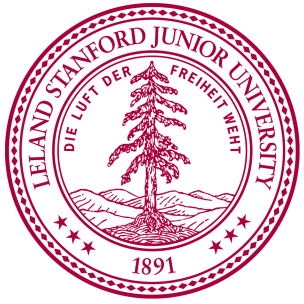
# Statistics for the Big Data Era

Emmanuel Candès




*Leçons Jacques-Louis Lions 2017, Paris, March 2017*

# Goals for these lectures

(1) Emphasize the importance of modern statistical inference problems

(2) Discuss ways in which the statistical community responds to pressing contemporary problems

# Agenda

## Lecture 2: Controlled variable selection via the knockoff filter

- Feature selection
- Model-free knockoffs
- Feature selection via the knockoff filter
- FDR control
- Examples
- Knockoff filter for linear regression with fixed design

# Collaborators

Knockoffs v1.0

- Rina Foygel Barber (Chicago)

Knockoffs v2.0

- Lucas Janson (Stanford)
- Yingying Fan (USC)
- Jinchi Lv (USC)

# Yesterday: big data and a new scientific paradigm

Collect data first $\Longrightarrow$ Ask questions later

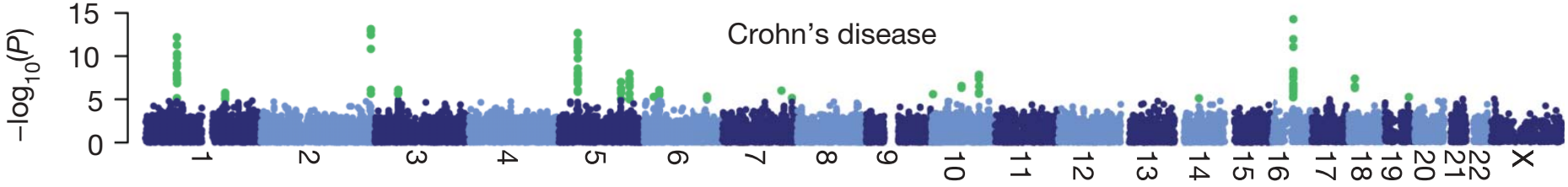
- Large data sets available prior to formulation of hypotheses
- Need to quantify "reliability" of hypotheses generated by data snooping

Very different from hypothesis-driven research

# If ignored... selective reporting/statistical bias

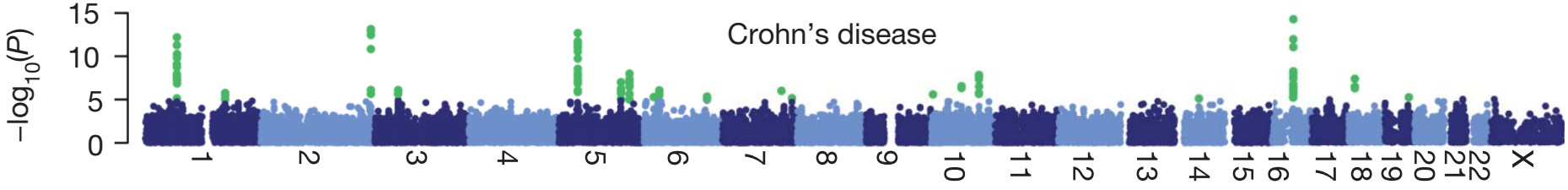
# *The Feature Selection Problem*

# Contemporary problem




- $Y \in \{0, 1\}$: disease status (e.g. Crohn's disease)
- $X_j \in \{0, 1, 2\}$: # of minor alleles at marker $j$ (SNP info)

# Contemporary problem




- $Y \in \{0, 1\}$: disease status (e.g. Crohn's disease)
- $X_j \in \{0, 1, 2\}$: # of minor alleles at marker $j$ (SNP info)

Which genetic variations are important for understanding the risk of a disease?

# A master problem of (big) data science

- Which genetic features influence the risk of a disease?
- Which electronic medical record entries influence future medical costs?
- Which demographic or socioeconomic variables influence political opinions?
- Which software characteristics predict user engagement?

# A master problem of (big) data science

- Which genetic features influence the risk of a disease?
- Which electronic medical record entries influence future medical costs?
- Which demographic or socioeconomic variables influence political opinions?
- Which software characteristics predict user engagement?

Response $Y$ and thousands/millions of covariates $X$

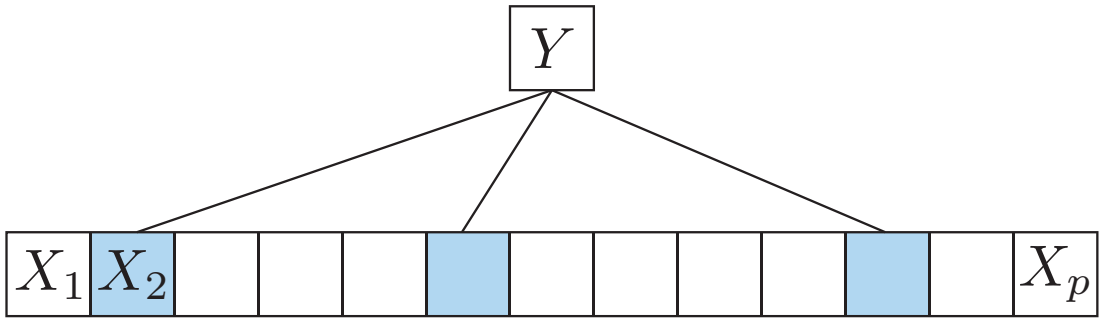
Which ones are important?

Distribution of $Y \mid X$ depends on $X$ through which variables?

# Selection problem

Subset $\mathcal{S}$ of relevant variables

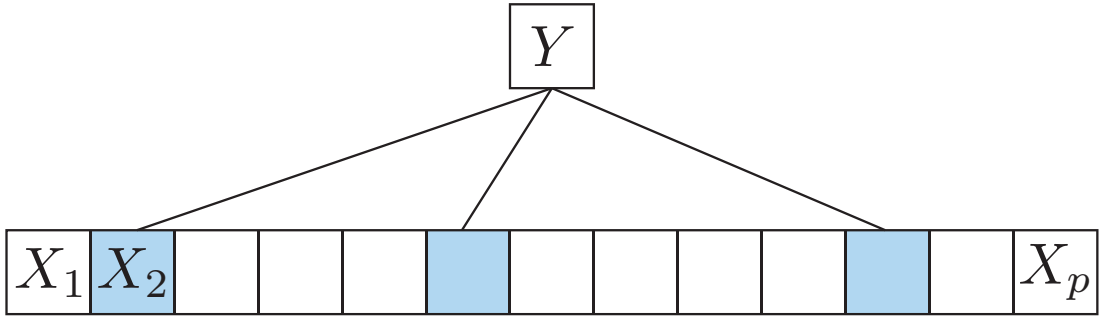
$$Y \mid \{X_j\}_{j \in 1:p} \quad \stackrel{d}{=} \quad Y \mid \{X_j\}_{j \in \mathcal{S}}$$



# Selection problem

Subset $\mathcal{S}$ of relevant variables

$$Y \,|\, \{X_j\}_{j \in 1:p} \quad \stackrel{d}{=} \quad Y \,|\, \{X_j\}_{j \in \mathcal{S}}$$

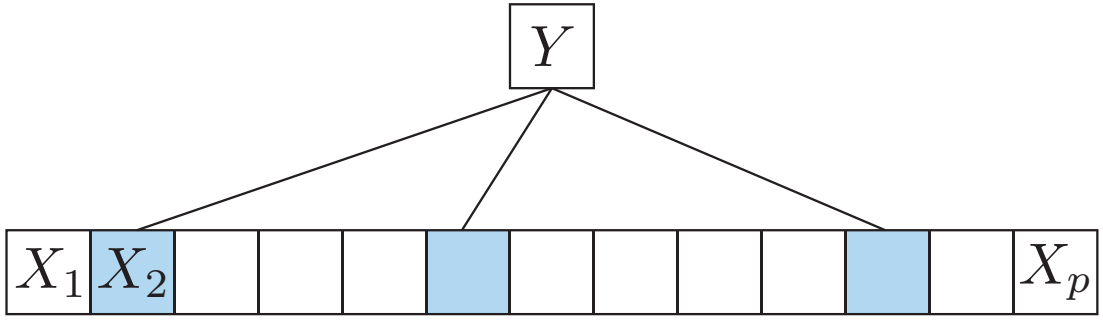



Goal: select set $\hat{\mathcal{S}}$ of features $X_j$ that are likely to be relevant
without too many false positives – do not run into the problem of irreproducibilty

# Selection problem

Subset $\mathcal{S}$ of relevant variables

$$Y \,|\, \{X_j\}_{j \in 1:p} \quad \stackrel{d}{=} \quad Y \,|\, \{X_j\}_{j \in \mathcal{S}}$$




Goal: select set $\hat{\mathcal{S}}$ of features $X_j$ that are likely to be relevant
without too many false positives – do not run into the problem of irreproducibilty

$$\underbrace{\text{FDR}}_{\text{False discovery rate}} = \mathbb{E} \underbrace{\frac{\#\text{ false positives}}{\#\text{ features selected}}}_{\text{False discovery proportion}} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \setminus \mathcal{S}|}{|\hat{\mathcal{S}}|}\right]$$

# Problem statement

## Working definition of null (unimportant) variables

Say $j \in \mathcal{H}_0$ is null iff $Y \perp\!\!\!\perp X_j \,|\, X_{-j}$

# Problem statement

## Working definition of null (unimportant) variables

Say $j \in \mathcal{H}_0$ is null iff $Y \perp\!\!\!\perp X_j \,|\, X_{-j}$

Local Markov property $\implies$ non nulls are smallest subset $\mathcal{S}$ (Markov blanket) s.t.

$$Y \perp\!\!\!\perp \{X_j\}_{j \in \mathcal{S}^c} \,|\, \{X_j\}_{j \in \mathcal{S}}$$

# Problem statement

**Working definition of null (unimportant) variables**

Say $j \in \mathcal{H}_0$ is null iff $Y \perp\!\!\!\perp X_j \,|\, X_{-j}$

Local Markov property $\Longrightarrow$ non nulls are smallest subset $\mathcal{S}$ (Markov blanket) s.t.

$$Y \perp\!\!\!\perp \{X_j\}_{j \in \mathcal{S}^c} \,|\, \{X_j\}_{j \in \mathcal{S}}$$

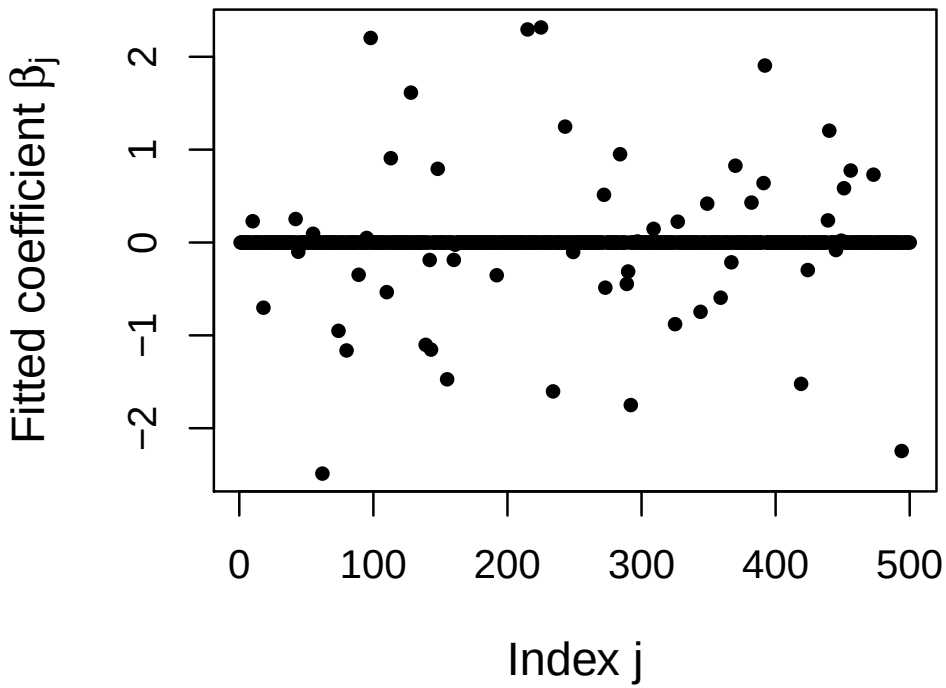
Logistic model: $$\mathbb{P}(Y = 0|X) = \frac{1}{1 + e^{X^\top \beta}}$$

If variables $X_{1:p}$ are not perfectly dependent, then $j \in \mathcal{H}_0 \iff \beta_j = 0$

# Subtlety of the selection problem

(Logistic) LASSO model

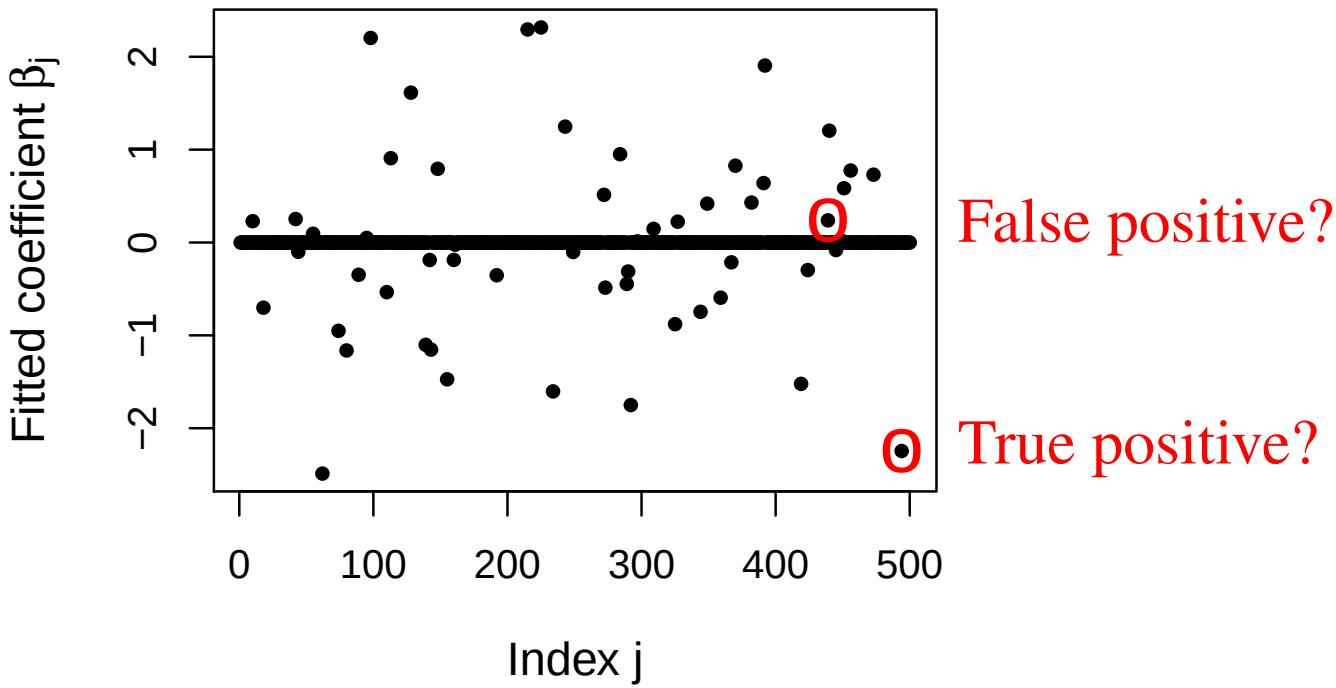
$$\min_{\boldsymbol{b}\in\mathbb{R}^p} \quad -\ell(\boldsymbol{y};\, \boldsymbol{X}\boldsymbol{b}) + \lambda\|\boldsymbol{b}\|_1$$



# Subtlety of the selection problem

(Logistic) LASSO model

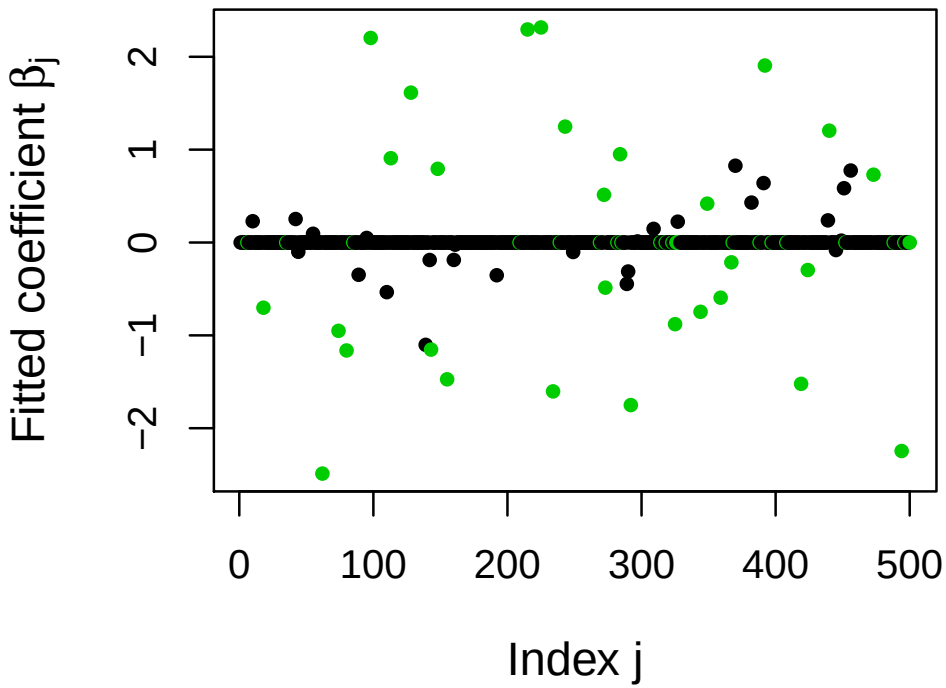
$$\min_{\boldsymbol{b}\in\mathbb{R}^p} \quad -\ell(\boldsymbol{y};\, \boldsymbol{X}\boldsymbol{b}) + \lambda\|\boldsymbol{b}\|_1$$



# Subtlety of the selection problem

(Logistic) LASSO model

$$\min_{\boldsymbol{b}\in\mathbb{R}^p} \quad -\ell(\boldsymbol{y};\, \boldsymbol{X}\boldsymbol{b}) + \lambda\|\boldsymbol{b}\|_1$$

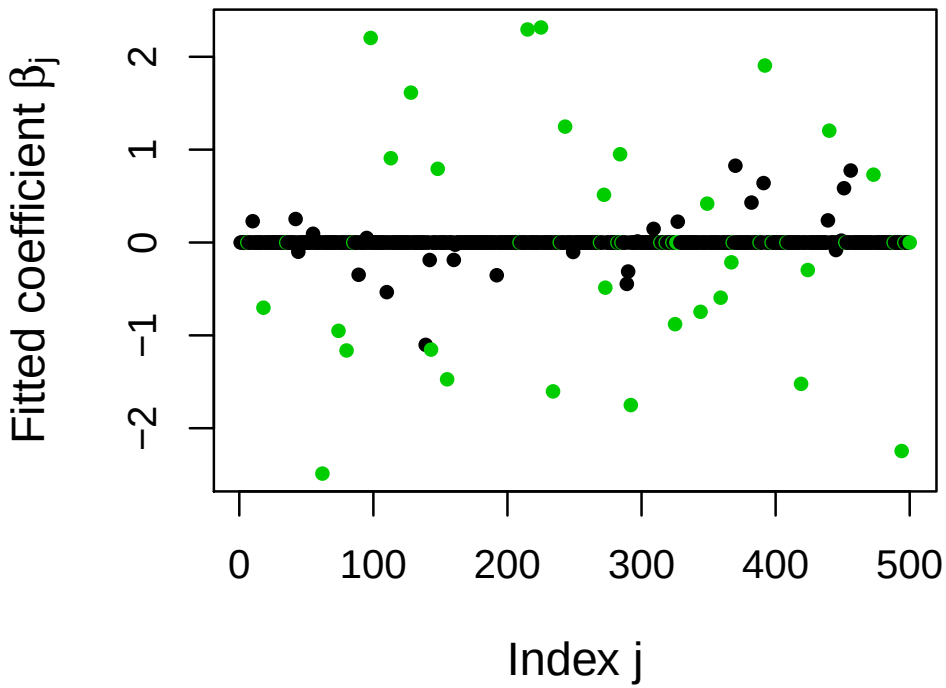



$\text{FDP} = \frac{26}{55} = 47\%$

# Subtlety of the selection problem

(Logistic) LASSO model

$$\min_{\boldsymbol{b}\in\mathbb{R}^p} \quad -\ell(\boldsymbol{y};\, \boldsymbol{X}\boldsymbol{b}) + \lambda\|\boldsymbol{b}\|_1$$



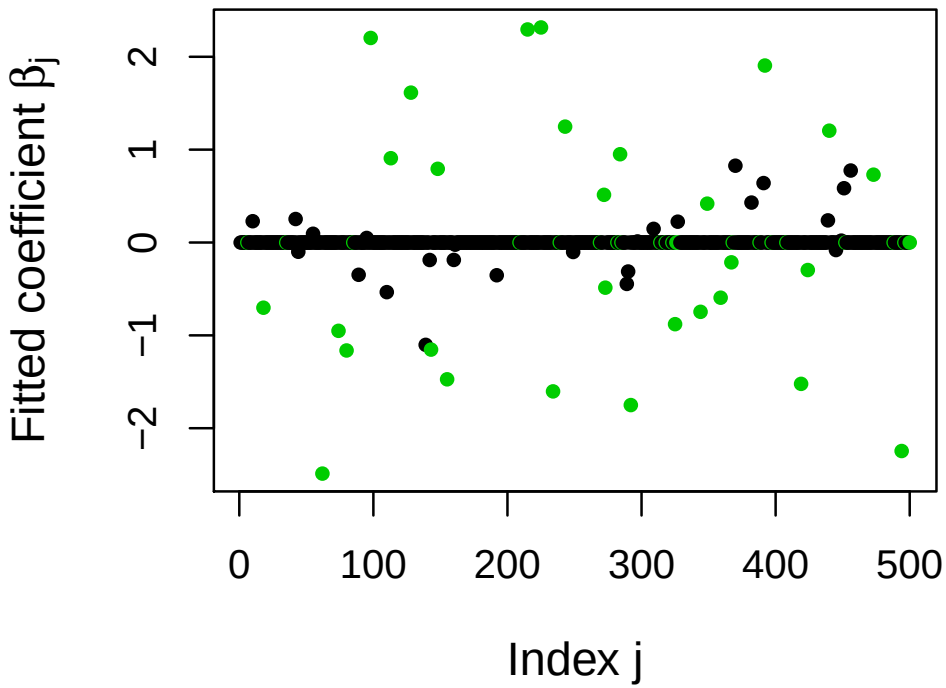

$\text{FDP} = \frac{26}{55} = 47\%$

Estimate FDP?

