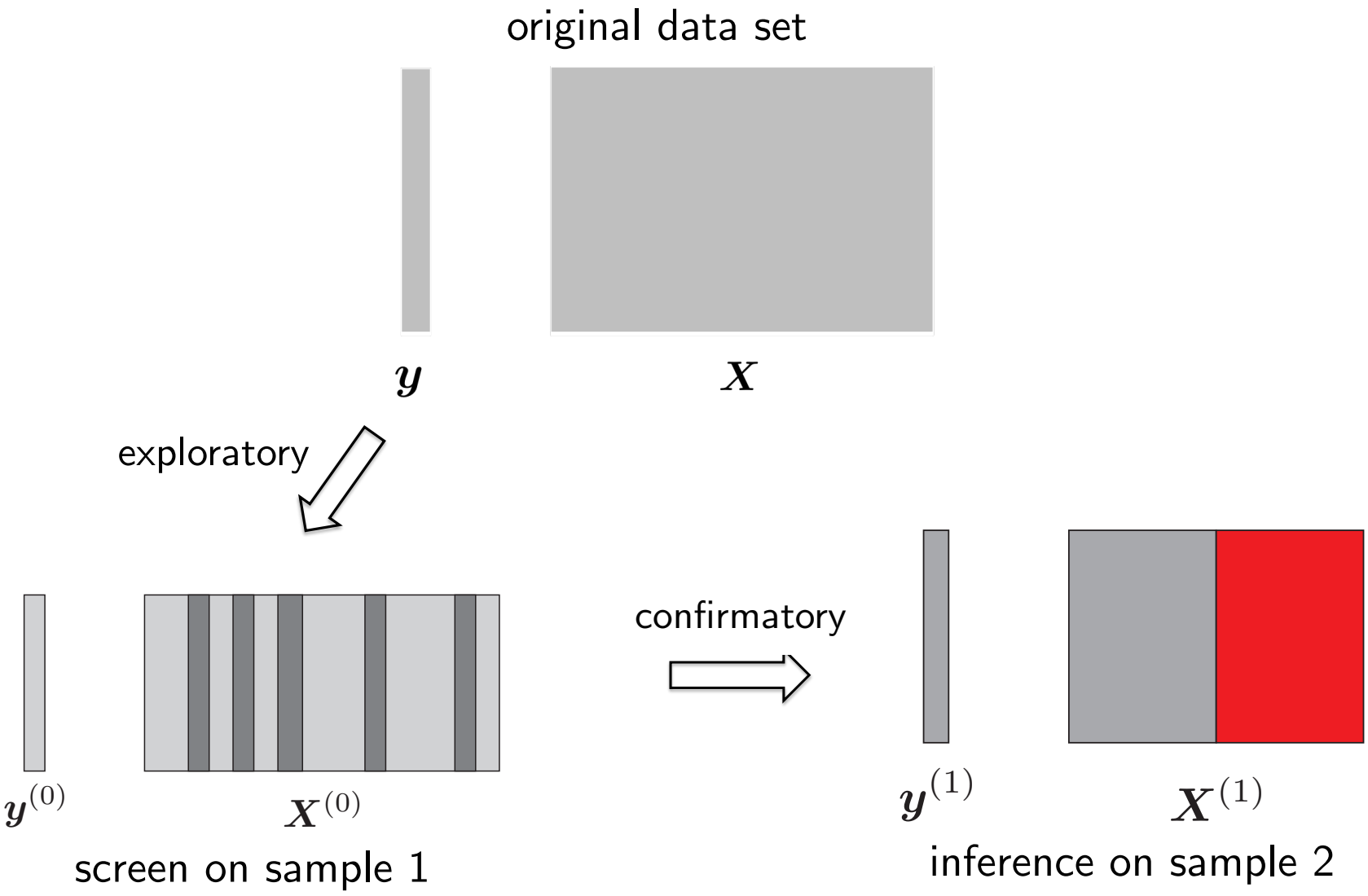
$$\begin{aligned} \tilde{X}_j' \tilde{X}_k &= X_j' X_k \qquad \forall\, j, k \\ \tilde{X}_j' X_k &= X_j' X_k \qquad \forall\, j \neq k \end{aligned} \quad \Longrightarrow \quad \tilde{X}_j = X_j \quad \forall j$$