# Subtlety of the selection problem

(Logistic) LASSO model

$$\min_{\boldsymbol{b}\in\mathbb{R}^p} \quad -\ell(\boldsymbol{y};\, \boldsymbol{X}\boldsymbol{b}) + \lambda\|\boldsymbol{b}\|_1$$




FDP $= \frac{26}{55} = 47\%$

Estimate FDP? ...

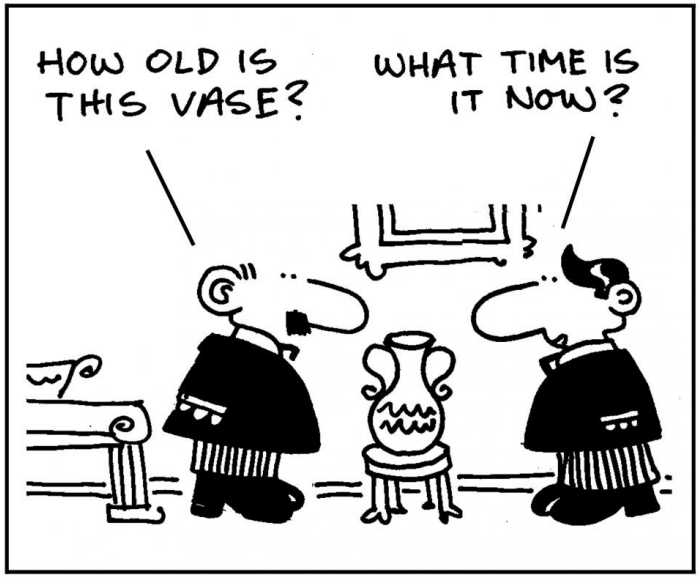
# *Model-Free Knockoffs*

*Joint with Janson, Fan and Lv*

# Knockoffs

*It's not a name brand bag*
*just a cheap knockoff*

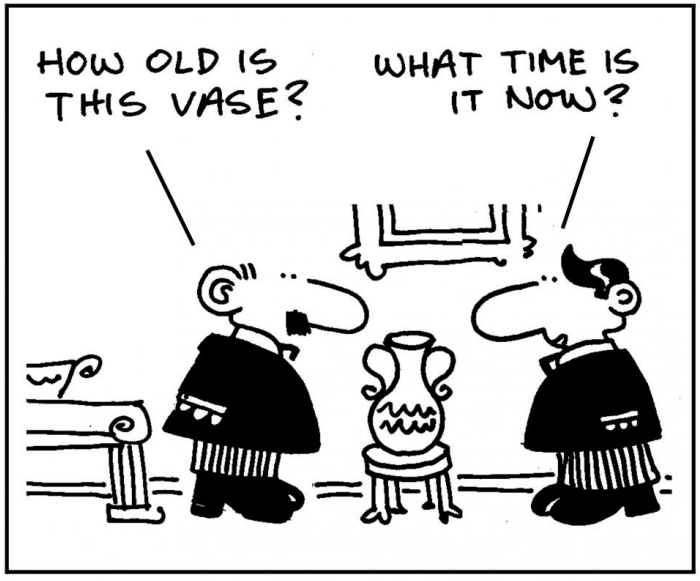
Thesaurize.com



# Knockoffs

*It's not a name brand bag*
*just a cheap knockoff*

Thesaurize.com




For each feature $X_j$, construct a knockoff version $\tilde{X}_j$
The knockoffs serve as a "control group" $\Rightarrow$ can estimate FDP

# Model-free knockoff variables

i.i.d. samples $(X^{(i)}, Y^{(i)}) \sim F_{XY}$

- Distribution of $X$ known
- Distribution of $Y \,|\, X$ (likelihood) completely unknown

# Model-free knockoff variables

i.i.d. samples $(X^{(i)}, Y^{(i)}) \sim F_{XY}$

- Distribution of $X$ known
- Distribution of $Y \,|\, X$ (likelihood) completely unknown

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

# Model-free knockoff variables

i.i.d. samples $(X^{(i)}, Y^{(i)}) \sim F_{XY}$

- Distribution of $X$ known
- Distribution of $Y \,|\, X$ (likelihood) completely unknown

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

(1) Pairwise exchangeability

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

e.g.

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

# Model-free knockoff variables

i.i.d. samples $(X^{(i)}, Y^{(i)}) \sim F_{XY}$

- Distribution of $X$ known
- Distribution of $Y \,|\, X$ (likelihood) completely unknown

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

(1) Pairwise exchangeability

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

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

# Model-free knockoff variables

i.i.d. samples $(X^{(i)}, Y^{(i)}) \sim F_{XY}$

- Distribution of $X$ known
- Distribution of $Y \,|\, X$ (likelihood) completely unknown

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

(1) Pairwise exchangeability

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

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

No need for new data or experiment

# Why?

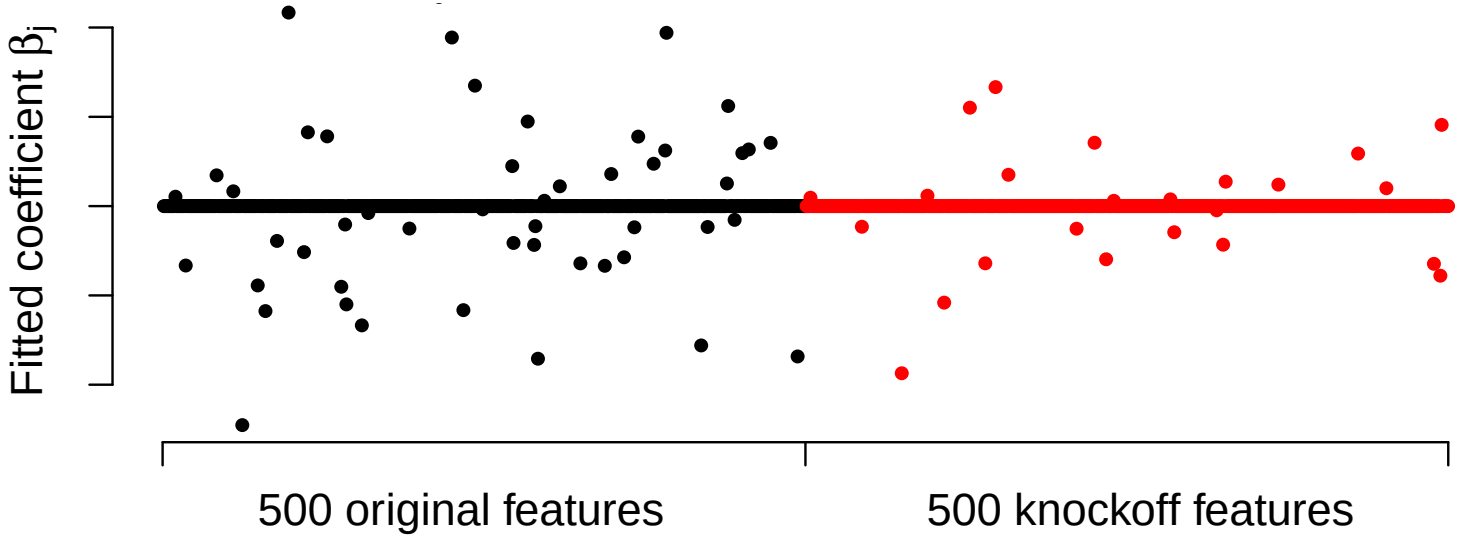
Run same procedure on original and knockoff variables 'serving as controls'

$$\min_{\boldsymbol{b} \in \mathbb{R}^{2p}} \quad -\ell(\boldsymbol{y};\, [\boldsymbol{X}\ \tilde{\boldsymbol{X}}]\, \boldsymbol{b}) + \lambda \|\boldsymbol{b}\|_1$$

# Why?

Run same procedure on original and knockoff variables 'serving as controls'

$$\min_{\boldsymbol{b} \in \mathbb{R}^{2p}} \quad -\ell(\boldsymbol{y};\, [\boldsymbol{X}\ \tilde{\boldsymbol{X}}]\, \boldsymbol{b}) + \lambda \|\boldsymbol{b}\|_1$$




Lasso selects 49 original features & 24 knockoff features

$\Longrightarrow$ probably $\approx$ 24 false positives among 49 original features

# Why?

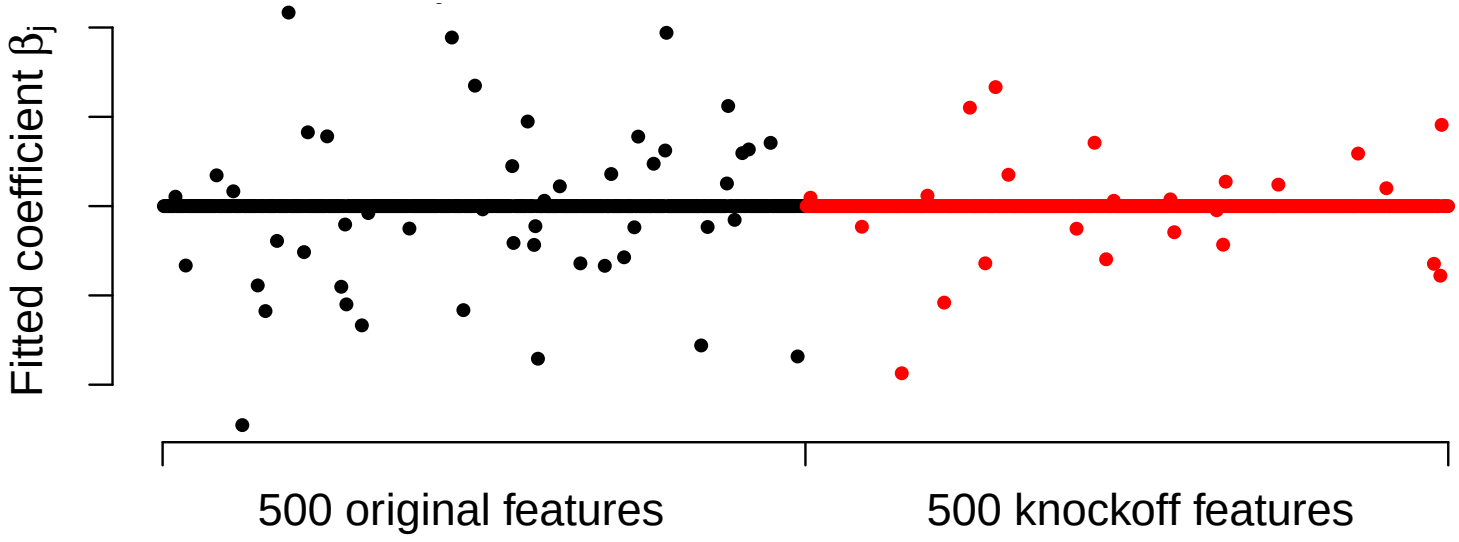
Run same procedure on original and knockoff variables 'serving as controls'

$$\min_{\boldsymbol{b} \in \mathbb{R}^{2p}} \quad -\ell(\boldsymbol{y};\, [\boldsymbol{X}\ \tilde{\boldsymbol{X}}]\, \boldsymbol{b}) + \lambda \|\boldsymbol{b}\|_1$$

# Why?

Run same procedure on original and knockoff variables 'serving as controls'

$$\min_{\boldsymbol{b} \in \mathbb{R}^{2p}} \quad -\ell(\boldsymbol{y};\, [\boldsymbol{X}\ \tilde{\boldsymbol{X}}]\, \boldsymbol{b}) + \lambda \|\boldsymbol{b}\|_1$$

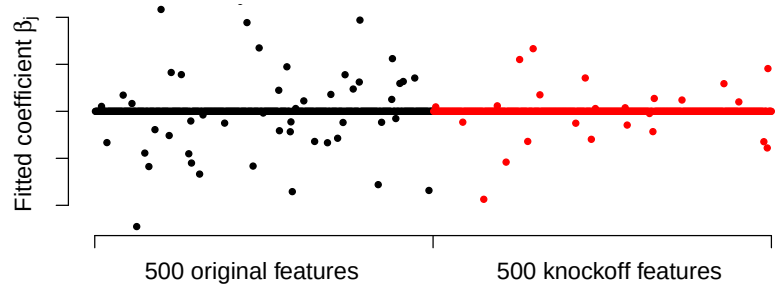



Lasso selects 49 original features & 24 knockoff features

$\Longrightarrow$ probably $\approx$ 24 false positives among 49 original features

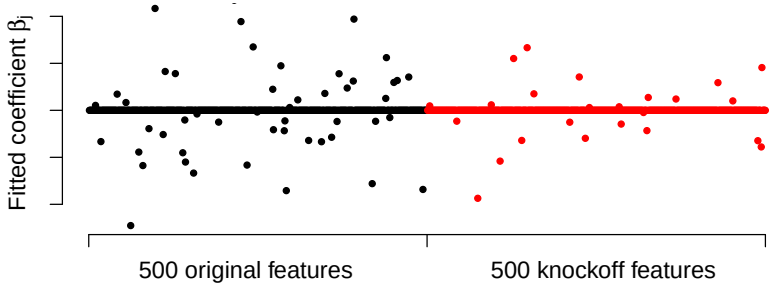
# Why?

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$



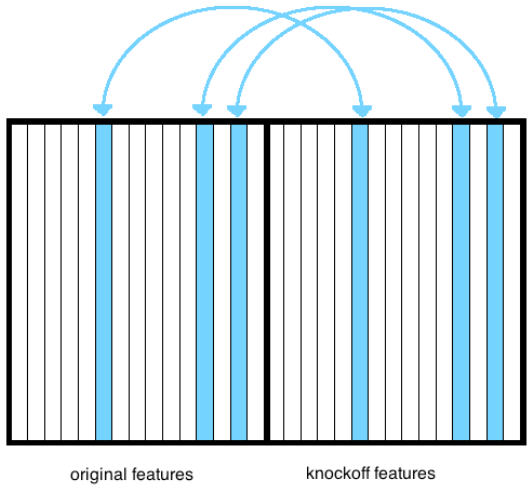
# Why?

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$



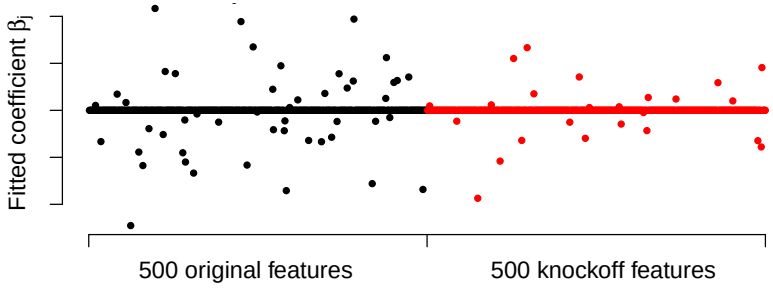

Swapping originals and knockoffs swaps the $Z$'s

$$\overbrace{(Z_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_1, Z_2, Z_3)}^{(Z, \tilde{Z})_{\text{swap}\{2,3\}}} = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}\{2,3\}}, \boldsymbol{y})$$



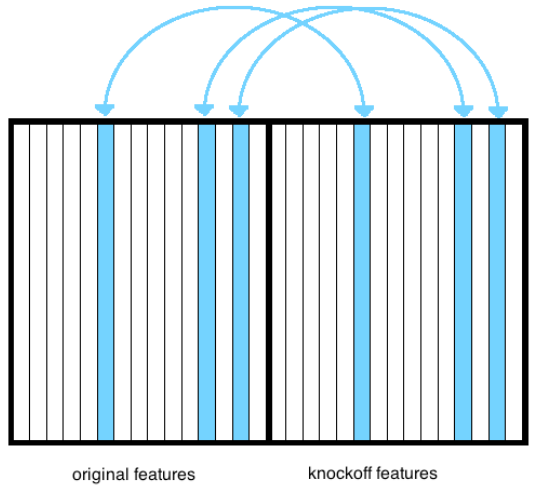
# Why?

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$




Swapping originals and knockoffs swaps the $Z$'s

$$\overbrace{(Z_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_1, Z_2, Z_3)}^{(Z,\, \tilde{Z})_{\text{swap}\{2,3\}}} = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}]_{\text{swap}\{2,3\}}, \boldsymbol{y})$$



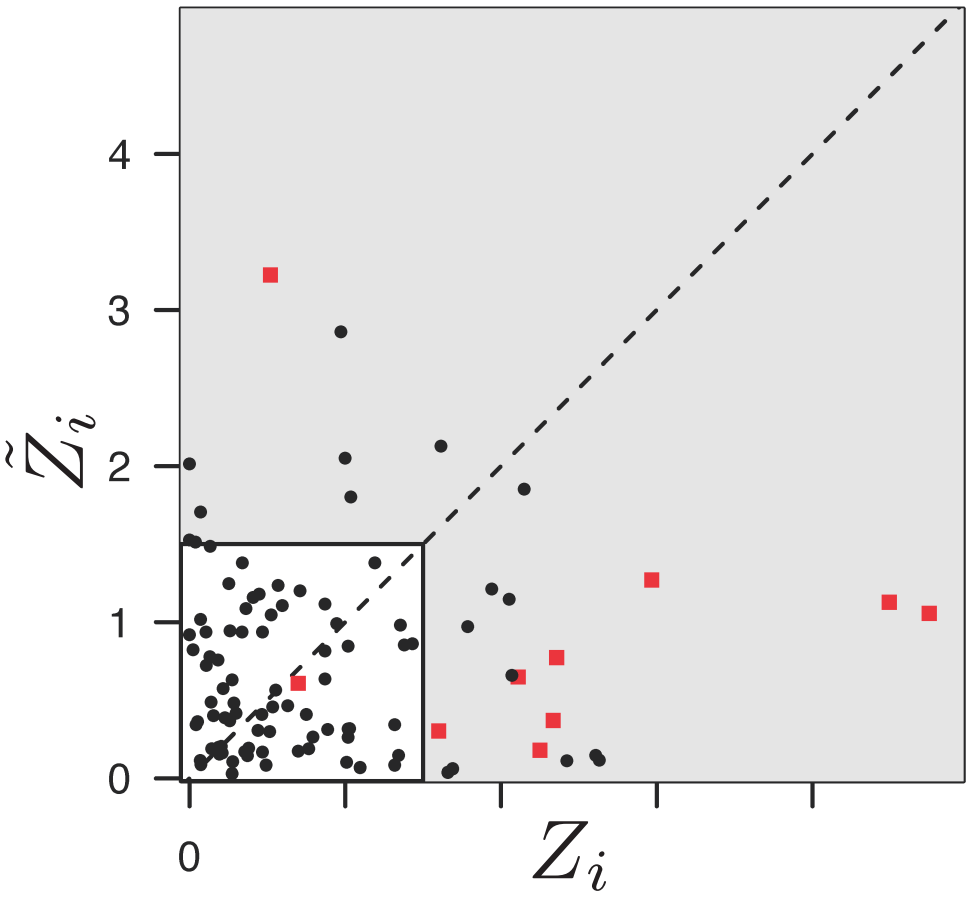

## Theorem (C., Fan, Janson Lv ('16))

*For any subset $\mathcal{T} \subset \mathcal{H}_0$ of nulls ($j \in \mathcal{H}_0$ iff $Y \perp\!\!\!\perp X_j \,|\, X_{-j}$)*

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

*This holds no matter the relationship between $Y$ and $X$*

# Exchangeability of the nulls




Signal
Null

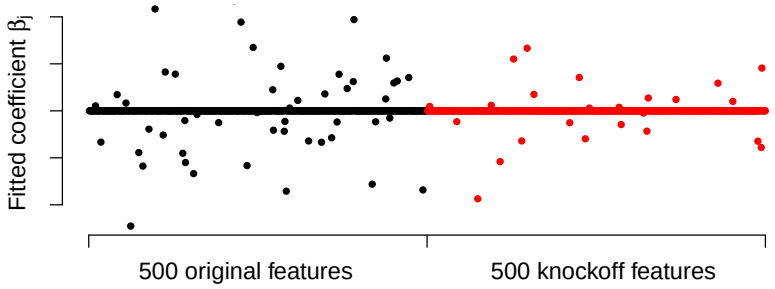
exchangeable

$$(Z_1, Z_2, Z_3, \ldots, Z_p, \tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \ldots, \tilde{Z}_p)$$

exchangeable

# Feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$



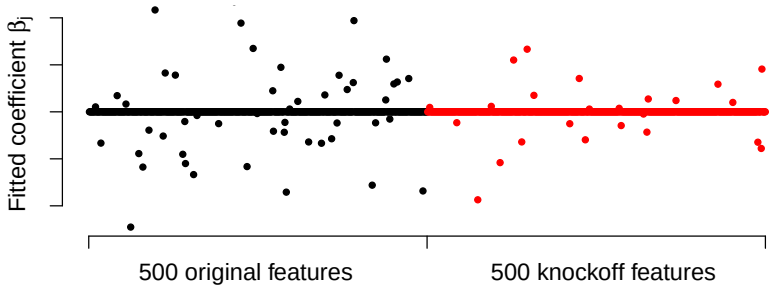

## Statistics $W_j$ (anti-symmetric)

$$W_j = W_j(Z_j, \tilde{Z}_j) \qquad W_j(\tilde{Z}_j, Z_j) = -W_j(Z_j, \tilde{Z}_j)$$

e.g. $W_j = Z_j - \tilde{Z}_j$

# Feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$




## Statistics $W_j$ (anti-symmetric)

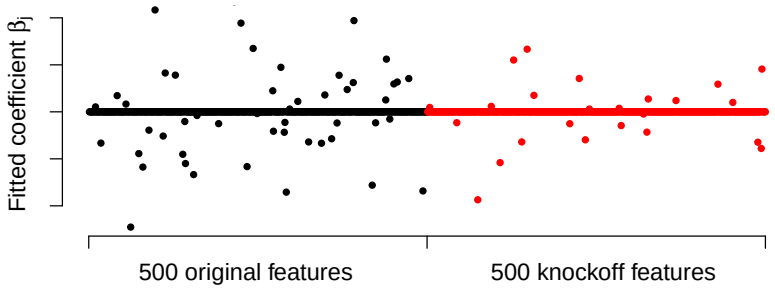
$$W_j = W_j(Z_j, \tilde{Z}_j) \qquad W_j(\tilde{Z}_j, Z_j) = -W_j(Z_j, \tilde{Z}_j)$$

e.g. $W_j = Z_j - \tilde{Z}_j$

- The null $W_j$'s are symmetrically distributed
- Conditional on $|W|$, the signs of the null $W_j$'s are i.i.d. coin flips

# Feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$

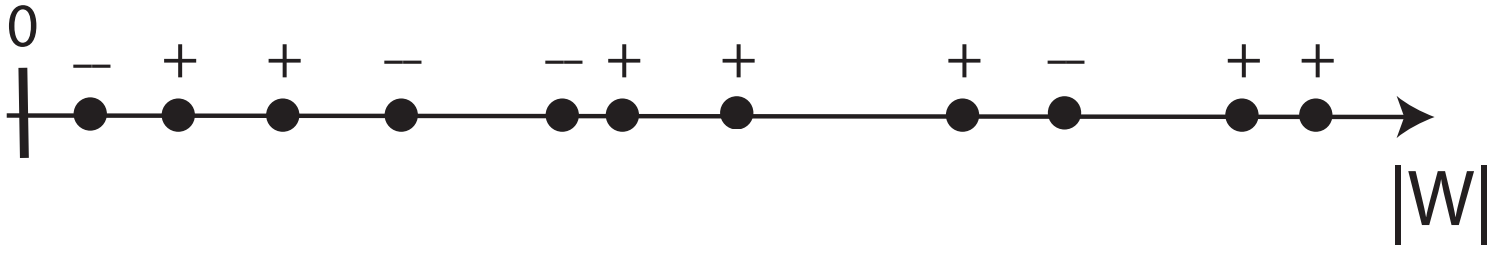



## Statistics $W_j$ (anti-symmetric)

$$W_j = W_j(Z_j, \tilde{Z}_j) \qquad W_j(\tilde{Z}_j, Z_j) = -W_j(Z_j, \tilde{Z}_j)$$

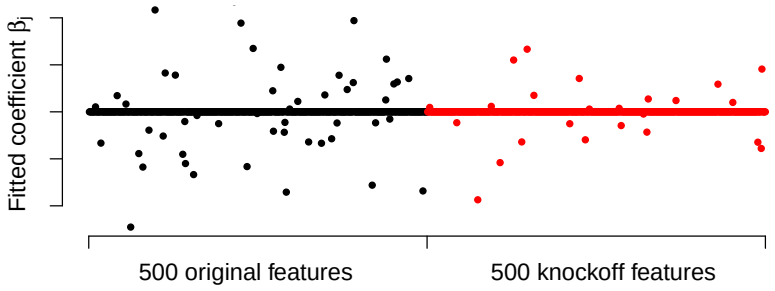
e.g. $W_j = Z_j - \tilde{Z}_j$

- The null $W_j$'s are symmetrically distributed
- Conditional on $|W|$, the signs of the null $W_j$'s are i.i.d. coin flips



# Feature importance statistics

$$(\underbrace{Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X}, \tilde{\boldsymbol{X}}], \boldsymbol{y})$$

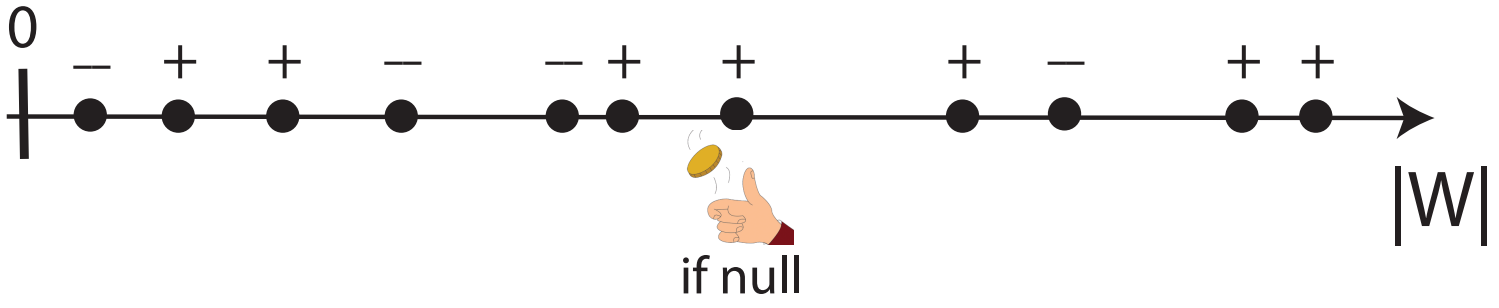



## Statistics $W_j$ (anti-symmetric)

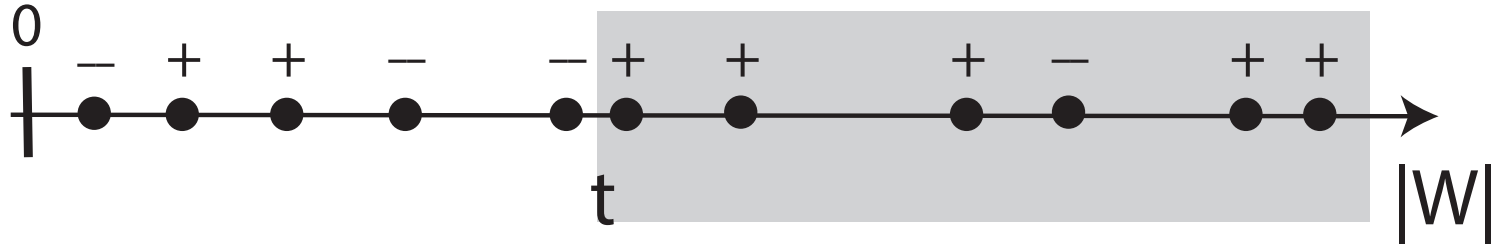
$$W_j = W_j(Z_j, \tilde{Z}_j) \qquad W_j(\tilde{Z}_j, Z_j) = -W_j(Z_j, \tilde{Z}_j)$$

e.g. $W_j = Z_j - \tilde{Z}_j$

- The null $W_j$'s are symmetrically distributed
- Conditional on $|W|$, the signs of the null $W_j$'s are i.i.d. coin flips

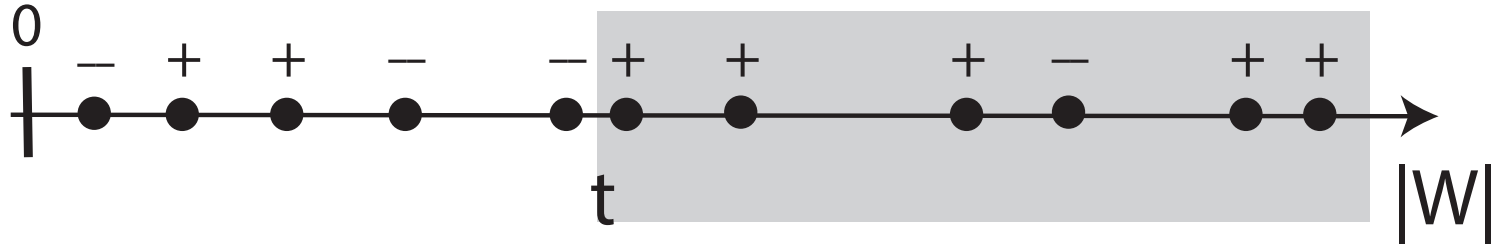


# Knockoff estimate of FDR




Interested in selecting $\{j : W_j \geq t\}$

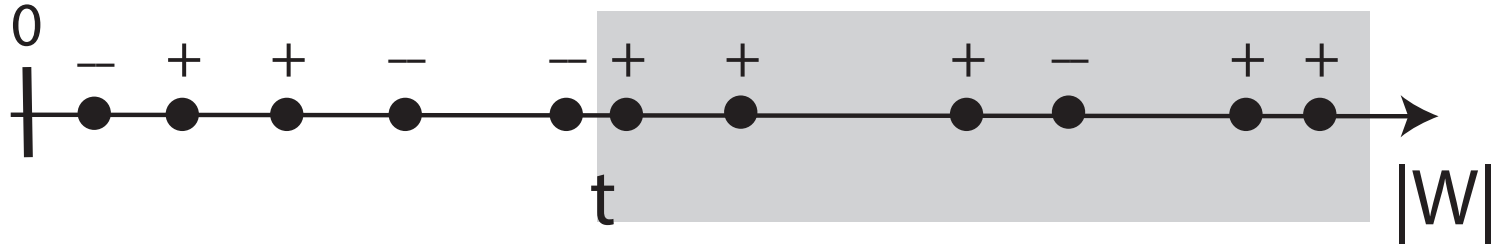
# Knockoff estimate of FDR




Interested in selecting $\{j : W_j \geq t\}$

$$\text{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1}$$

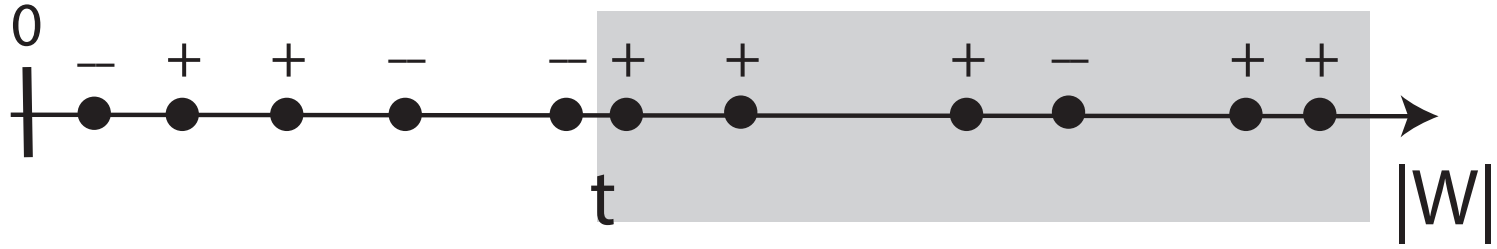
# Knockoff estimate of FDR




Interested in selecting $\{j : W_j \geq t\}$

$$\text{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j \text{ null } : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

# Knockoff estimate of FDR

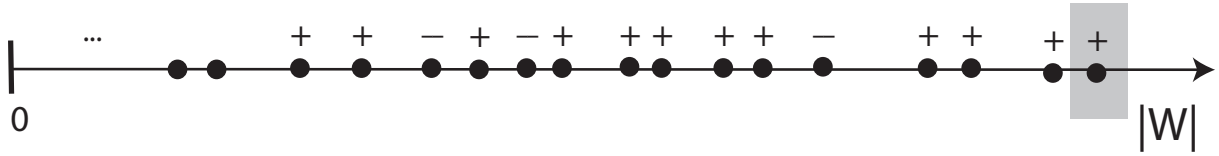



Interested in selecting $\{j : W_j \geq t\}$

$$\mathsf{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j \text{ null } : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

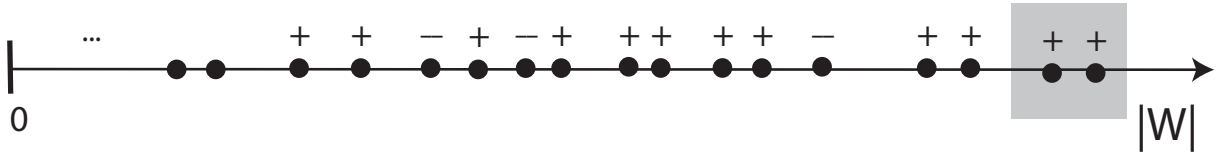
$$\leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} := \widehat{\mathsf{FDP}}(t)$$

# Selection


...
+ + − + − + + + + + − + + + +
0
|W|

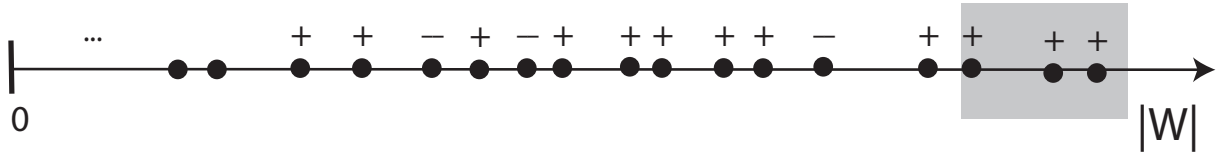

FEDERAL RESERVE NOTE
CB 77960327 B
B2
THE UNITED STATES OF AMERICA
100
CB 77960327 B
ONE HUNDRED DOLLARS

# Selection


...
+ + − + − + + + + + − + + + +
0
|W|

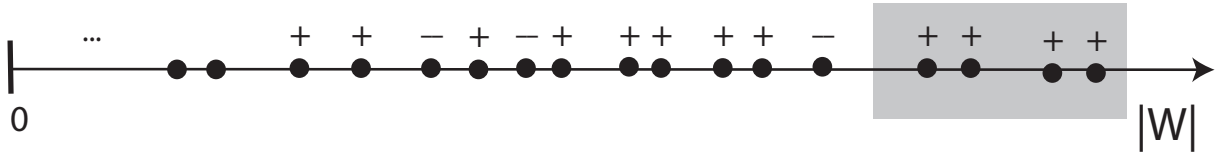

FEDERAL RESERVE NOTE
CB 77960327 B
B2
THE UNITED STATES OF AMERICA
100
CB 77960327 B
100
ONE HUNDRED DOLLARS
100
ONE HUNDRED DOLLARS

# Selection


...
+ + − + − + + + + + − + + + +
0
|W|

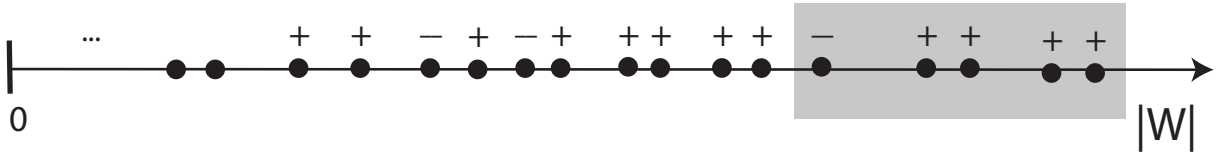

FEDERAL RESERVE NOTE
CB 77960327 B
B2
THE UNITED STATES OF AMERICA
100
ONE HUNDRED DOLLARS
FRANKLIN

# Selection


...
+ + − + − + + + + + − + + + +
0
|W|


FEDERAL RESERVE NOTE
CB 77960327 B
B2
THE UNITED STATES OF AMERICA
100
ONE HUNDRED DOLLARS

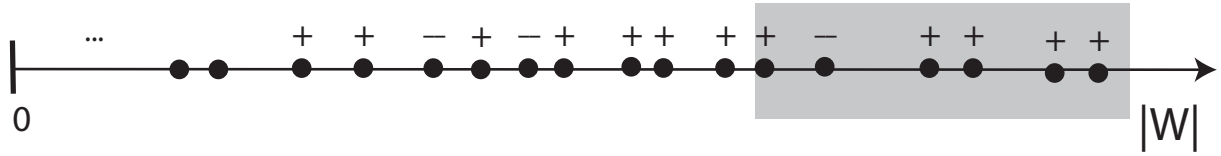
# Selection


...
+ + − + − + + + + + − + + + +
0
|W|


FEDERAL RESERVE NOTE
CB 77960327 B
B2
THE UNITED STATES OF AMERICA
100
ONE HUNDRED DOLLARS


MONOPOLY
100
100
100

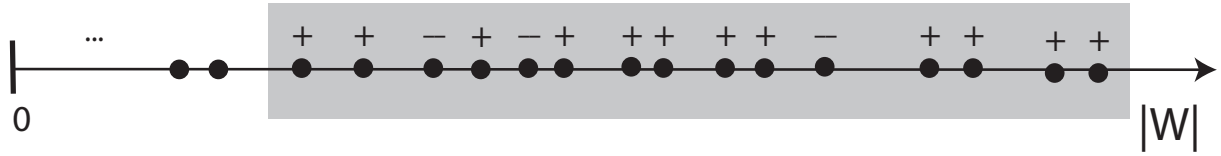
# Selection


...
+ + − + − + + + + + − + + + +
0
|W|


FEDERAL RESERVE NOTE
CB 77960327 B
B2
THE UNITED STATES OF AMERICA
100
ONE HUNDRED DOLLARS


MONOPOLY
100
100
100

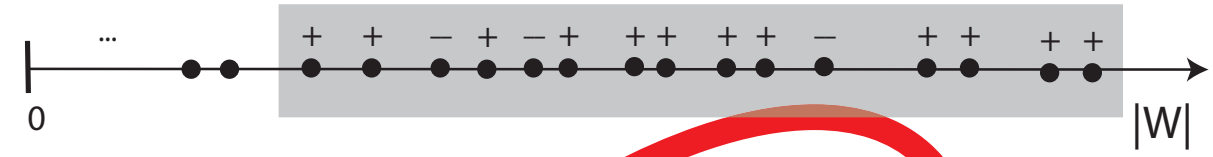
# Selection

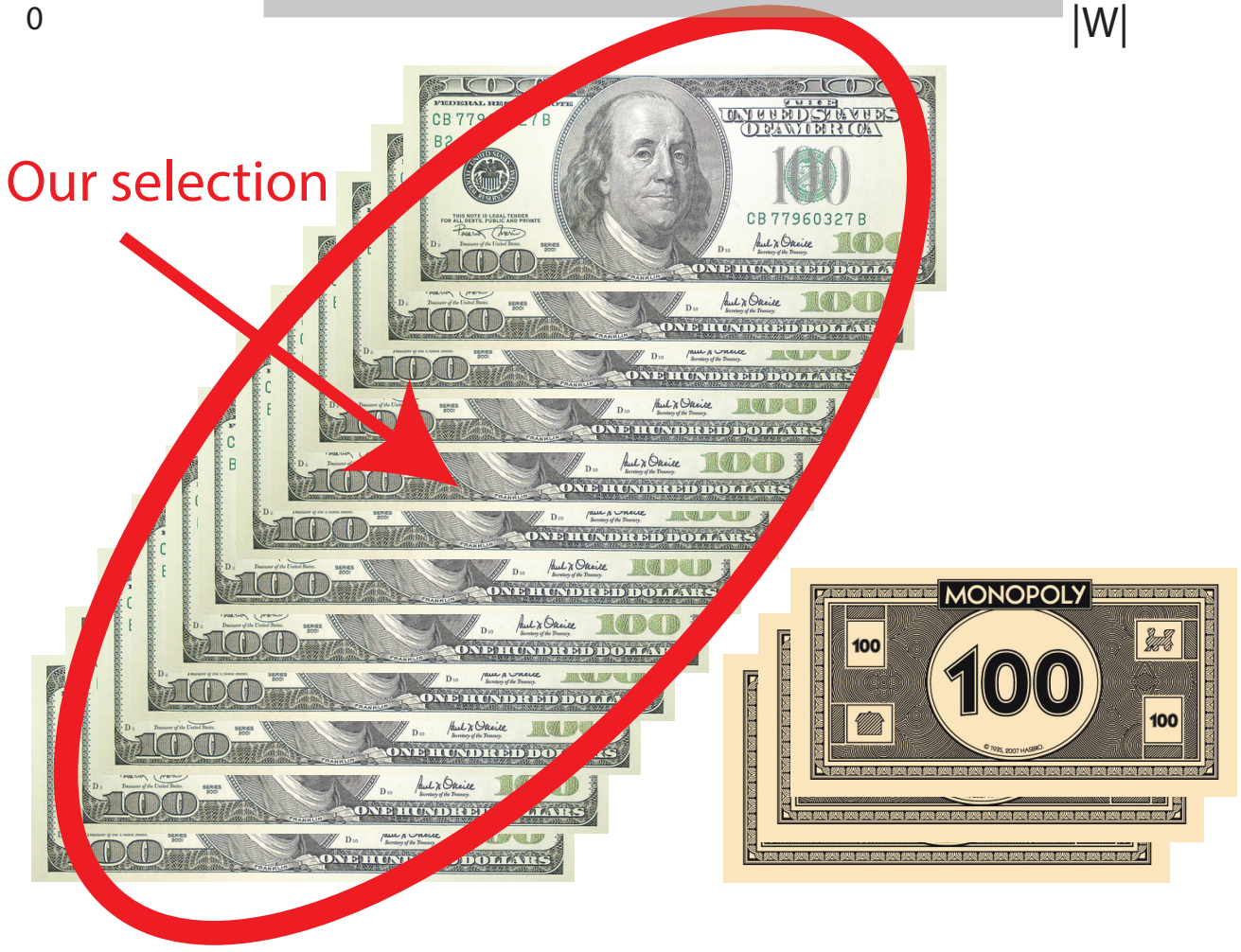




Stop last time ratio between '-' and '+' below target FDR level

# Selection



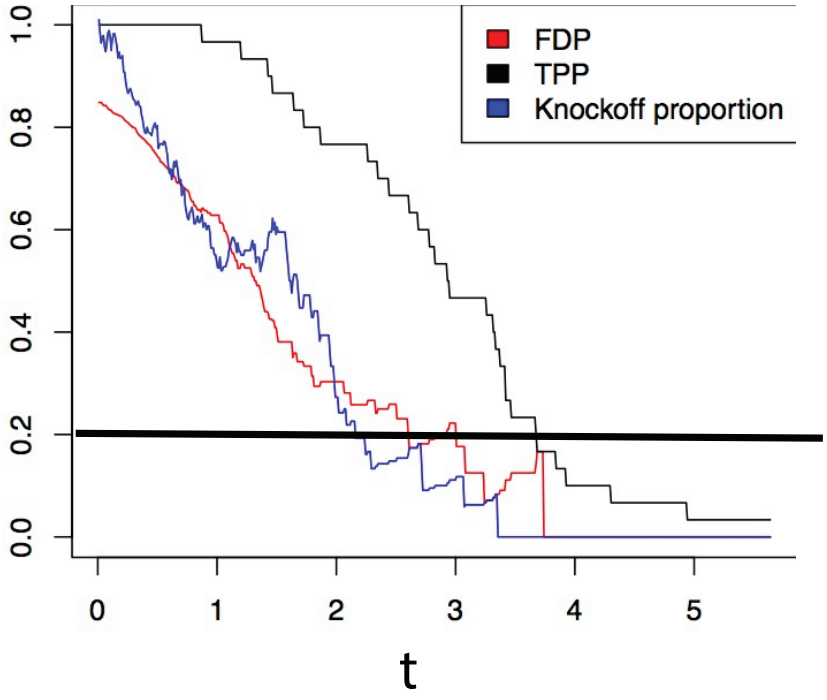

Select '+'s

# FDR control

$$N^{\pm}(t) = \#\{j : |W_j| \geq t \text{ and } \operatorname{sgn}(W_j) = \pm\}$$

$$T = \min\left\{t : \widehat{\mathsf{FDP}}(t) = \frac{1+N^{-}(t)}{1 \vee N^{+}(t)} \leq q\right\}$$