High dimensional knockoffs: screen and confirm

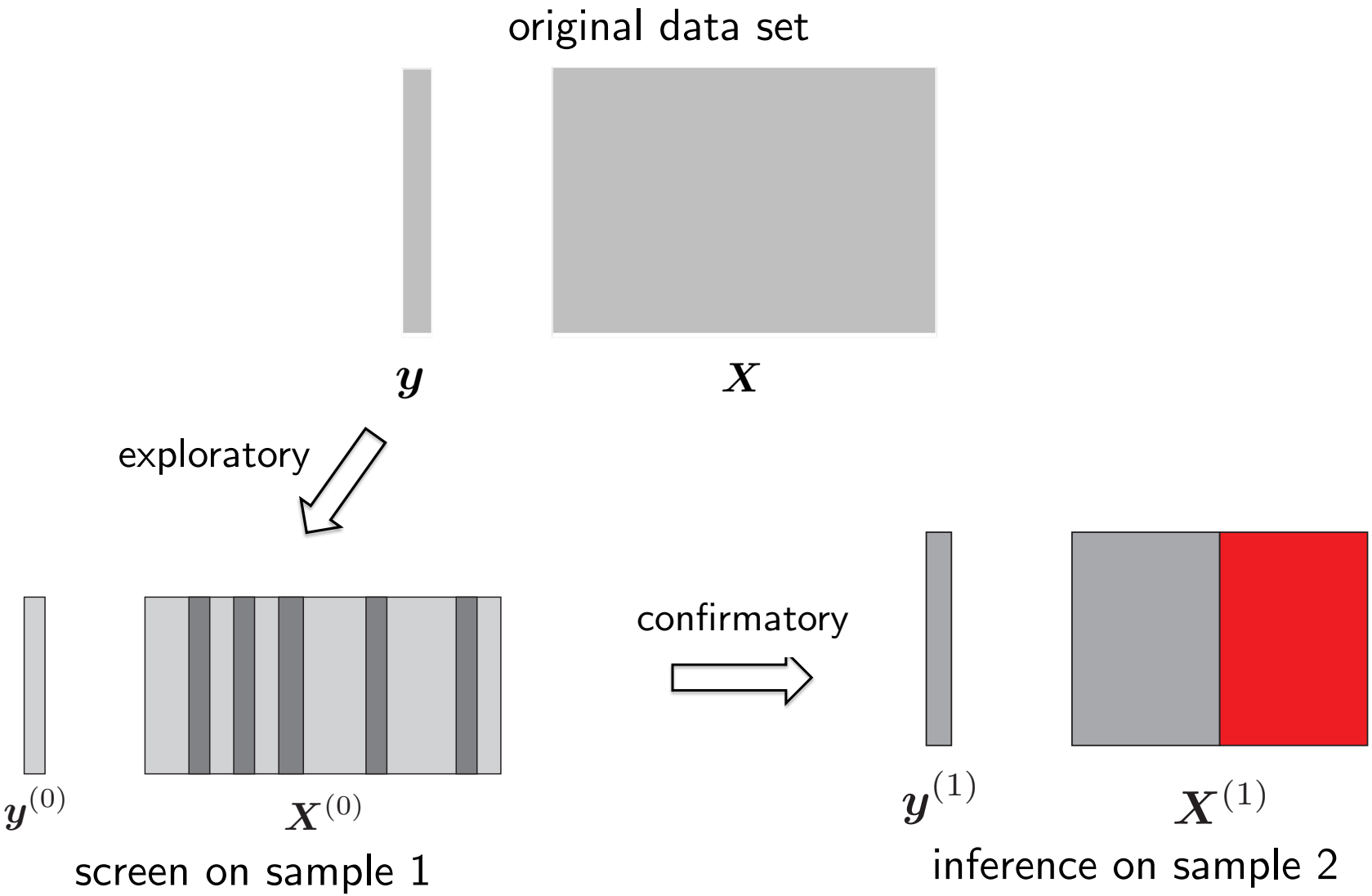
High dimensional knockoffs: screen and confirm



High dimensional knockoffs: screen and confirm



High dimensional knockoffs: screen and confirm



- Theory (Barber and C., '16)
- Safe data re-use to improve power (Barber and C., '16)

Some extensions

$$y = \underbrace{\left(X_1 \right)}_{n \times p_1} \cdot \beta_1 + \underbrace{\left(X_2 \right)}_{n \times p_2} \cdot \beta_2 + \cdots + \mathcal{N}(0, \sigma^2 I_n)$$

- Group sparsity — build knockoffs at the group-wise level

  Dai & Barber 2015

- Identify key groups with PCA — build knockoffs only for the top PC in each group

  Chen, Hou, Hou 2017

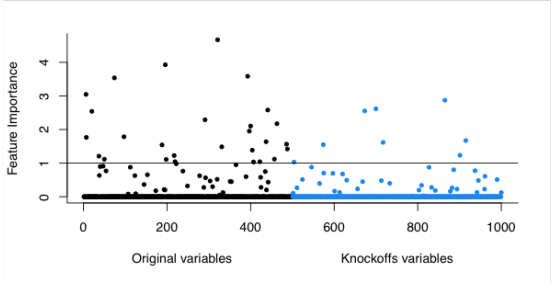
- Build knockoffs only for prototypes selected from each group

  Reid & Tibshirani 2015

- Multilayer knockoffs to control FDR at the individual and group levels simultaneously

  Katsevich & Sabatti 2017

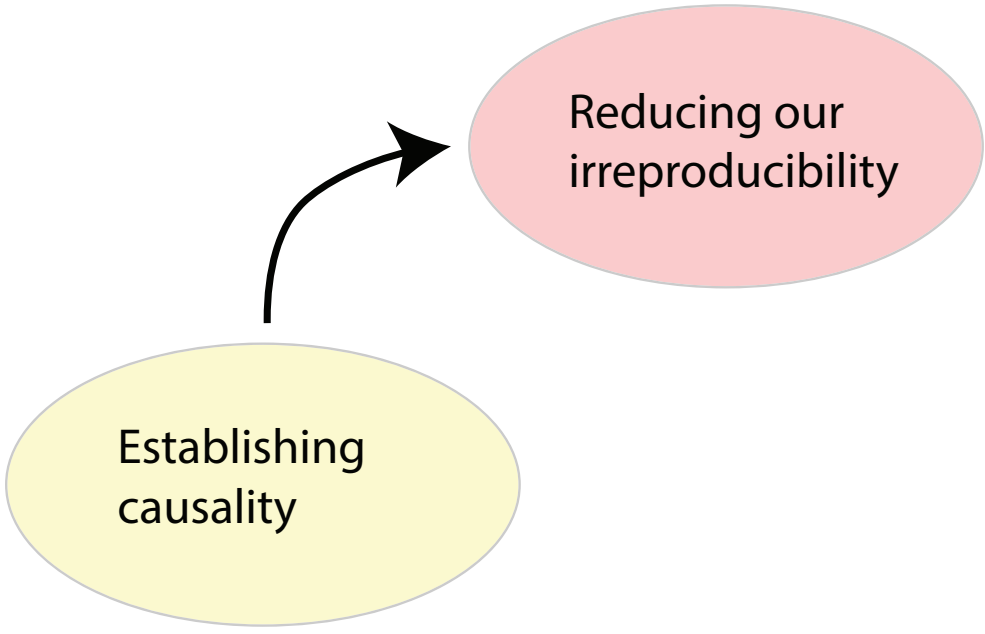
Learning from data is not trivial



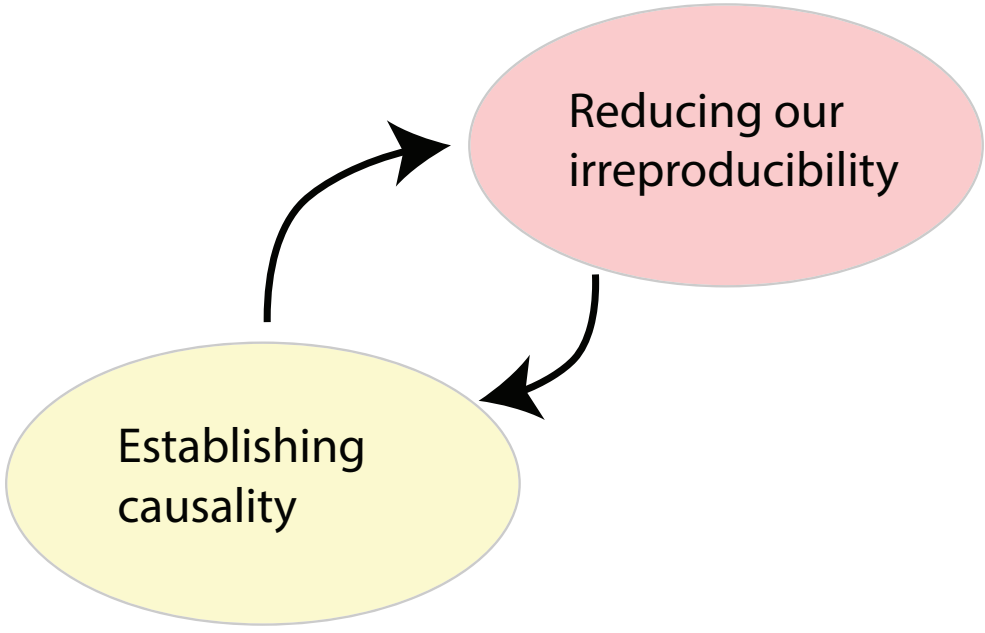
'Wrapper' around black-box algorithm rigorously addresses reproducibility issue