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

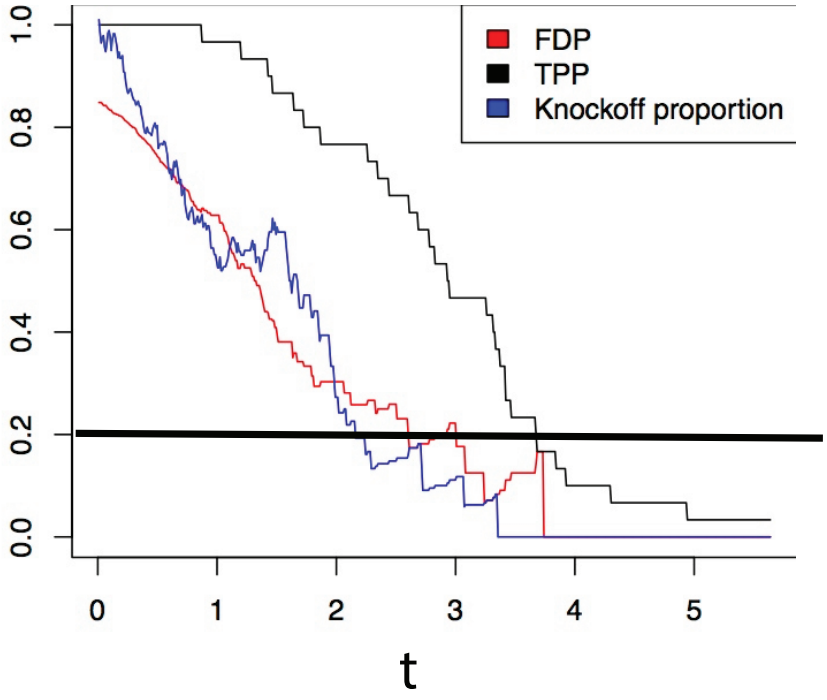


# FDR control

$$N^{\pm}(t) = \#\{j : |W_j| \geq t \text{ and } \operatorname{sgn}(W_j) = \pm\}$$

$$T = \min\left\{t : \widehat{\mathsf{FDP}}(t) = \frac{1+N^{-}(t)}{1 \vee N^{+}(t)} \leq q\right\}$$

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




## Theorem

- *Knockoff*

$$\mathbb{E}\left[\frac{V}{R + q^{-1}}\right] \leq q$$

$V$ : # false positives
$R$ : total # of selections

- *Knockoff+*

$$\mathbb{E}\left[\frac{V}{R \vee 1}\right] \leq q$$

# Not just dummy variables

For linear models, A.J. Miller ('84) creates "dummy" variables whose entries are drawn i.i.d. at random

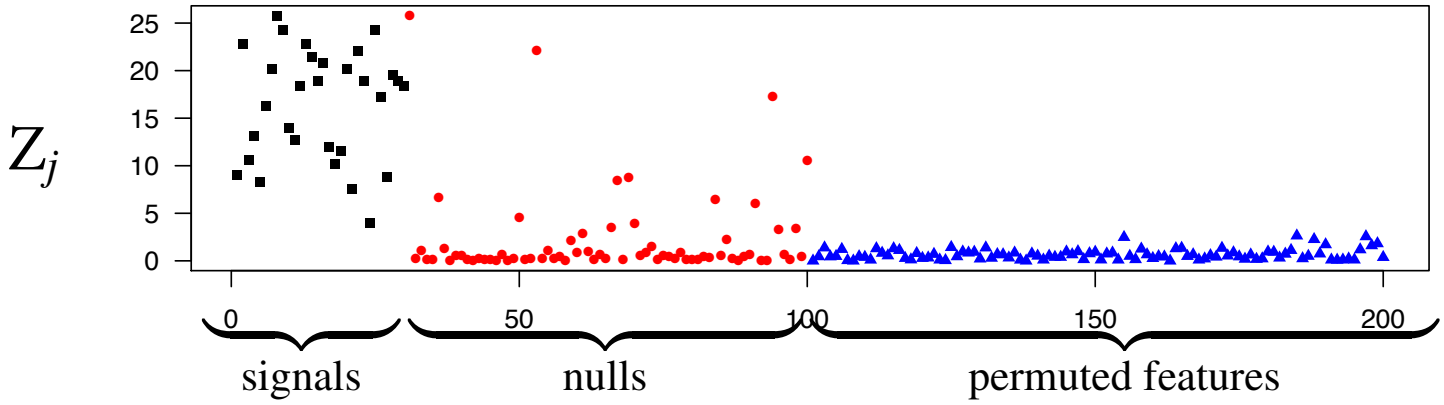
- Forward selection procedure is applied to augmented list of variables
- Stop when selects a dummy variable for the first time

# Not just dummy variables

For linear models, A.J. Miller ('84) creates "dummy" variables whose entries are drawn i.i.d. at random

- Forward selection procedure is applied to augmented list of variables
- Stop when selects a dummy variable for the first time

Importance statistics for originals and dummies




Dummies cannot serve as controls

# Towards pairwise exchangeability

To serve as controls, pairwise correlations need to be preserved

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(\tilde{X}_i, X_j) \quad i \neq j$$

# Towards pairwise exchangeability

To serve as controls, pairwise correlations need to be preserved

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(\tilde{X}_i, X_j) \quad i \neq j$$

For non-linear models, want something like

$$X_i \,|\, X_{-i} \quad \stackrel{d}{=} \quad \tilde{X}_i \,|\, X_{-i}$$

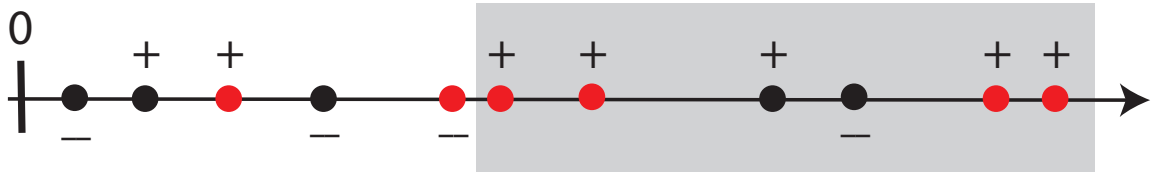
Implied by

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

*Martingales: for Mathematicians... (Hopefully)*

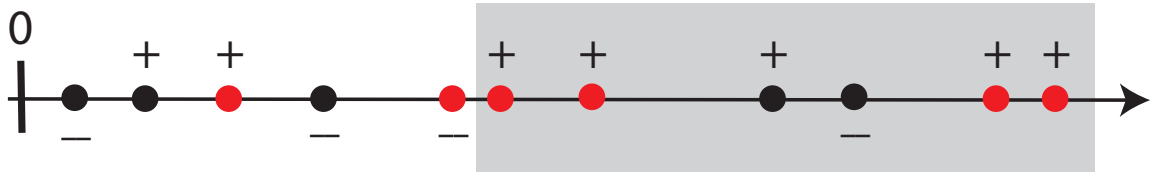
# Why does all this work?

$$T = \min\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\} \qquad \begin{aligned}\mathcal{S}^+(t) &= \{j : W_j \geq t\} \\ \mathcal{S}^-(t) &= \{j : W_j \leq -t\}\end{aligned}$$



# Why does all this work?

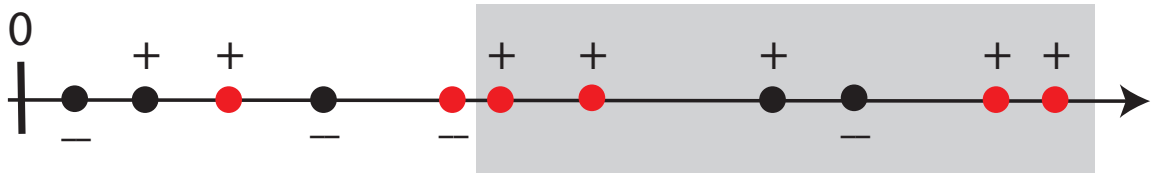
$$T = \min\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\} \qquad \begin{aligned}\mathcal{S}^+(t) &= \{j : W_j \geq t\} \\ \mathcal{S}^-(t) &= \{j : W_j \leq -t\}\end{aligned}$$




$$\mathsf{FDP}(t) = \frac{\#\{j \text{ null} : j \in \mathcal{S}^+(t)\}}{\#\{j : j \in \mathcal{S}^+(t)\} \vee 1}$$

# Why does all this work?

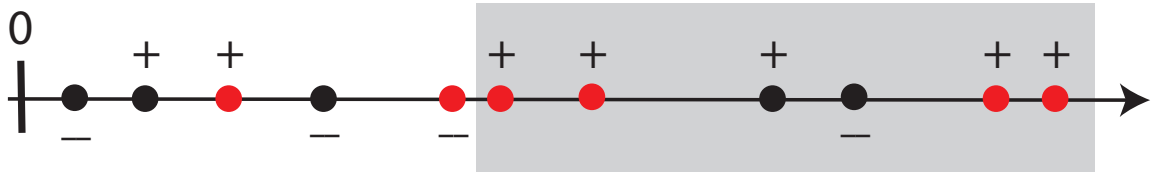
$$T = \min\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\} \qquad \begin{aligned}\mathcal{S}^+(t) &= \{j : W_j \geq t\} \\ \mathcal{S}^-(t) &= \{j : W_j \leq -t\}\end{aligned}$$




$$\mathsf{FDP}(t) = \frac{\#\{j \text{ null} : j \in \mathcal{S}^+(t))\}}{\#\{j : j \in \mathcal{S}^+(t)\} \vee 1} \cdot \frac{1 + \#\{j \text{ null} : j \in \mathcal{S}^-(t)\}}{1 + \#\{j \text{ null} : j \in \mathcal{S}^-(t)\}}$$

# Why does all this work?

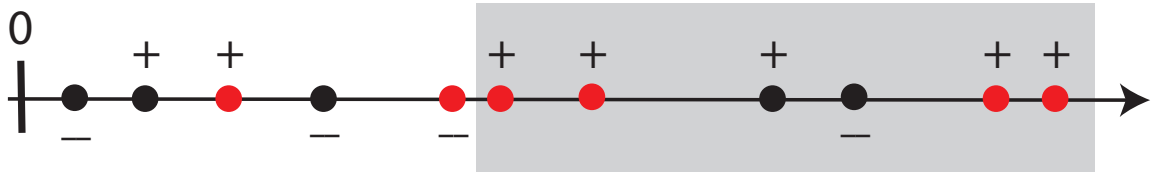
$$T = \min\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\} \qquad \begin{aligned}\mathcal{S}^+(t) &= \{j : W_j \geq t\} \\ \mathcal{S}^-(t) &= \{j : W_j \leq -t\}\end{aligned}$$




$$\mathsf{FDP}(t) \leq q \cdot \frac{\overbrace{\#\{j \text{ null} : j \in \mathcal{S}^+(t)\}}^{V^+(t)}}{1 + \underbrace{\#\{j \text{ null} : j \in \mathcal{S}^-(t)\}}_{V^-(t)}}$$

# Why does all this work?

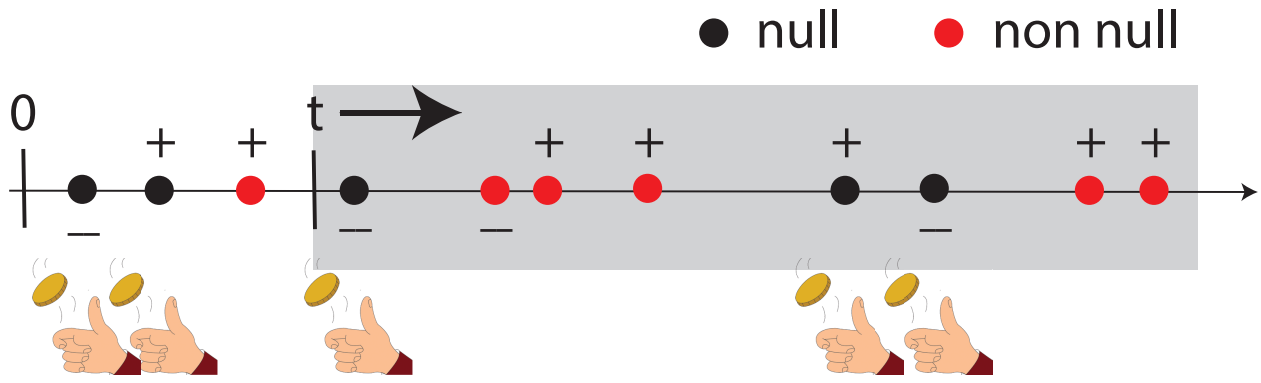
$$T = \min\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\} \qquad \begin{aligned}\mathcal{S}^+(t) &= \{j : W_j \geq t\} \\ \mathcal{S}^-(t) &= \{j : W_j \leq -t\}\end{aligned}$$




$$\mathrm{FDP}(t) \leq q \cdot \frac{\overbrace{\#\{j \text{ null} : j \in \mathcal{S}^+(t)\}}^{V^+(t)}}{1 + \underbrace{\#\{j \text{ null} : j \in \mathcal{S}^-(t)\}}_{V^-(t)}}$$

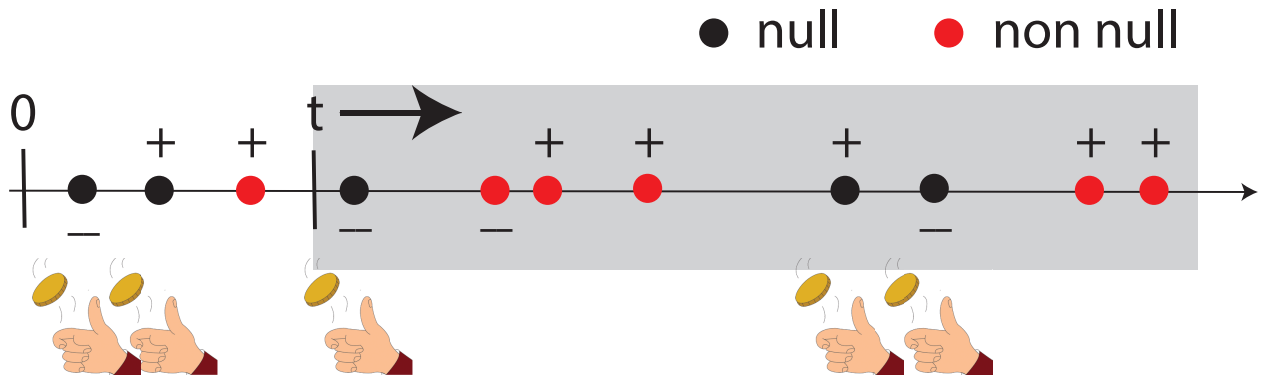
- $V^+(t)/(1+V^-(t))$ is a super-martingale w.r.t. well defined filtration
- $T$ is stopping time

# Optional stopping time theorem



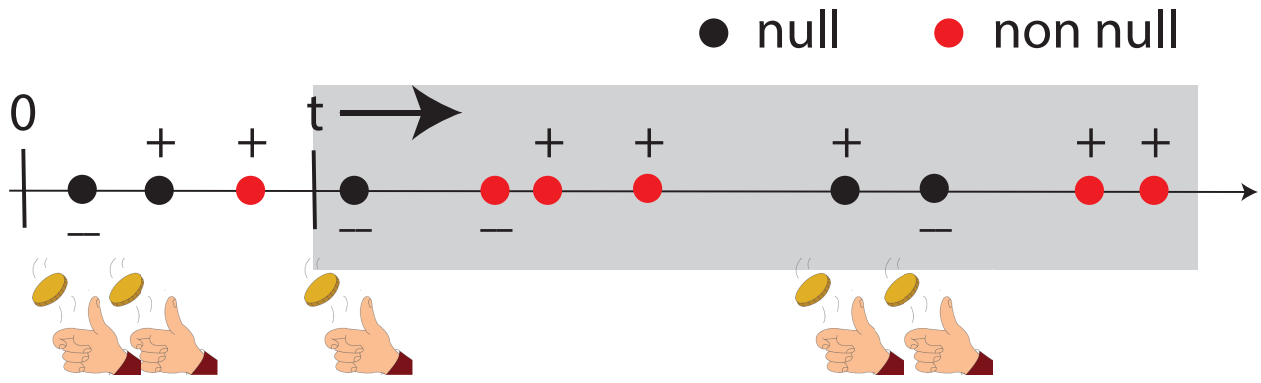

$$\text{FDR} \leq q\, \mathbb{E}\left[\frac{V^+(T)}{1+V^-(T)}\right]$$

# Optional stopping time theorem



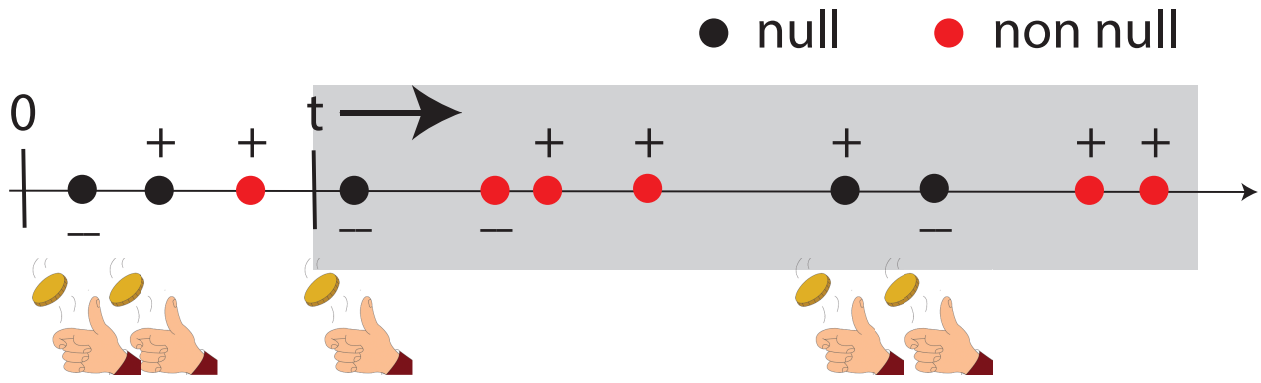

$$\text{FDR} \leq q\ \mathbb{E}\left[\frac{V^+(T)}{1+V^-(T)}\right] \leq q\ \mathbb{E}\left[\frac{V^+(0)}{1+V^-(0)}\right]$$

# Optional stopping time theorem




$$\text{FDR} \leq q\ \mathbb{E}\left[\frac{V^+(T)}{1+V^-(T)}\right] \leq q\ \mathbb{E}\left[\frac{V^+(0)}{1+V^-(0)}\right] = q\ \mathbb{E}\left[\frac{\overbrace{V^+(0)}^{\text{Ber}(\#\text{nulls},1/2)}}{1+\#\text{nulls}-V^+(0)}\right]$$

# Optional stopping time theorem




$$\text{FDR} \leq q\ \mathbb{E}\left[\frac{V^+(T)}{1+V^-(T)}\right] \leq q\ \mathbb{E}\left[\frac{V^+(0)}{1+V^-(0)}\right] = q\ \mathbb{E}\left[\frac{\overbrace{V^+(0)}^{\text{Ber}(\#\text{nulls},1/2)}}{1+\#\text{nulls}-V^+(0)}\right] \leq q$$

# *Some Remarks*

# Panning for gold

- Thousands/millions of variables
- No idea how response depends on all of these

Model-free knockoff filters controls FDR

$$\text{FDR} = \mathbb{E}\,\frac{\#\text{ variables falsely selected}}{\text{total }\#\text{ of selections}}$$

# Panning for gold

- Thousands/millions of variables
- No idea how response depends on all of these

**Model-free knockoff filters controls FDR**

$$\text{FDR} = \mathbb{E}\,\frac{\#\text{ variables falsely selected}}{\text{total }\#\text{ of selections}}$$

Pros:

- No parameters
- No p-values
- Holds for finite samples
- No matter the dependence between $Y$ and $X$
- No matter the dimensionality

# Panning for gold

- Thousands/millions of variables
- No idea how response depends on all of these

Model-free knockoff filters controls FDR

$$\text{FDR} = \mathbb{E}\,\frac{\#\text{ variables falsely selected}}{\text{total }\#\text{ of selections}}$$

Pros:

- No parameters
- No p-values
- Holds for finite samples
- No matter the dependence between $Y$ and $X$
- No matter the dimensionality

Cons: Need to know something about the distribution of covariates

# Obstacles to obtaining p-values

$$Y \mid X \sim \text{Bernoulli}(\text{logit}(X^\top \beta))$$

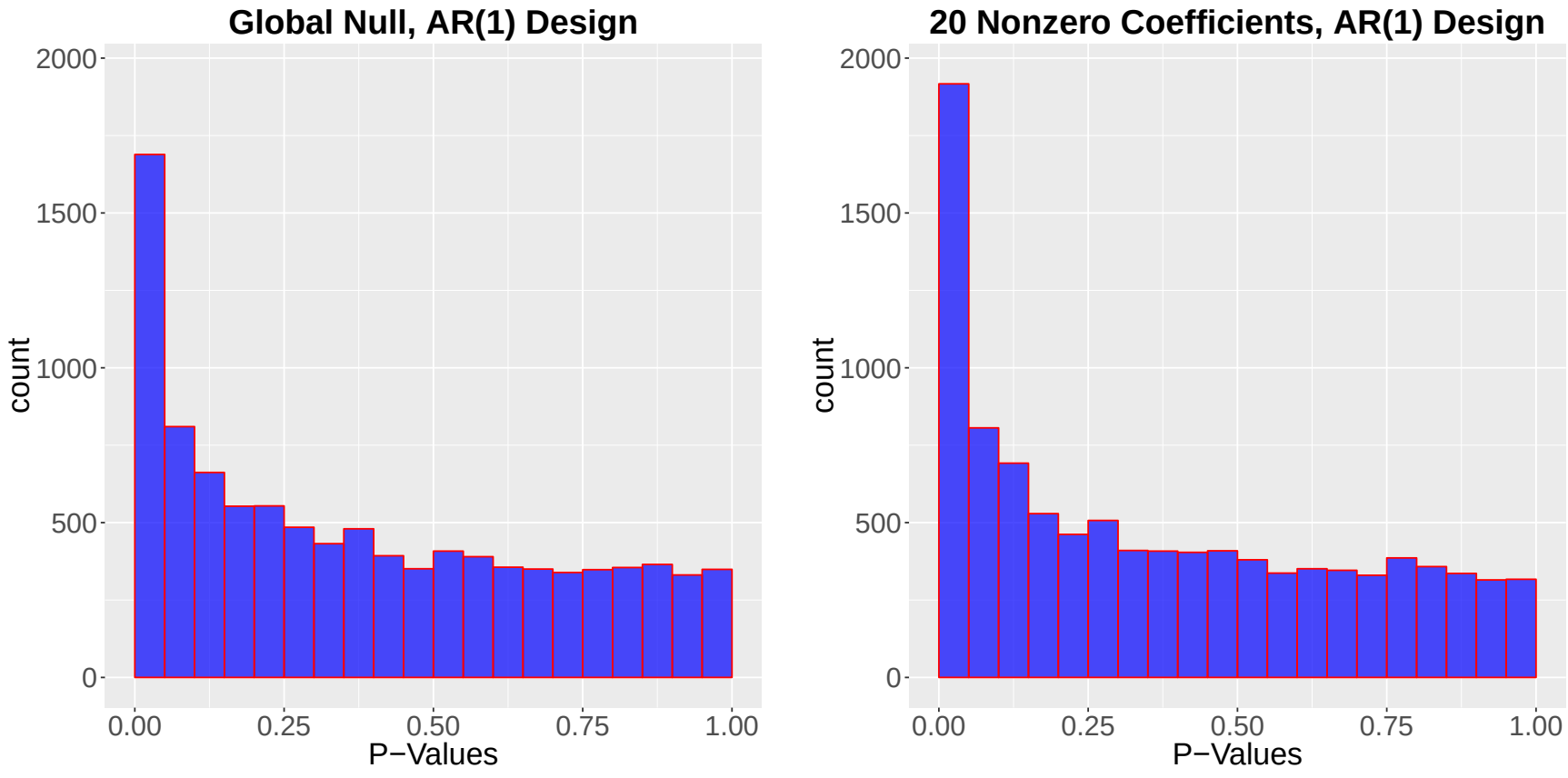



Figure: Distribution of null logistic regression p-values with $n = 500$ and $p = 200$

# Obstacles to obtaining p-values

| $\mathbb{P}\{p\text{-val} \leq \ldots\%\}$ | Sett. (1) | Sett. (2) | Sett. (3) | Sett. (4) |
|---|---|---|---|---|
| 5% | 16.89% (0.37) | 19.17% (0.39) | 16.88% (0.37) | 16.78% (0.37) |
| 1% | 6.78% (0.25) | 8.49% (0.28) | 7.02% (0.26) | 7.03% (0.26) |
| 0.1% | 1.53% (0.12) | 2.27% (0.15) | 1.87% (0.14) | 2.04% (0.14) |

Table: Inflated p-value probabilities with estimated Monte Carlo SEs

# Relationship with classical setup

| Classical | MF Knockoffs |
| --- | --- |

# Relationship with classical setup

| Classical | MF Knockoffs |
| --- | --- |
| Observations of $X$ are fixed<br>Inference is conditional on obs. values | Observations of $X$ are random[1] |

1 Often appropriate in 'big' data apps: e.g. SNPs of subjects randomly sampled

# Relationship with classical setup

| Classical | MF Knockoffs |
|---|---|
| Observations of $X$ are fixed<br>Inference is conditional on obs. values | Observations of $X$ are random[1] |
| Strong model linking $Y$ and $X$ | Model free[2] |

1. Often appropriate in 'big' data apps: e.g. SNPs of subjects randomly sampled
2. Shifts the 'burden' of knowledge

# Relationship with classical setup

| Classical | MF Knockoffs |
|---|---|
| Observations of $X$ are fixed<br>Inference is conditional on obs. values | Observations of $X$ are random[1] |
| Strong model linking $Y$ and $X$ | Model free[2] |
| Useful inference even if model inexact | Useful inference even if model inexact[3] |

1. Often appropriate in 'big' data apps: e.g. SNPs of subjects randomly sampled
2. Shifts the 'burden' of knowledge
3. More later

# Shift in the burden of knowledge

When are our assumptions useful?

# Shift in the burden of knowledge

When are our assumptions useful?

- When we have large amounts of unsupervised data (e.g. economic studies with same covariate info but different responses)

# Shift in the burden of knowledge

When are our assumptions useful?

- When we have large amounts of unsupervised data (e.g. economic studies with same covariate info but different responses)
- When we have more prior information about the covariates than about their relationship with a response (e.g. GWAS)

# Shift in the burden of knowledge

When are our assumptions useful?

- When we have large amounts of unsupervised data (e.g. economic studies with same covariate info but different responses)
- When we have more prior information about the covariates than about their relationship with a response (e.g. GWAS)
- When we control the distribution of $X$ (experimental crosses in genetics, gene knockout experiments,...)

# Shift in the burden of knowledge

When are our assumptions useful?

- When we have large amounts of unsupervised data (e.g. economic studies with same covariate info but different responses)
- When we have more prior information about the covariates than about their relationship with a response (e.g. GWAS)
- When we control the distribution of $X$ (experimental crosses in genetics, gene knockout experiments,...)
- ...

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \Longrightarrow \quad \tilde{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \tilde{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **)$$

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \tilde{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \tilde{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

$\boldsymbol{s}$ such that $** \succeq 0$

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \tilde{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

  $\boldsymbol{s}$ such that $** \succeq 0$

- Given $X$, sample $\tilde{X}$ from $\tilde{X} \,|\, X$ (regression formula)

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \tilde{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

  $\boldsymbol{s}$ such that $** \succeq 0$

- Given $X$, sample $\tilde{X}$ from $\tilde{X} \,|\, X$ (regression formula)

Different from knockoffs of Foygel Barber and C. (2015)

# Gaussian knockoffs

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

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \implies \tilde{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X,\, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

  $\boldsymbol{s}$ such that $** \succeq 0$

- Given $X$, sample $\tilde{X}$ from $\tilde{X} \,|\, X$ (regression formula)

Different from knockoffs of Foygel Barber and C. (2015)

Must always have

$$\mathbb{E}\, X = \mathbb{E}\, \tilde{X} \qquad \mathrm{Cov}(X, \tilde{X}) = **$$

Knockoffs obeying this only $\longrightarrow$ *second-order knockoffs*

# *Some Examples*

# Logistic model with dependent covariates

LCD (Lasso coeff. difference)

$$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\mathsf{CV}})|$$

$$W_j = Z_j - \tilde{Z}_j$$

# Logistic model with dependent covariates

LCD (Lasso coeff. difference)

$$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\mathsf{CV}})|$$

$$W_j = Z_j - \tilde{Z}_j$$

LSM (Lasso sign max)

$$Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$$

$$W_j = (Z_j \vee \tilde{Z}_j)\, \mathrm{sgn}(Z_j - \tilde{Z}_j)$$

# Logistic model with dependent covariates

LCD (Lasso coeff. difference)

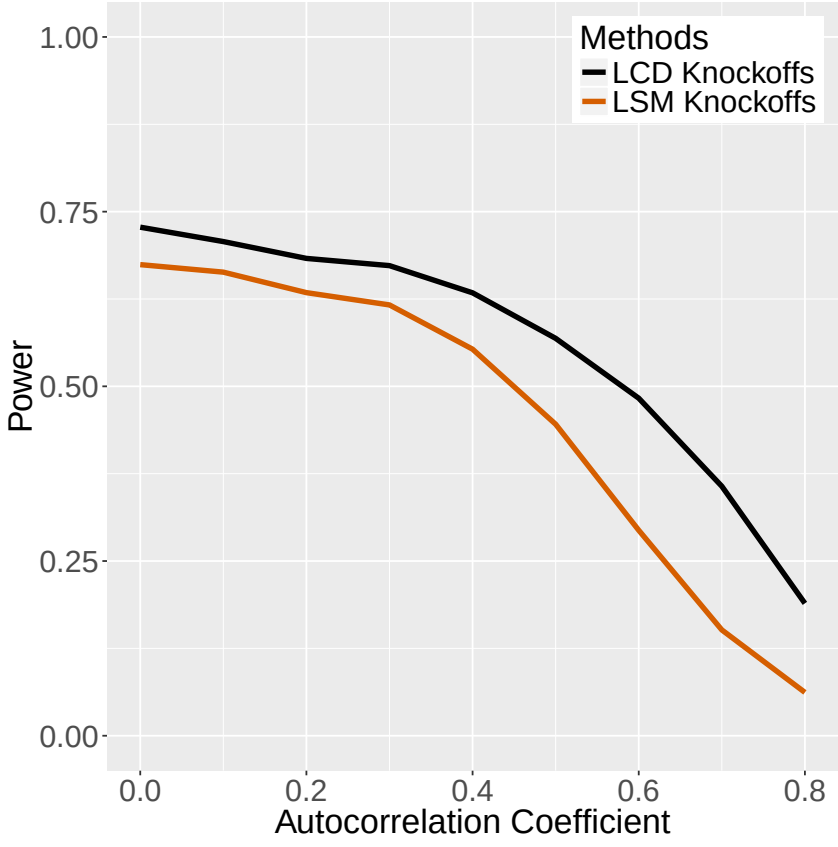
$$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\mathrm{CV}})|$$

$$W_j = Z_j - \tilde{Z}_j$$

LSM (Lasso sign max)

$$Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$$

$$W_j = (Z_j \vee \tilde{Z}_j)\ \mathrm{sgn}(Z_j - \tilde{Z}_j)$$



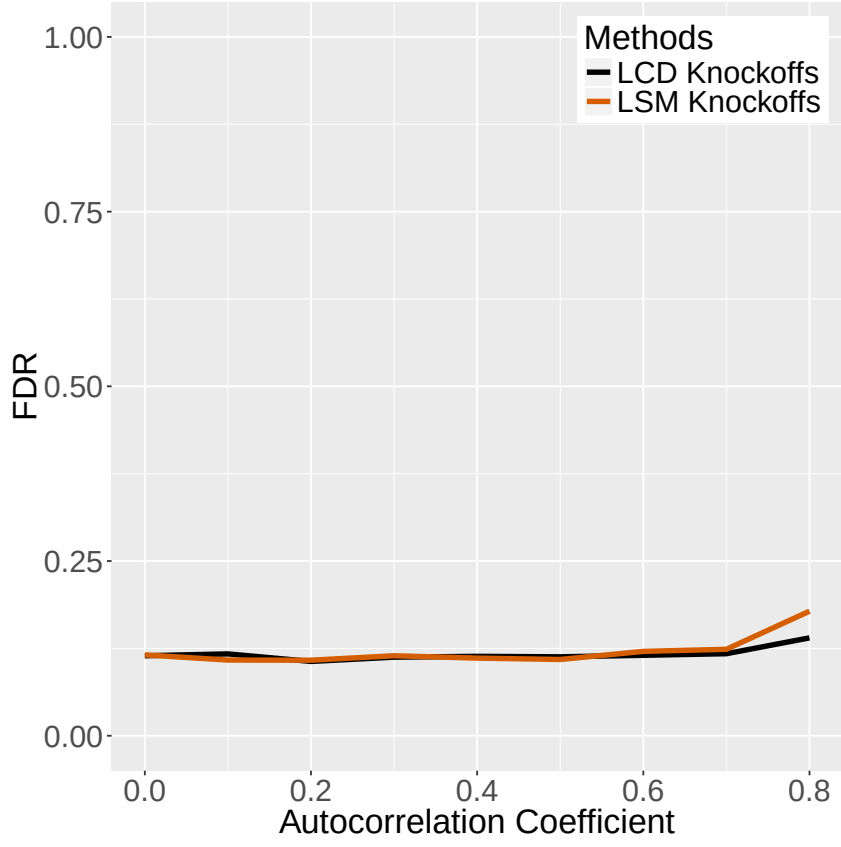




Figure: $n = 3000$, $p = 1000$, 60 regression coeffs. of magnitude 3.5 and random signs

# Bayesian knockoff statistics

LCD (Lasso coeff. difference)

BVS (Bayesian variable selection)

$$Z_j = \mathbb{P}(\beta_j \neq 0 \,|\, \boldsymbol{y}, \boldsymbol{X})$$

$$W_j = Z_j - \tilde{Z}_j$$

# Bayesian knockoff statistics

LCD (Lasso coeff. difference)

BVS (Bayesian variable selection)

$$Z_j = \mathbb{P}(\beta_j \neq 0 \,|\, \boldsymbol{y}, \boldsymbol{X})$$

$$W_j = Z_j - \tilde{Z}_j$$

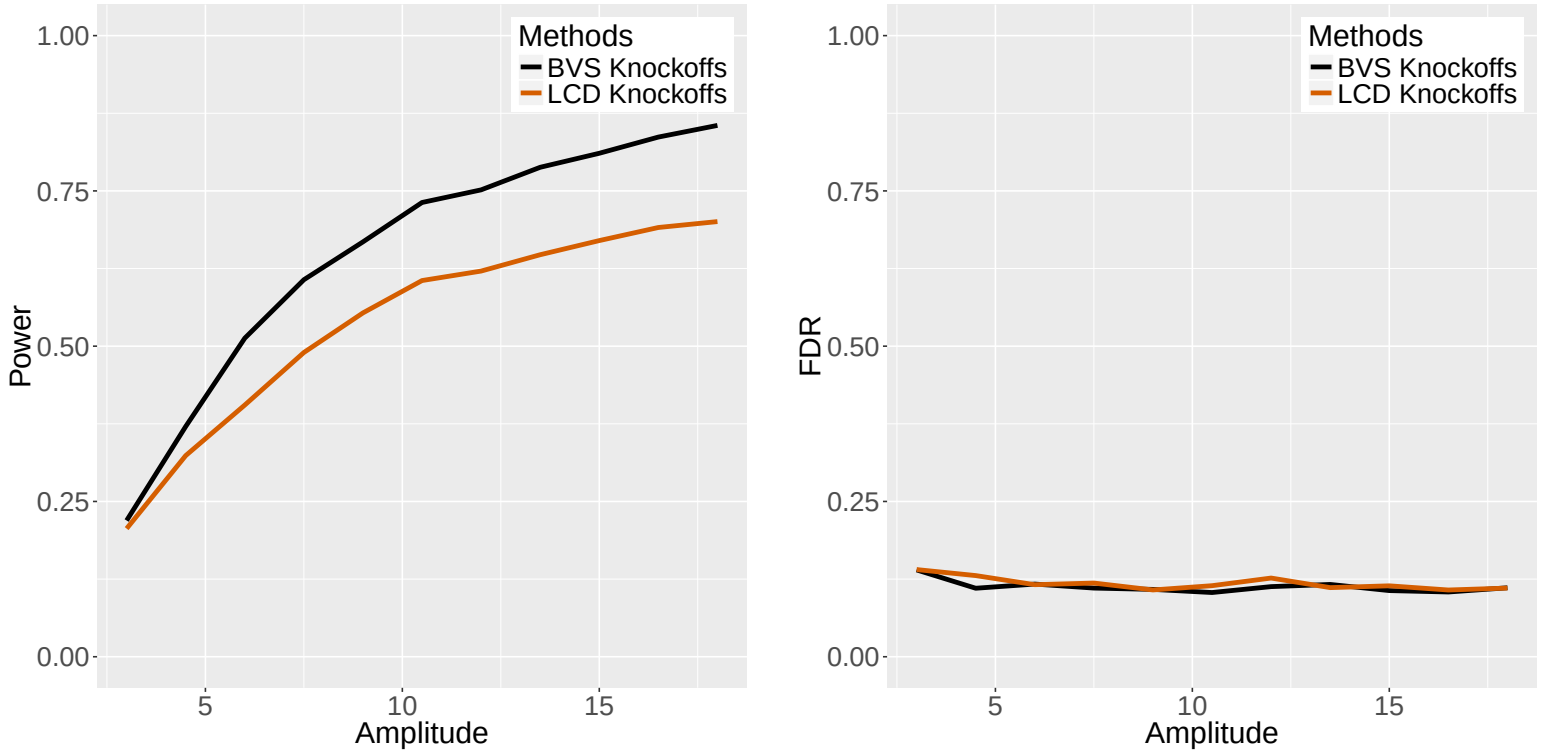



Figure: $n = 300$, $p = 1000$ and Bayesian linear model with 60 expected variables

Inference is correct even if prior is wrong or MCMC has not converged

# Flexibility and adaptivity

Test statistic $W_j(Z_j, \tilde{Z}_j)$ can be anything...

# Flexibility and adaptivity

Test statistic $W_j(Z_j, \tilde{Z}_j)$ can be anything...

e.g. $Z_j$ $=$ absolute value of LASSO coeff.
or
some random forest feature importance
depending on which model has lower CV error