How to make valid knockoffs (controls)?	Importance of correct statistical reasoning
Which level of significance is appropriate?	Importance of mathematics (martingale theory)
Sensitivity to modeling assumptions	Importance of mathematics

Beyond replicability: grand challenges in data-driven science



Beyond replicability: grand challenges in data-driven science

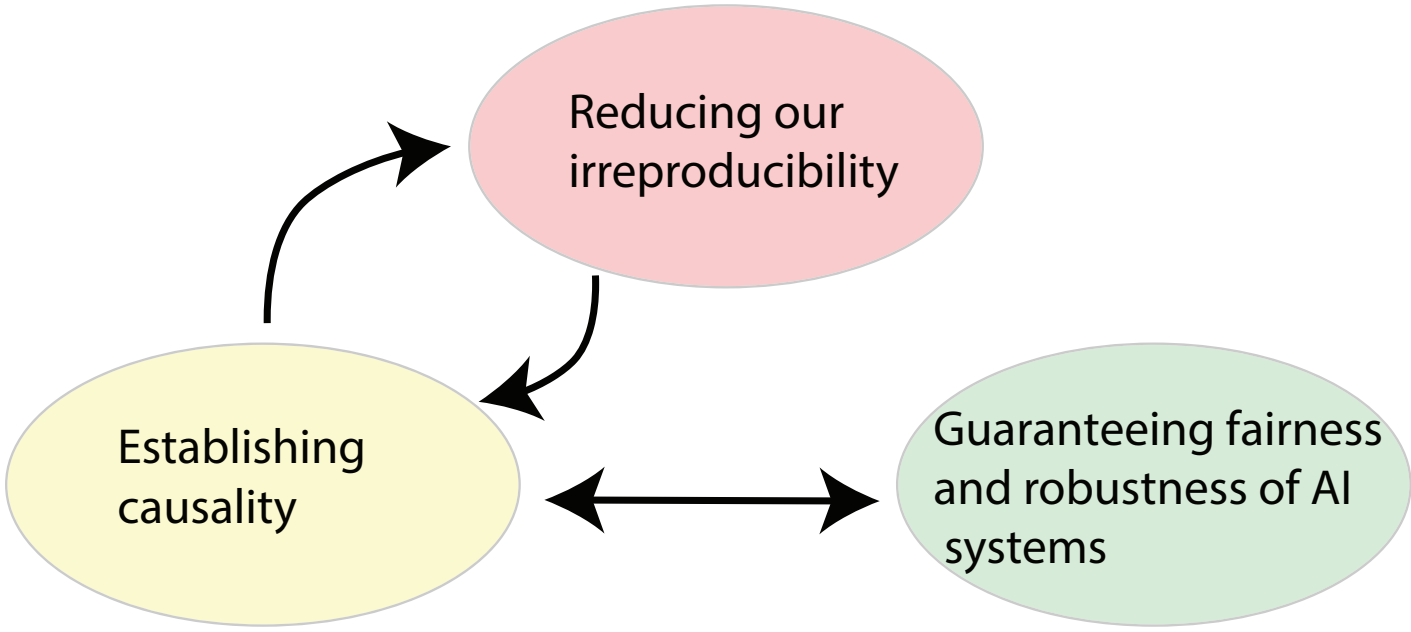


In some cases, variables with the property

$$p(\text{response} \,|\, \text{variable}, \text{ others}) \neq p(\text{response} \,|\, \text{others})$$

are 'causal'