# Power comparisons: linear model with indep. covariates

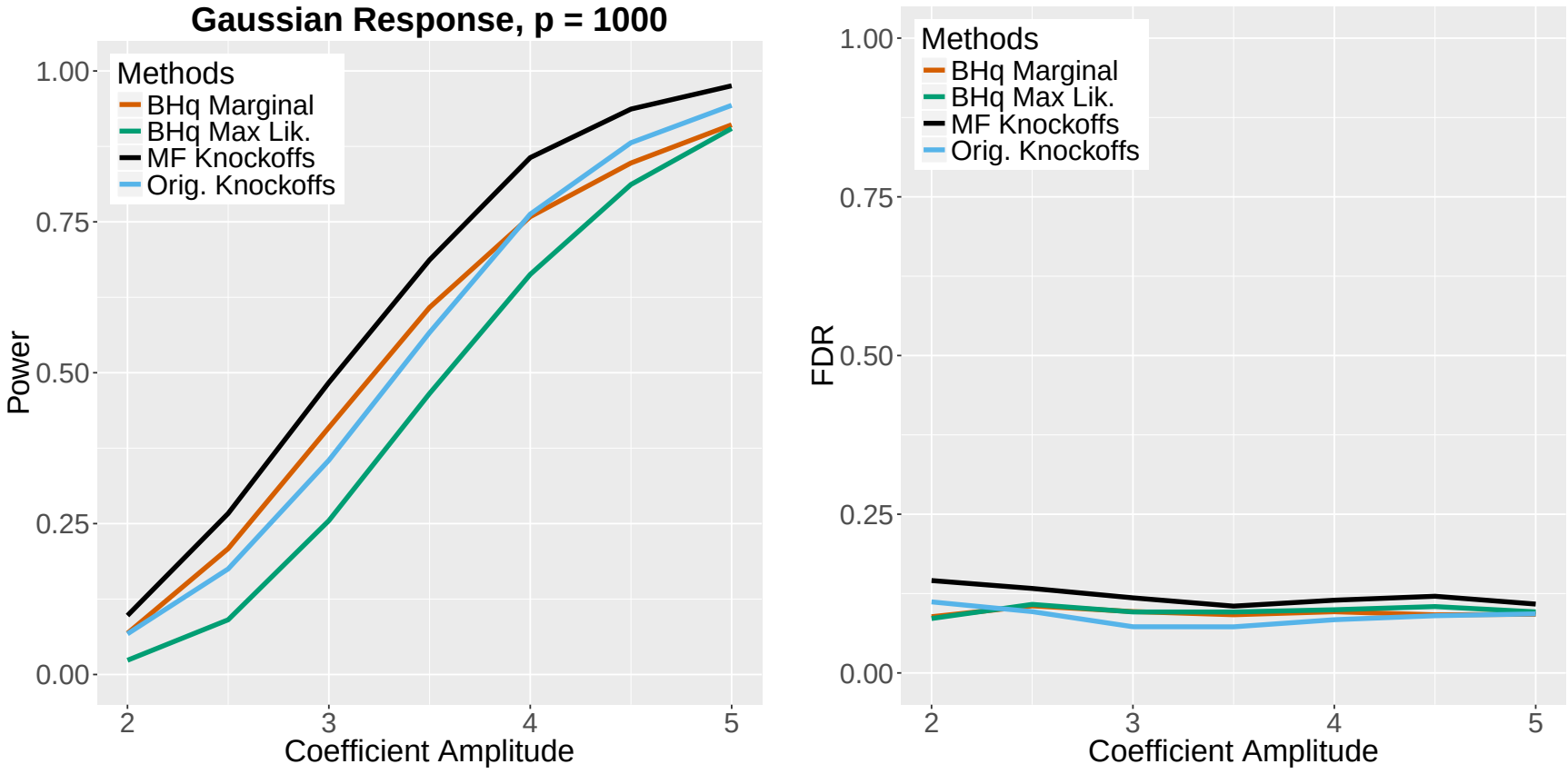



Figure: Low-dimensional setting: $n = 3000$, $p = 1000$

# Power comparisons: linear model with indep. covariates

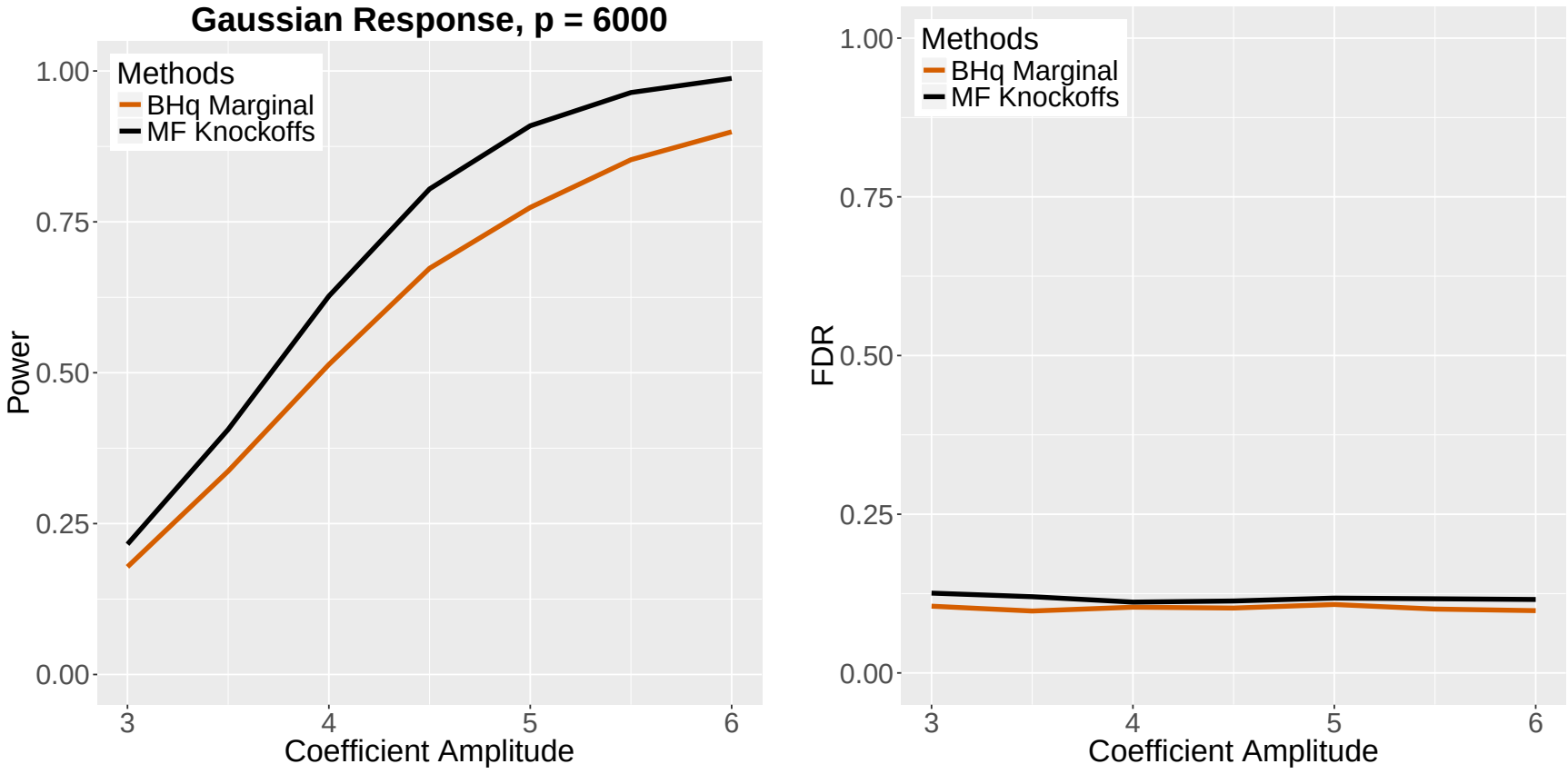



Figure: High-dimensional setting: $n = 3000$, $p = 6000$

# Power comparisons: logistic model with indep. covariates

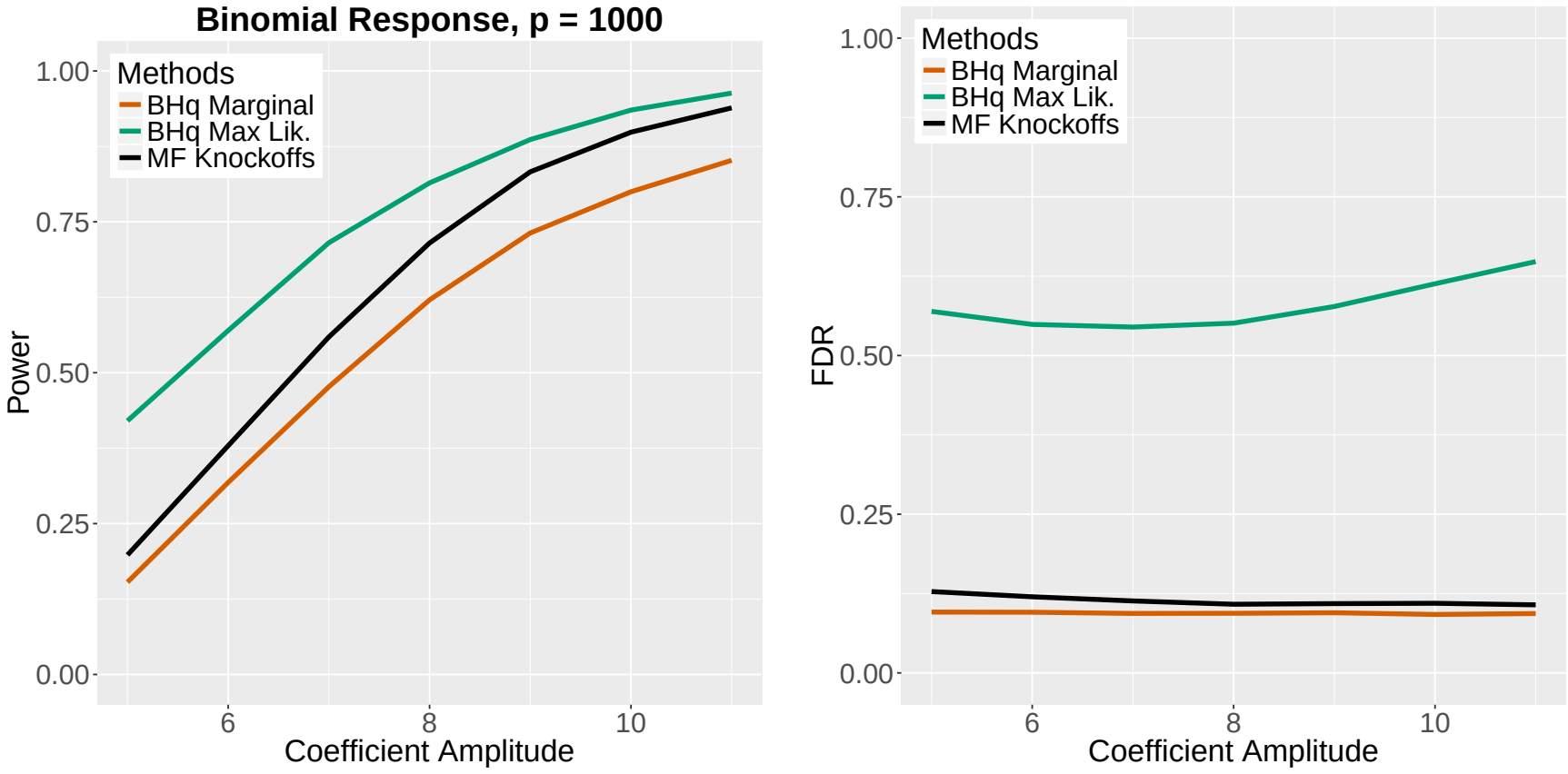



Figure: Low-dimensional setting: $n = 3000$, $p = 1000$

# Power comparisons: logistic model with indep. covariates

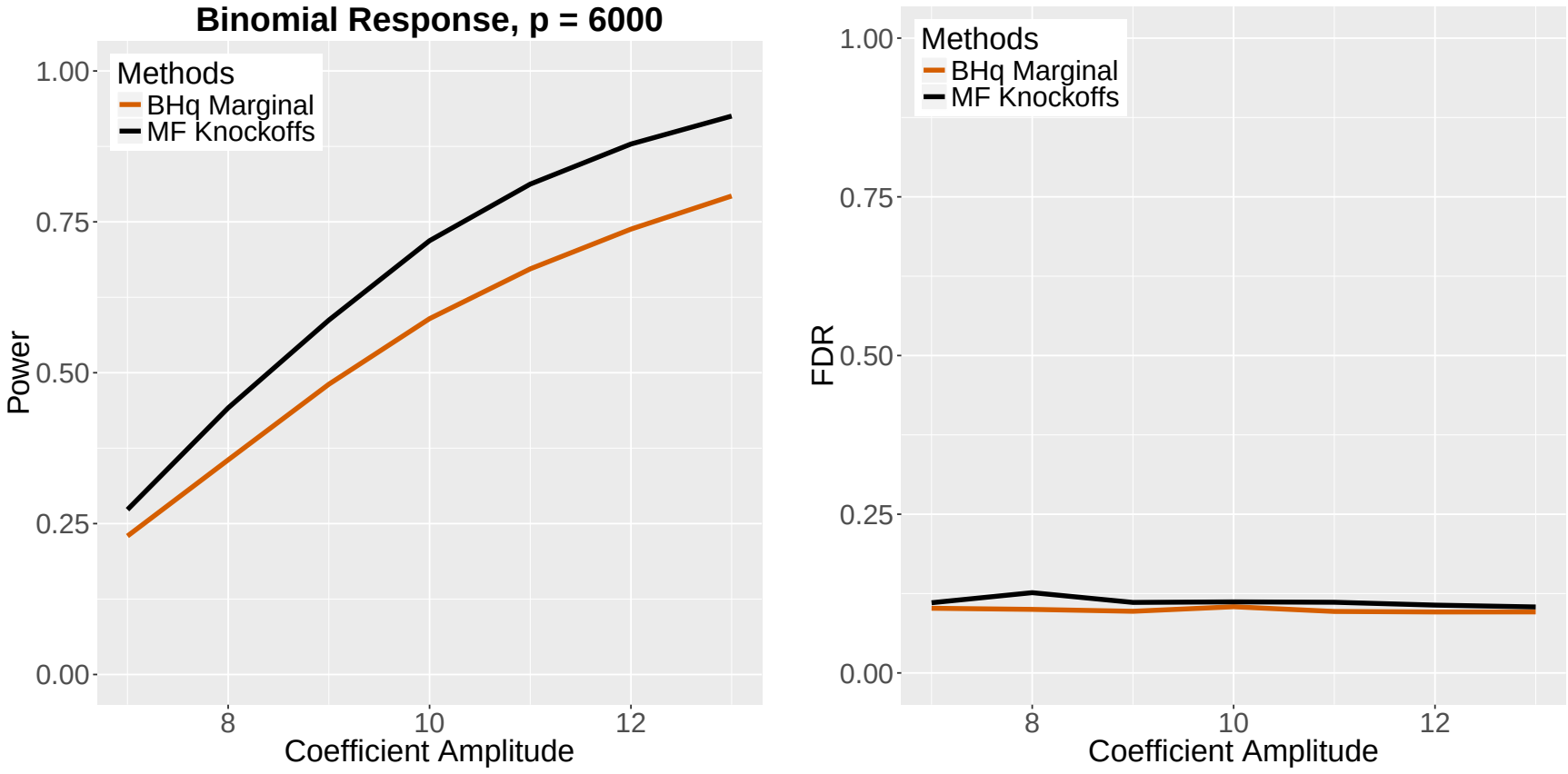



Figure: High-dimensional setting: $n = 3000$, $p = 6000$

# Linear model with dependent covariates

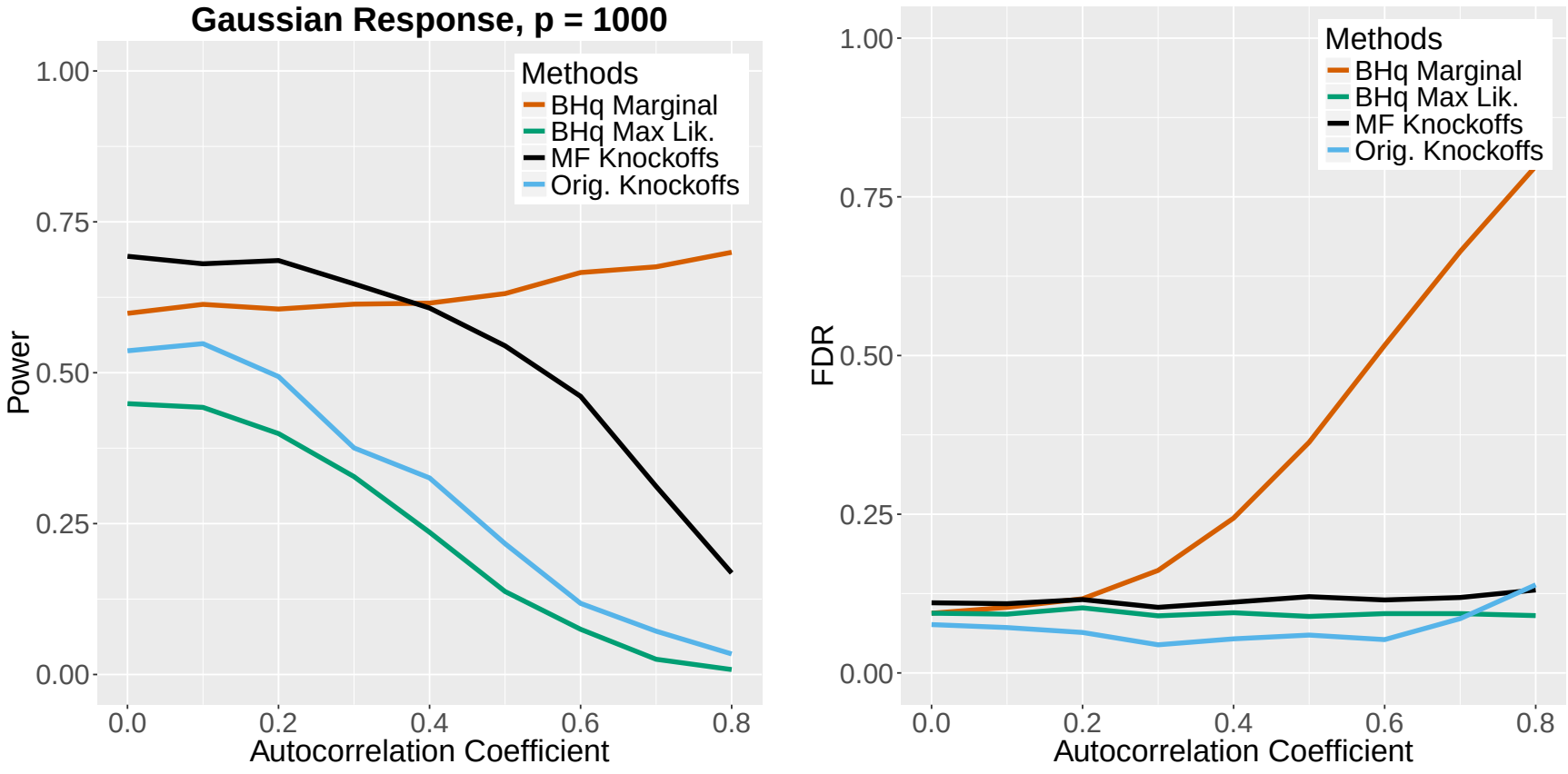



Figure: Low-dimensional setting: $n = 3000$, $p = 1000$

# Logistic model with dependent covariates

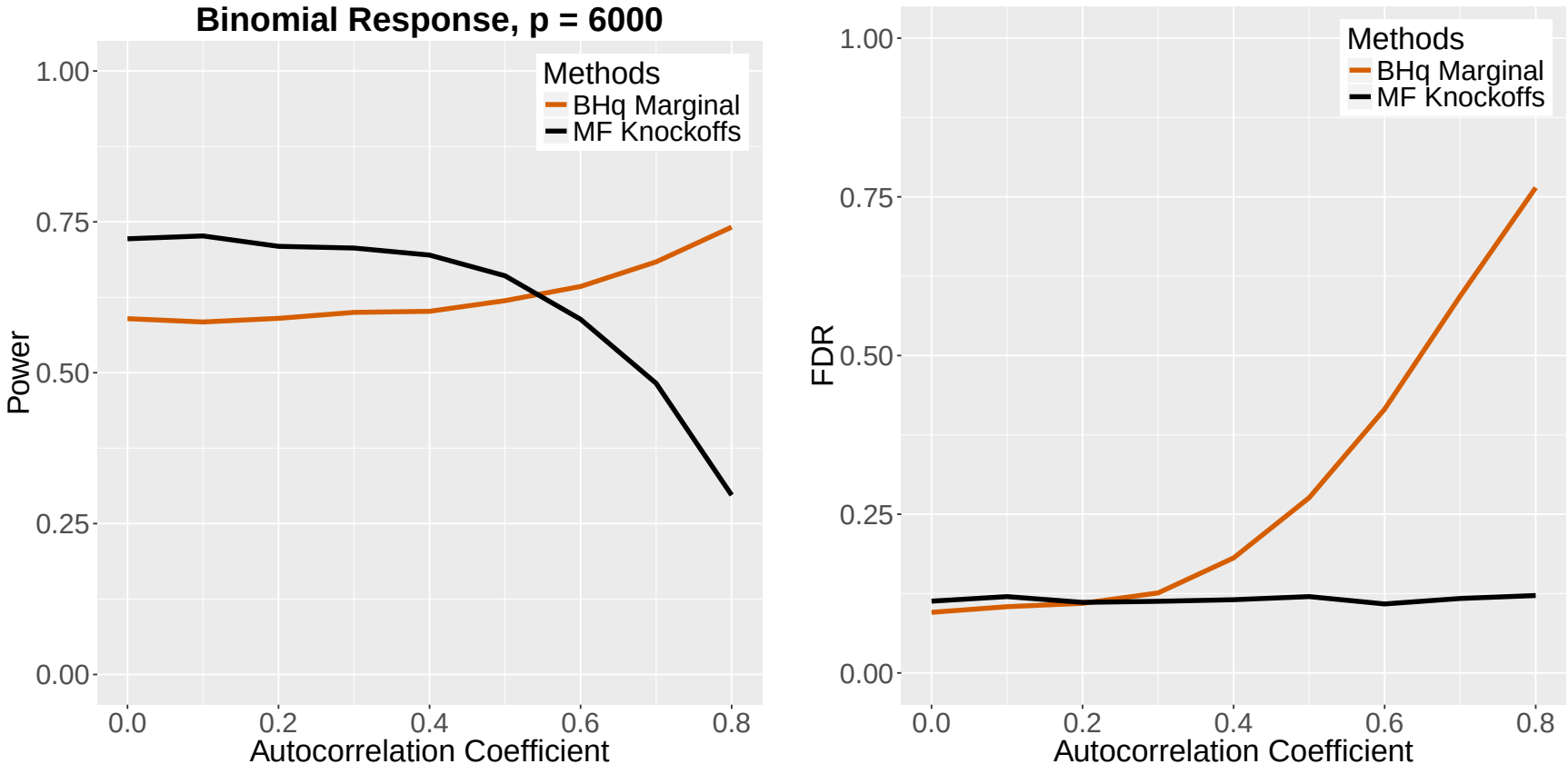



Figure: High-dimensional setting: $n = 3000$, $p = 6000$

# Robustness

$X$ is AR(1) and autocorrelation $\rho = 0.3$

$$\mathbf{\Sigma}_{\mathsf{EC}} = (1 - \alpha)\mathbf{\Sigma} + \alpha\hat{\mathbf{\Sigma}}$$

# Robustness

$X$ is AR(1) and autocorrelation $\rho = 0.3$

$$\mathbf{\Sigma}_{\text{EC}} = (1-\alpha)\mathbf{\Sigma} + \alpha\hat{\mathbf{\Sigma}}$$

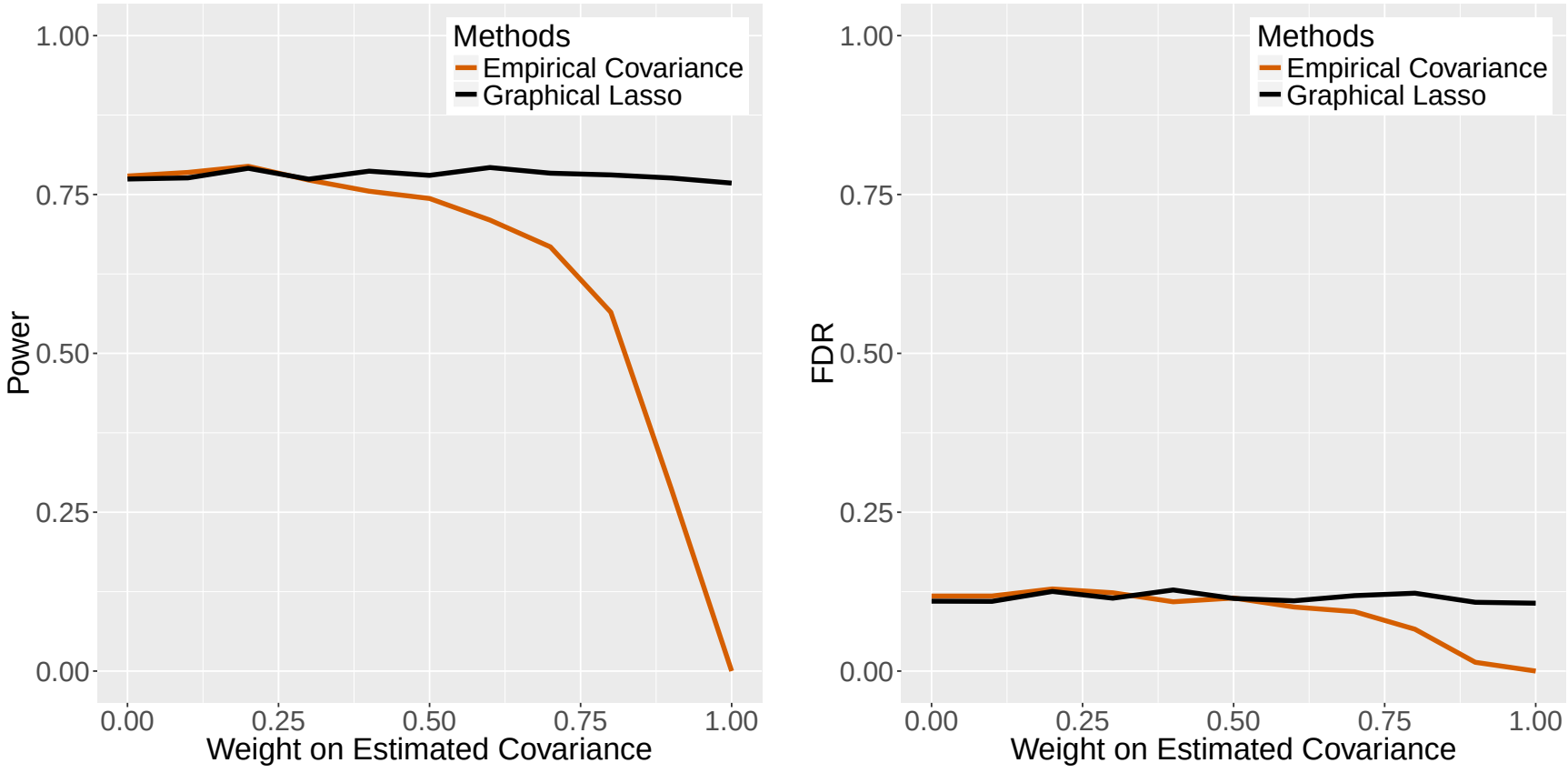



Figure: High-dimensional logistic model with $n = 800$, $p = 1500$

*Real Data Analysis*

# Genetic analysis of Crohn's disease (CD)

Data provided by the Wellcome Trust Case Control Consortium (WTCCC)

- $n \approx 5,000$ subjects
  - $\approx 2,000$ CD patients
  - $\approx 3,000$ healthy controls
- $p \approx 400,000$ SNPs
- Previously analyzed in WTCCC (2007)

# Peek at the results

*Model-free knockoffs* with nominal FDR level of 10%

# Peek at the results

*Model-free knockoffs* with nominal FDR level of 10%

- Power is much higher: WTCCC ('07) made 9 discoveries while knockoffs made 18 on average

# Peek at the results

*Model-free knockoffs* with nominal FDR level of 10%

- Power is much higher: WTCCC ('07) made 9 discoveries while knockoffs made 18 on average
- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10) and were not discovered in WTCCC ('07)

# Peek at the results

*Model-free knockoffs* with nominal FDR level of 10%

- Power is much higher: WTCCC ('07) made 9 discoveries while knockoffs made 18 on average
- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10) and were not discovered in WTCCC ('07)
- Knockoffs made a number of discoveries not found in either WTCCC ('07) or Franke et al. ('10)

# Peek at the results

*Model-free knockoffs* with nominal FDR level of 10%

- Power is much higher: WTCCC ('07) made 9 discoveries while knockoffs made 18 on average
- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10) and were not discovered in WTCCC ('07)
- Knockoffs made a number of discoveries not found in either WTCCC ('07) or Franke et al. ('10)
  - Expect some (roughly 10%) of these to be false discoveries

# Peek at the results

*Model-free knockoffs* with nominal FDR level of 10%

- Power is much higher: WTCCC ('07) made 9 discoveries while knockoffs made 18 on average
- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10) and were not discovered in WTCCC ('07)
- Knockoffs made a number of discoveries not found in either WTCCC ('07) or Franke et al. ('10)
  - Expect some (roughly 10%) of these to be false discoveries
  - It is likely that many of these correspond to true discoveries

# Peek at the results

*Model-free knockoffs* with nominal FDR level of 10%

- Power is much higher: WTCCC ('07) made 9 discoveries while knockoffs made 18 on average
- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10) and were not discovered in WTCCC ('07)
- Knockoffs made a number of discoveries not found in either WTCCC ('07) or Franke et al. ('10)
  - Expect some (roughly 10%) of these to be false discoveries
  - It is likely that many of these correspond to true discoveries
  - Evidence from independent studies about adjacent genes shows some of the top unconfirmed hits to be promising candidates

| Selection frequency | SNP (Cluster Size) | Chrom. | Confirmed in Franke et al.'10? | Selected in WTCCC '07? |
|---|---|---|---|---|
| 100% | rs11805303 (16) | 1 | Yes | Yes |
| 100% | rs11209026 (2) | 1 | Yes | Yes |
| 100% | rs6431654 (20) | 2 | Yes | Yes |
| 100% | rs6601764 (1) | 10 | No | No |
| 100% | rs7095491 (18) | 10 | Yes | Yes |
| 90% | rs6688532 (33) | 1 | Yes | No |
| 90% | rs17234657 (1) | 5 | Yes | Yes |
| 90% | rs3135503 (16) | 16 | Yes | Yes |
| 80% | rs9783122 (234) | 10 | No | No |
| 80% | rs11627513 (7) | 14 | No | No |
| 60% | rs4437159 (4) | 3 | No | No |
| 60% | rs7768538 (1145) | 6 | Yes | No |
| 60% | rs6500315 (4) | 16 | Yes | Yes |
| 60% | rs2738758 (5) | 20 | Yes | No |
| 50% | rs7726744 (46) | 5 | Yes | Yes |
| 50% | rs4246045 (46) | 5 | Yes | Yes |
| 50% | rs2390248 (13) | 7 | No | No |
| 50% | rs7186163 (6) | 16 | Yes | Yes |

Table: SNP clusters discovered to be important for CD over 10 repetitions of knockoffs. Clusters not found in Franke et al. ('10) represent promising sites, especially rs6601764 and rs4692386, whose nearest genes have been independendently linked to CD

# Data analysis issues

(1) Covariance estimation
(2) Highly correlated SNPs
(3) Computational issues (knockoffs construction)

# Data analysis issues

(1) Covariance estimation
(2) Highly correlated SNPs
(3) Computational issues (knockoffs construction)

**Covariance estimation**

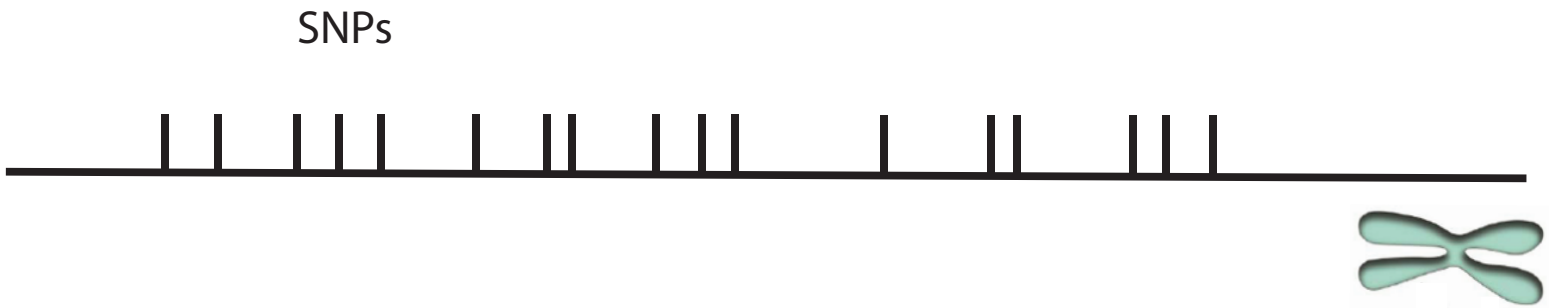
- Methodology of Wen and Stephens (2010)

  Shrink off-diagonal entries of empirical covariance matrix using genetic distance info estimated from the HapMap CEU population

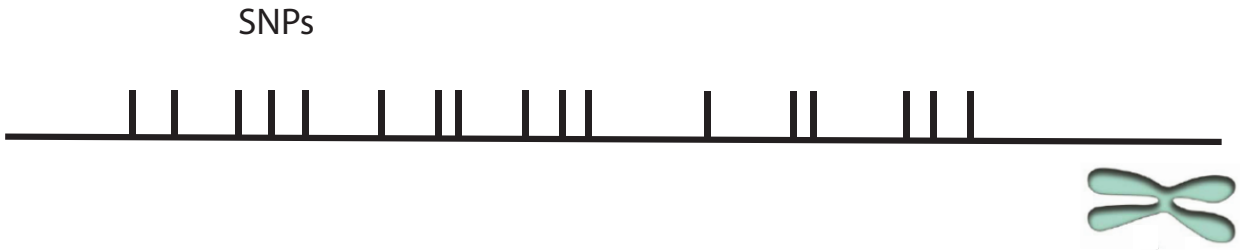
- Second-order knockoffs assuming $\hat{\mathbf{\Sigma}}$ as population covariance

# High correlations

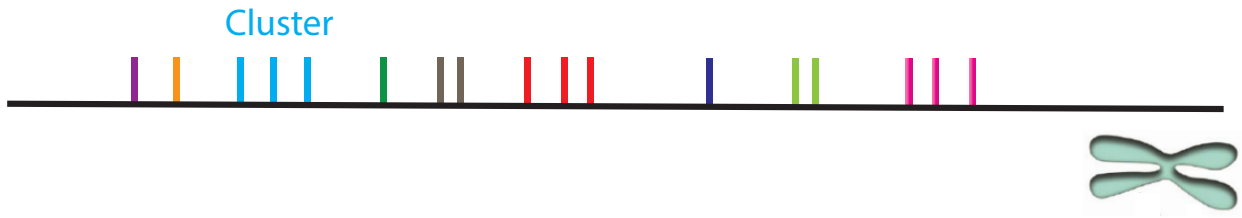
Hard to choose between two or more nearly-identical variables if the data supports at least one of them being selected



# Clustering

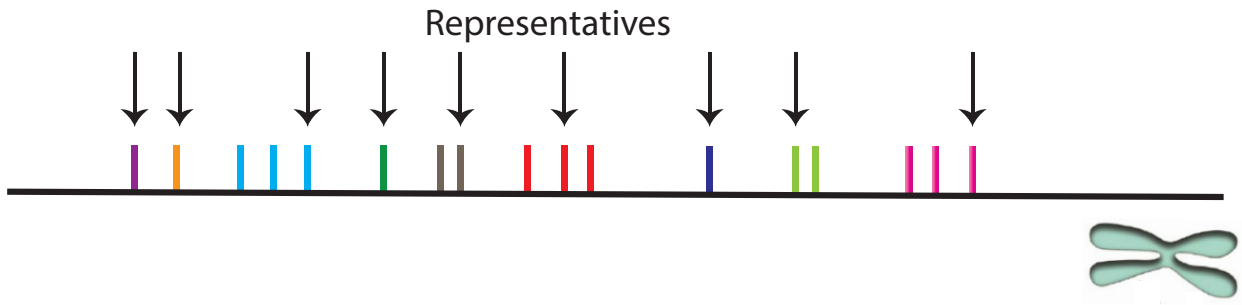


# Clustering



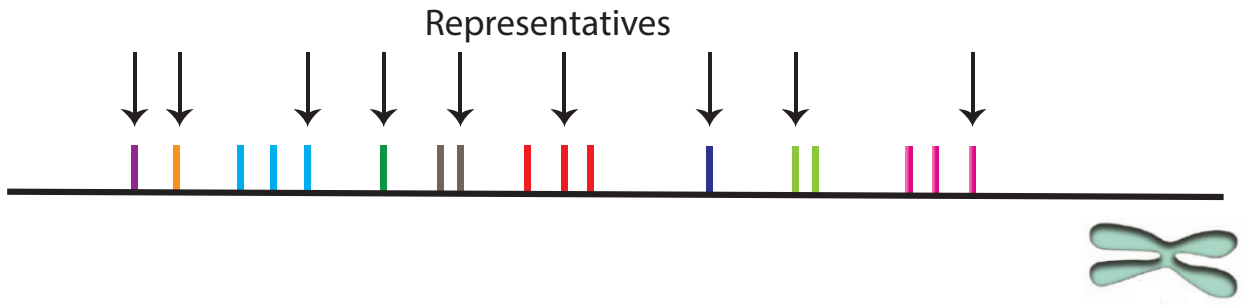

- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5
  - $\rightsquigarrow$ settle for discovering important SNP clusters among 71,145 candidates

# Clustering



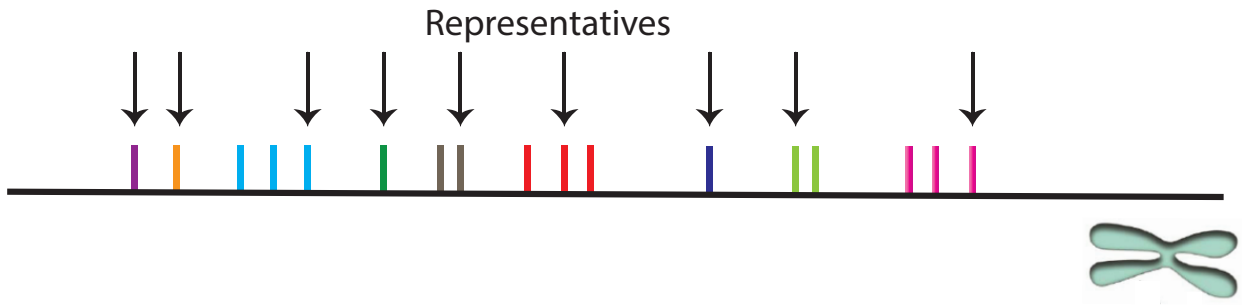

- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5
  - $\rightsquigarrow$ settle for discovering important SNP clusters among 71,145 candidates
- Cluster variables? Choose a representative SNP from each cluster
  - $\rightsquigarrow$ approximate null: cluster rep $\perp\!\!\!\perp Y \,|\,$ other reps

# Clustering



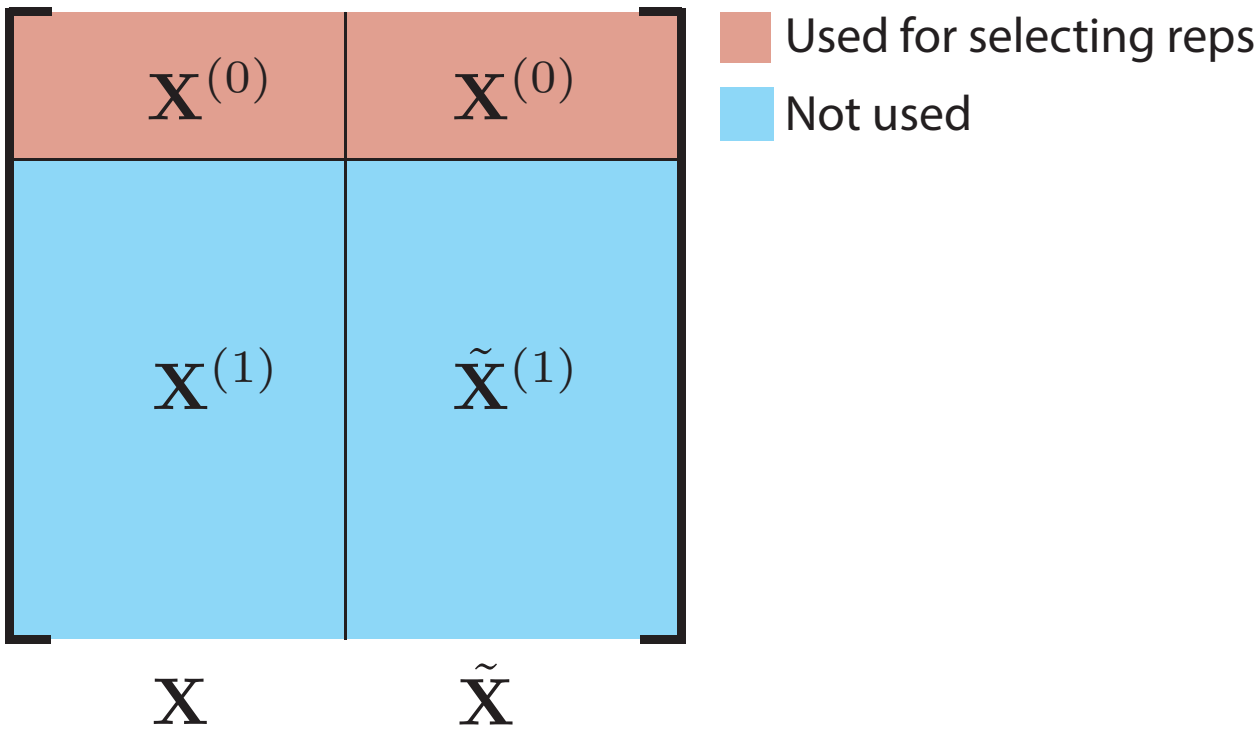

- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5
  $\rightsquigarrow$ settle for discovering important SNP clusters among 71,145 candidates
- Cluster variables? Choose a representative SNP from each cluster
  $\rightsquigarrow$ approximate null: cluster rep $\perp\!\!\!\perp Y \,|\,$ other reps
- Which rep? Most significant SNP as computed on 20% of the samples
  $\rightsquigarrow$ test all clusters (no cluster selection) $\neq$ inference after selection

# Clustering




- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5
  $\rightsquigarrow$ settle for discovering important SNP clusters among 71,145 candidates
- Cluster variables? Choose a representative SNP from each cluster
  $\rightsquigarrow$ approximate null: cluster rep $\perp\!\!\!\perp Y \,|\,$ other reps
- Which rep? Most significant SNP as computed on 20% of the samples
  $\rightsquigarrow$ test all clusters (no cluster selection) $\neq$ inference after selection
- Safe data re-use (optimize power) as in Foygel Barber and C. (2016)

# Safe data re-use



# Simulations with genetic covariates

- Real genetic covariates $X$
- Logistic conditional model $Y \mid X$ with 60 variables

# Simulations with genetic covariates

- Real genetic covariates $X$
- Logistic conditional model $Y \mid X$ with 60 variables

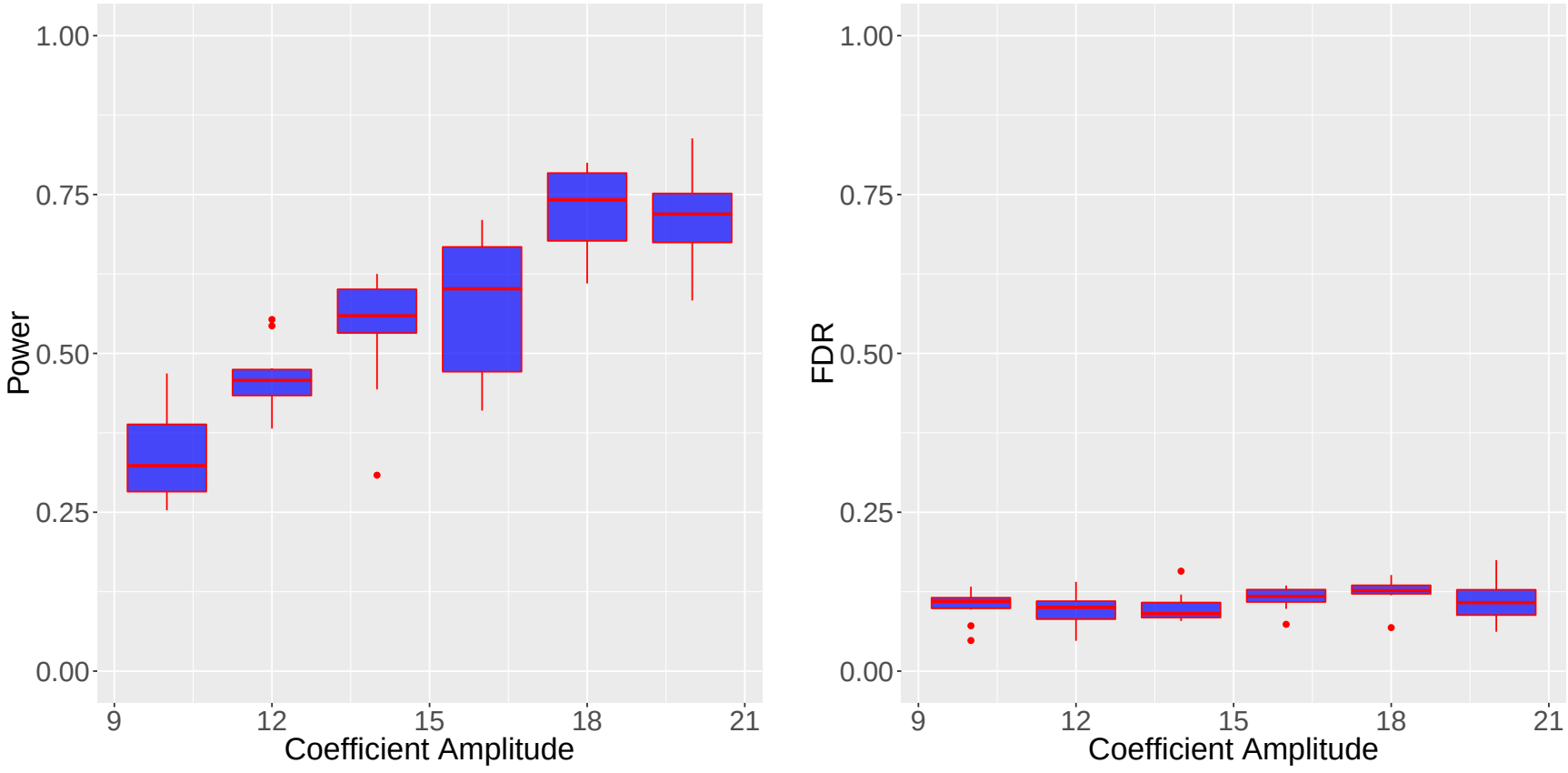



Figure: Power and FDR (target is 10%) for knockoffs with LCD statistic. Each boxplot represents 10 $\neq$ models with $\neq$ $X$'s and $Y$'s

| Selection frequency | SNP (Cluster Size) | Chrom. | Confirmed in Franke et al. '10? | Selected in WTCCC '07? |
|---|---|---|---|---|
| 100% | rs11805303 (16) | 1 | Yes | Yes |
| 100% | rs11209026 (2) | 1 | Yes | Yes |
| 100% | rs6431654 (20) | 2 | Yes | Yes |
| 100% | rs6601764 (1) | 10 | No | No |
| 100% | rs7095491 (18) | 10 | Yes | Yes |
| 90% | rs6688532 (33) | 1 | Yes | No |
| 90% | rs17234657 (1) | 5 | Yes | Yes |
| 90% | rs3135503 (16) | 16 | Yes | Yes |
| 80% | rs9783122 (234) | 10 | No | No |
| 80% | rs11627513 (7) | 14 | No | No |
| 60% | rs4437159 (4) | 3 | No | No |
| 60% | rs7768538 (1145) | 6 | Yes | No |
| 60% | rs6500315 (4) | 16 | Yes | Yes |
| 60% | rs2738758 (5) | 20 | Yes | No |
| 50% | rs7726744 (46) | 5 | Yes | Yes |
| 50% | rs4246045 (46) | 5 | Yes | Yes |
| 50% | rs2390248 (13) | 7 | No | No |
| 50% | rs7186163 (6) | 16 | Yes | Yes |

Table: SNP clusters discovered to be important for CD over 10 repetitions of knockoffs. Clusters not found in Franke et al. (2010) represent promising sites, especially rs6601764 and rs4692386, whose nearest genes have been independendently linked to CD

# *Knockoffs for Fixed Designs*

*Joint with Foygel Barber*

# Fixed designs

- Sometimes we do not know the distribution of $X$
- Sometimes $X$ is not random

Wish that inferential properties hold *conditionally on the value of* $X$

# Controlled variable selection in linear model

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

# Controlled variable selection in linear model

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

Goal: select set of features $X_j$ that are likely to be relevant without too many false positives

$$\underbrace{\text{FDR}}_{\text{False discovery rate}} = \mathbb{E}\Big[\underbrace{\frac{\#\text{ false positives}}{\#\text{ features selected}}}_{\text{False discovery proportion}}\Big] \qquad \text{`}0/0 = 0\text{'}$$

# Controlled variable selection in linear model

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

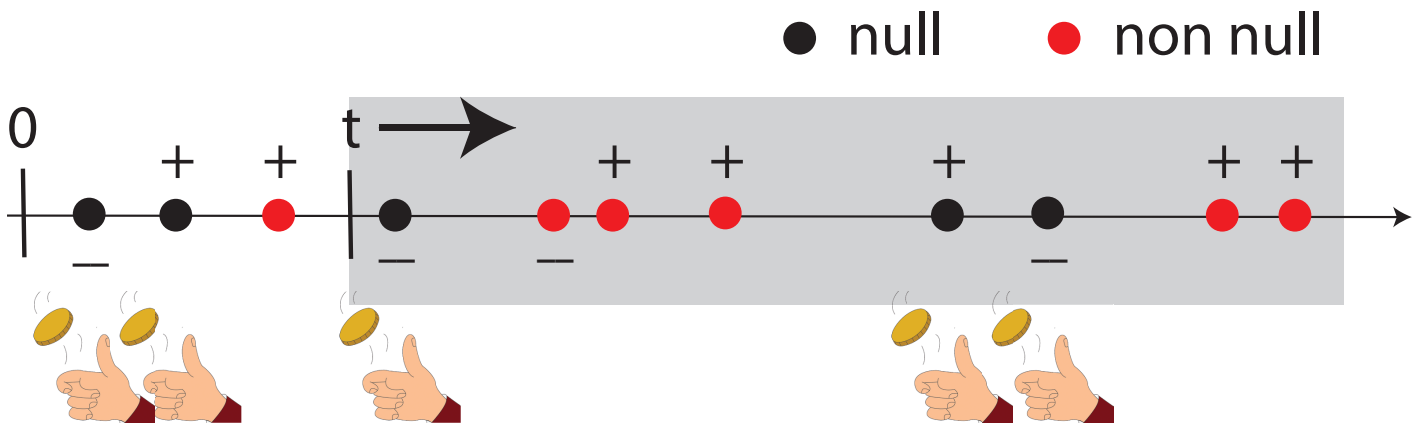
Goal: select set of features $X_j$ that are likely to be relevant without too many false positives

$$\underbrace{\text{FDR}}_{\text{False discovery rate}} = \mathbb{E}\Big[\underbrace{\frac{\#\text{ false positives}}{\#\text{ features selected}}}_{\text{False discovery proportion}}\Big] \qquad \text{`}0/0 = 0\text{'}$$