Beyond replicability: grand challenges in data-driven science



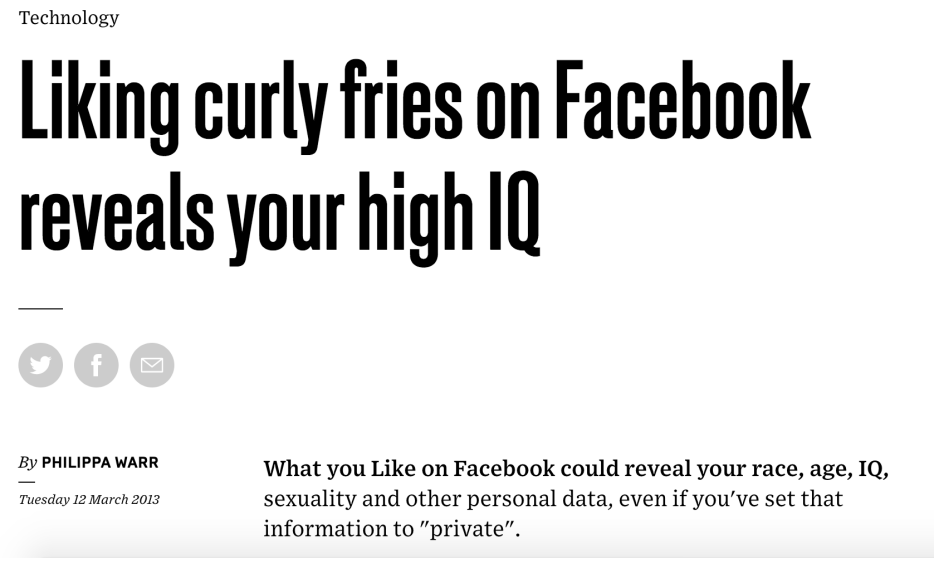
In some cases, variables with the property

$$p(\text{response} \,|\, \text{variable}, \text{ others}) \neq p(\text{response} \,|\, \text{others})$$

are 'causal'

If predictive algorithm uses causal variables, then it is likely to be fair

This is not just about not being wrong (irreproducibility)



Technology

Liking curly fries on Facebook reveals your high IQ

By **PHILIPPA WARR**

Tuesday 12 March 2013

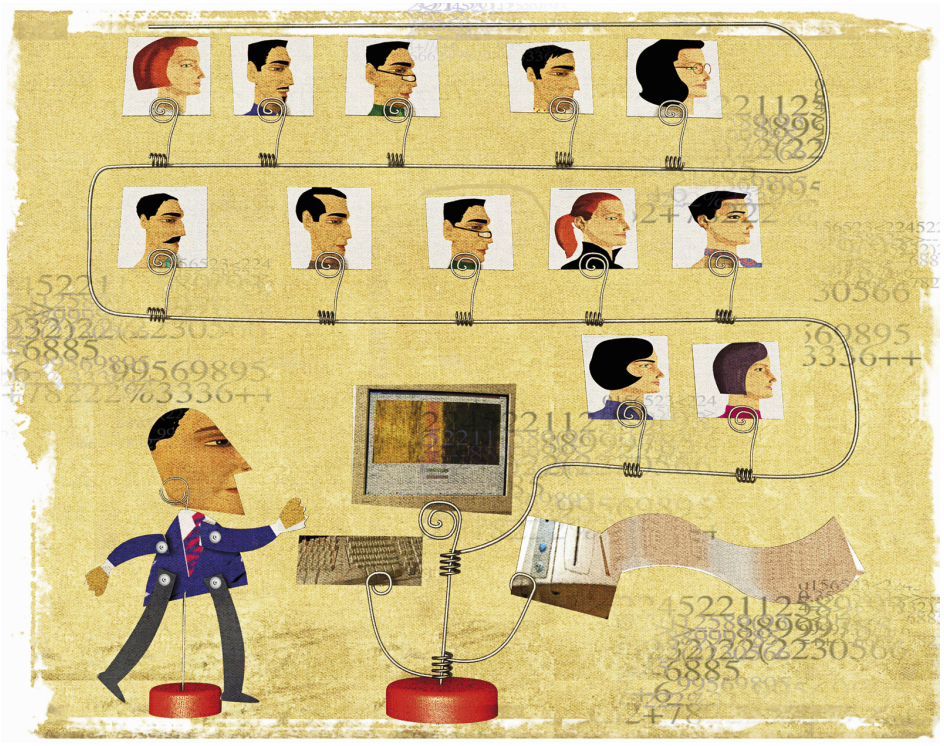
What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

Robustness?

Would want predictions to be valid in different samples collected in different circumstances

"Constant conjunction" is a property of causal effects (Hume)

Fairness: can computer programs be racist and sexist?



Guido Rosa/Getty Images/Ikon Images

Blind application of machine learning runs risk of amplifying biases and prejudices

Identifying variables $\rightsquigarrow$ chance to scrutinize model built from one sample:

- Do we believe these variables are "structurally" important, or are they just reflecting a spurious association in this sample?
- Are we learning something about the world or reifying our prejudices?