Context of multiple testing (with possibly many irrelevant variables)

$$H_j : \beta_j = 0 \qquad j = 1, \ldots, p$$

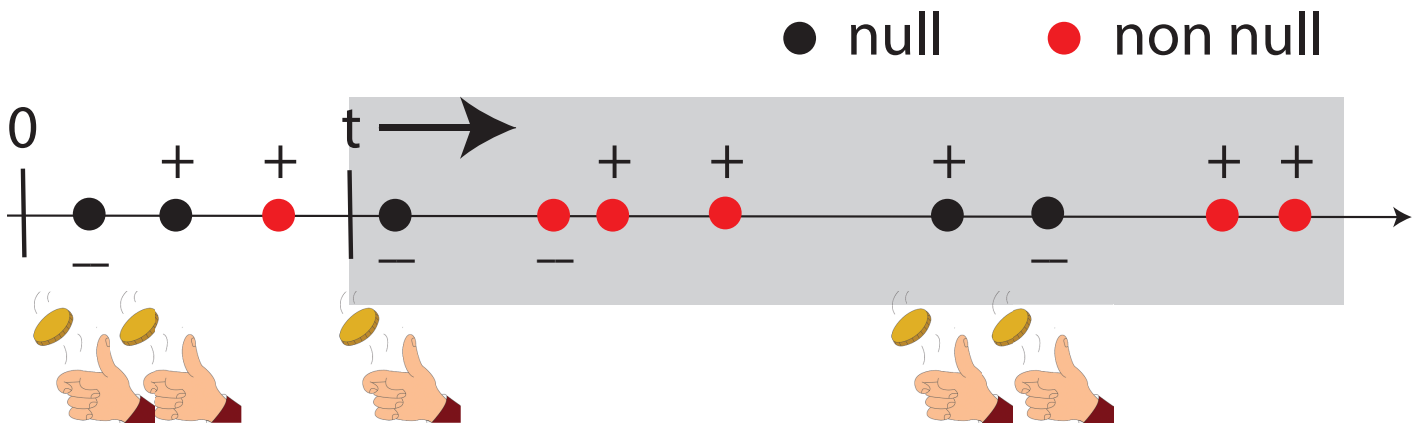
# Objective




(1) Signs of null statisitics are iid coin flips

(2) Good ordering

# Objective



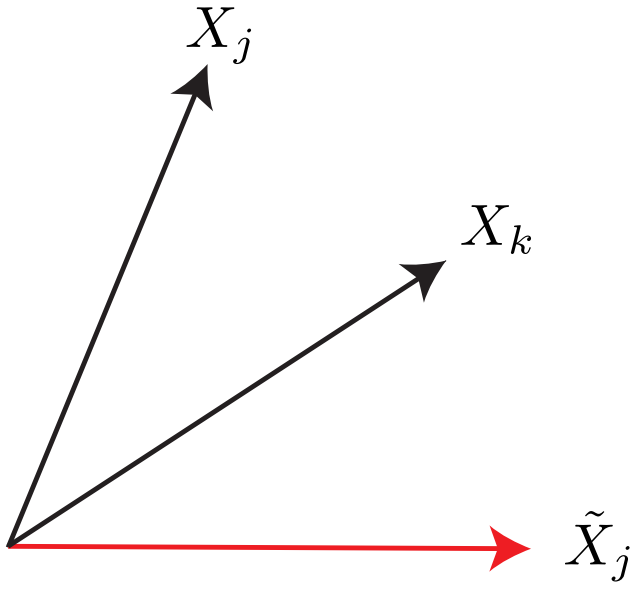

(1) Signs of null statisitics are iid coin flips

(2) Good ordering

If (1) then machinery applies $\Longrightarrow$ exact FDR control

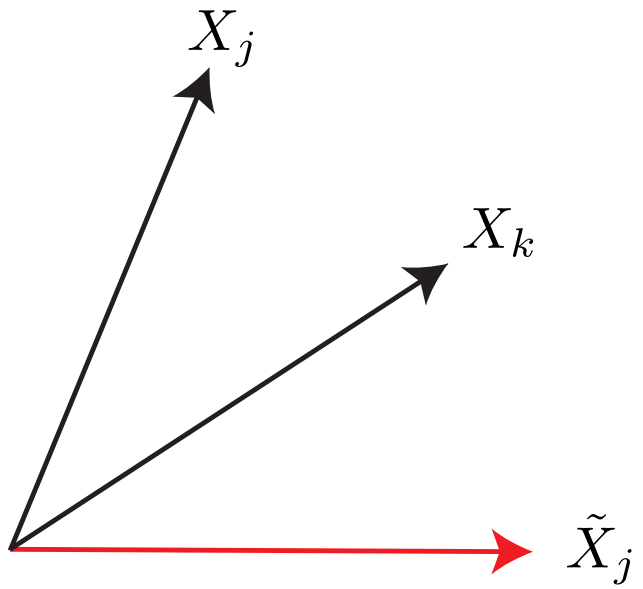
# Knockoff features for fixed designs

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



# Knockoff features for fixed designs

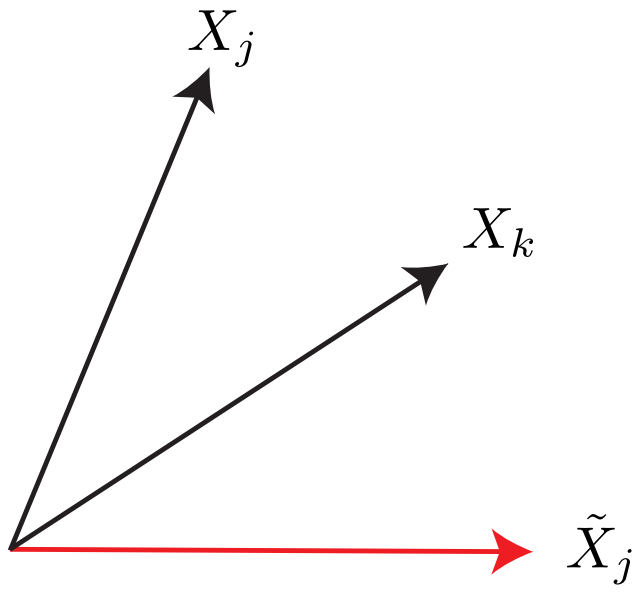
$$\tilde{X}_j' \tilde{X}_k = X_j' X_k \qquad \text{for all } j, k$$
$$\tilde{X}_j' X_k = X_j' X_k \qquad \text{for all } j \neq k$$




$$[X \;\; \tilde{X}]^\top [X \;\; \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} \succeq 0 \qquad s \in \mathbb{R}^p$$

# Knockoff features for fixed designs

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




$$[X \;\; \tilde{X}]^\top [X \;\; \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} \succeq 0 \qquad s \in \mathbb{R}^p$$

Construction via matrix computations and/or numerical optimization

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

# Symmetric statistics $\{W_j\}$

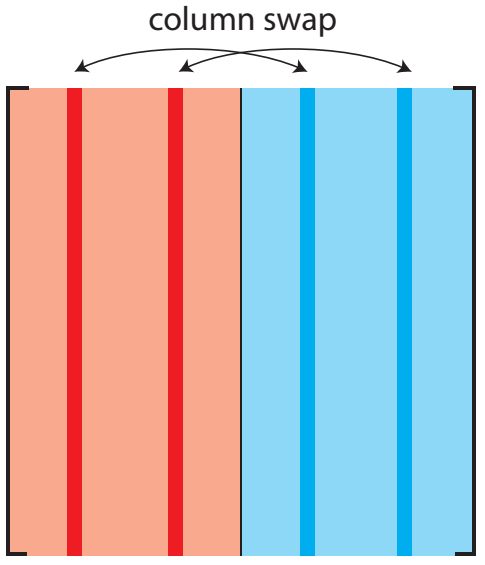
(1) *Anti-symmetry property:* swapping changes signs

$$W_j\left([X \;\; \tilde{X}]_{\mathsf{swap}(S)}, y\right) = W_j\left([X \;\; \tilde{X}], y\right) \cdot \begin{cases} +1 & j \notin S \\ -1 & j \in S \end{cases}$$

# Symmetric statistics $\{W_j\}$

(1) *Anti-symmetry property:* swapping changes signs

$$W_j\left(\left[X \;\; \tilde{X}\right]_{\mathsf{swap}(S)}, y\right) = W_j\left(\left[X \;\; \tilde{X}\right], y\right) \cdot \begin{cases} +1 & j \notin S \\ -1 & j \in S \end{cases}$$



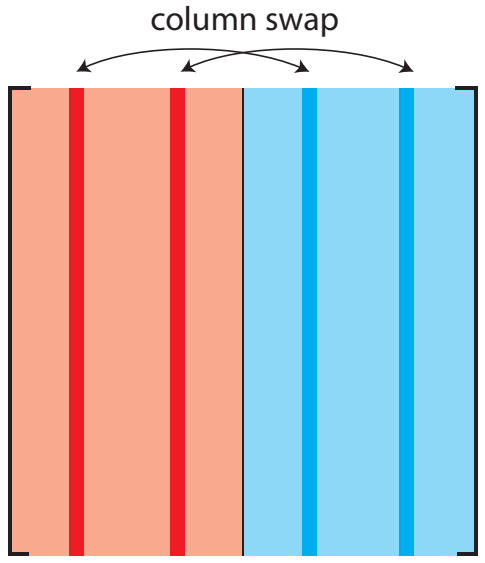

(2) *Sufficiency property:*

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

# Symmetric statistics $\{W_j\}$

(1) *Anti-symmetry property:* swapping changes signs

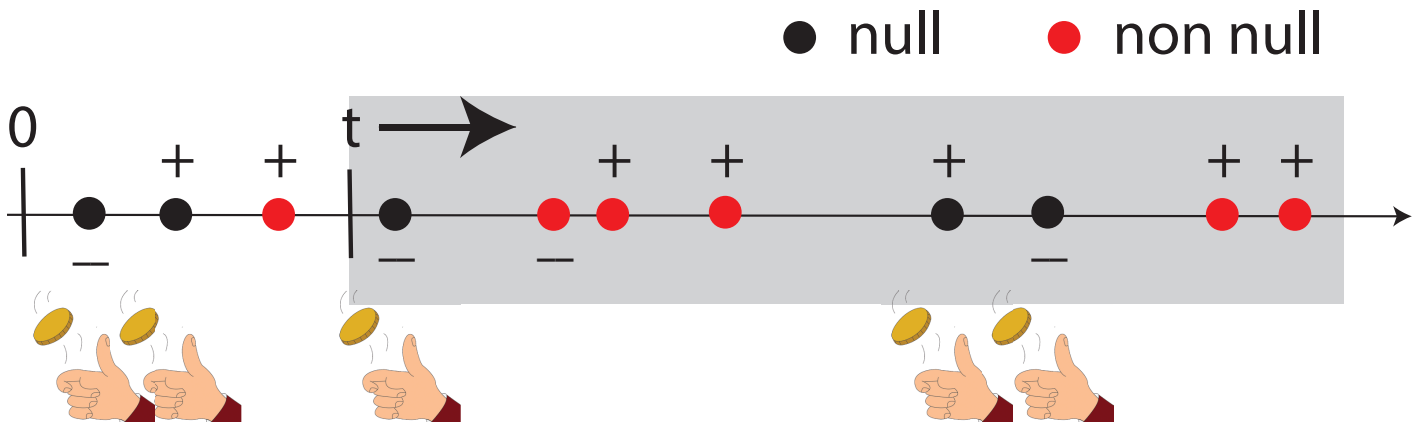
$$W_j\left([X \;\; \tilde{X}]_{\text{swap}(S)}, y\right) = W_j\left([X \;\; \tilde{X}], y\right) \cdot \begin{cases} +1 & j \notin S \\ -1 & j \in S \end{cases}$$




(2) *Sufficiency property:*

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

Guarantees iid coin flips for nulls



# Examples of statistics

- $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$, $j = 1, \ldots, 2p$, and $\hat{\beta}(\lambda)$ sol. to augmented Lasso

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p}) \qquad W_j = Z_j - Z_{j+p}$$

# Examples of statistics

- $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$, $j = 1, \ldots, 2p$, and $\hat{\beta}(\lambda)$ sol. to augmented Lasso

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p}) \qquad W_j = Z_j - Z_{j+p}$$

- Same with penalized estimate

$$\min \quad \tfrac{1}{2}\|y - Xb\|_2^2 + \lambda P(b)$$

# Examples of statistics

- $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$, $j = 1, \ldots, 2p$, and $\hat{\beta}(\lambda)$ sol. to augmented Lasso

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p}) \qquad W_j = Z_j - Z_{j+p}$$

- Same with penalized estimate

$$\min \quad \tfrac{1}{2}\|y - Xb\|_2^2 + \lambda P(b)$$

- Forward selection/orthogonal matching pursuit: $Z_1, \cdots, Z_{2p}$ (reverse) order in which $2p$ variables entered model

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p})$$

# Examples of statistics

- $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$, $j = 1, \ldots, 2p$, and $\hat{\beta}(\lambda)$ sol. to augmented Lasso

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p}) \qquad W_j = Z_j - Z_{j+p}$$

- Same with penalized estimate

$$\min \quad \tfrac{1}{2}\|y - Xb\|_2^2 + \lambda P(b)$$

- Forward selection/orthogonal matching pursuit: $Z_1, \cdots, Z_{2p}$ (reverse) order in which $2p$ variables entered model

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p})$$

- Statistics based on LS estimates

$$W_j = |\hat{\beta}_j^{\mathsf{LS}}|^2 - |\hat{\beta}_{j+p}^{\mathsf{LS}}|^2 \qquad |\hat{\beta}_j^{\mathsf{LS}}| - |\hat{\beta}_{j+p}^{\mathsf{LS}}|$$

# Examples of statistics

- $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$, $j = 1, \ldots, 2p$, and $\hat{\beta}(\lambda)$ sol. to augmented Lasso

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p}) \qquad W_j = Z_j - Z_{j+p}$$

- Same with penalized estimate

$$\min \quad \tfrac{1}{2}\|y - Xb\|_2^2 + \lambda P(b)$$

- Forward selection/orthogonal matching pursuit: $Z_1, \cdots, Z_{2p}$ (reverse) order in which $2p$ variables entered model

$$W_j = (Z_j \vee Z_{j+p}) \cdot \operatorname{sign}(Z_j - Z_{j+p})$$

- Statistics based on LS estimates

$$W_j = |\hat{\beta}_j^{\mathsf{LS}}|^2 - |\hat{\beta}_{j+p}^{\mathsf{LS}}|^2 \qquad |\hat{\beta}_j^{\mathsf{LS}}| - |\hat{\beta}_{j+p}^{\mathsf{LS}}|$$

- ... (endless possibilities)

# HIV drug resistance

| Drug type | # drugs | Sample size | # protease or RT positions genotyped | # mutations appearing $\geq 3$ 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 $\#j$

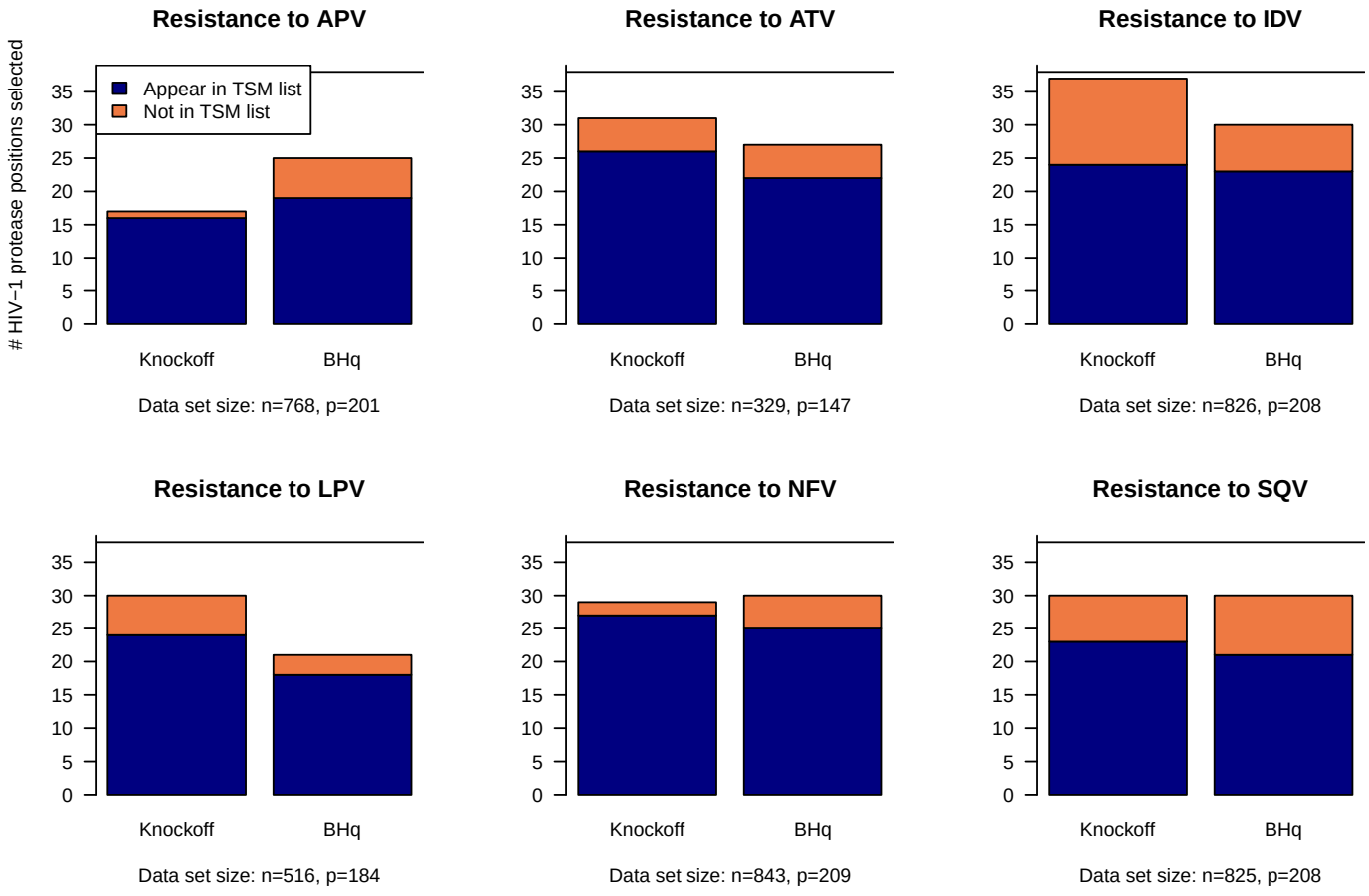
Data from R. Shafer (Stanford) available at:

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

# HIV data

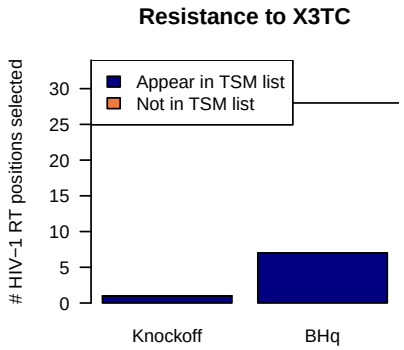
TSM list: mutations associated with the PI class of drugs in general, and is not specialized to the individual drugs in the class

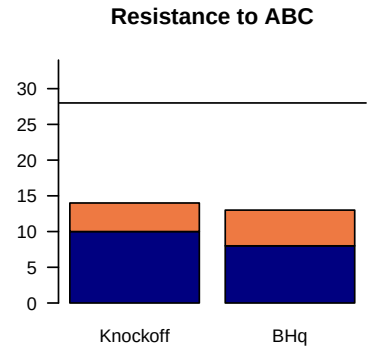
Results for PI type drugs

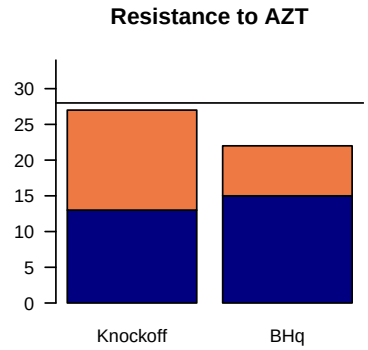


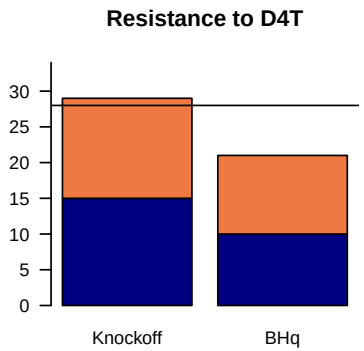
# HIV data

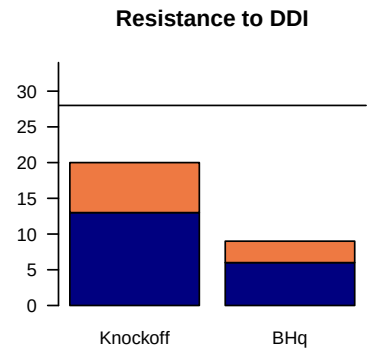
Results for
NRTI type drugs

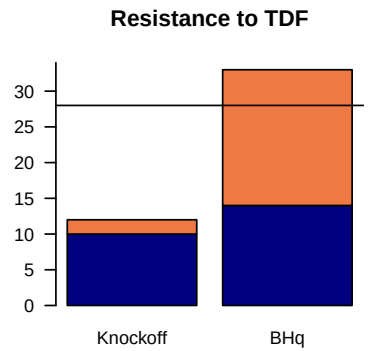


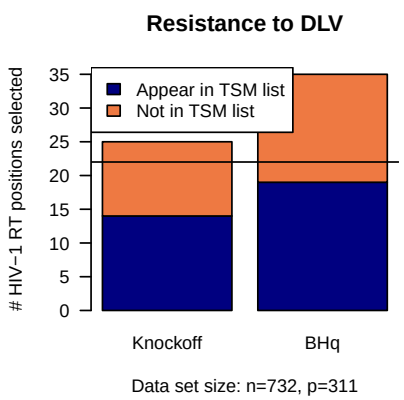



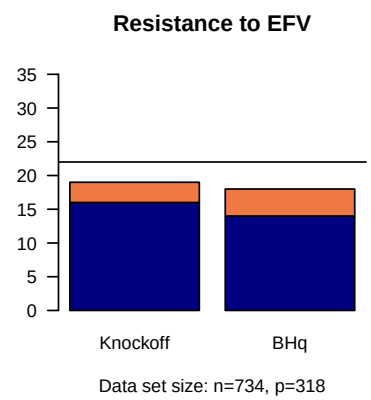



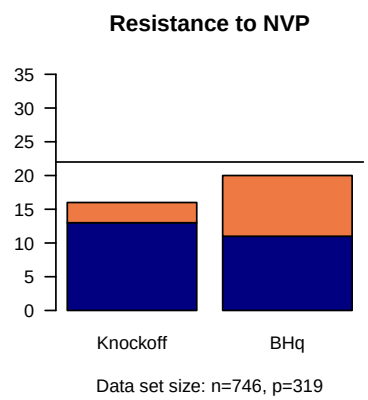




Results for
NNRTI type drugs



# Summary and open questions

**Knockoff filter = inference machine**

*You design the statistics, knockoffs take care of inference*

# Summary and open questions

**Knockoff filter = inference machine**

*You design the statistics, knockoffs take care of inference*

- Knockoffs construction (tomorrow)
- Robustness
- Which feature importance statistics should we use?
- Derandomization (multiple knockoffs)

# Summary and open questions

**Knockoff filter = inference machine**

*You design the statistics, knockoffs take care of inference*

- Knockoffs construction (tomorrow)
- Robustness
- Which feature importance statistics should we use?
- Derandomization (multiple knockoffs)

E. Candès, Y. Fan, L. Janson, and J. Lv. "Panning for gold: model-free knockoffs for high-dimensional controlled variable selection" (2016)

R. Foygel Barber and E. Candès. "Controlling the false discovery rate via knockoffs" *Annals of Statistics* (Oct. 2015)