Model-Free Knockoffs: High-Dimensional Variable Selection that Controls the False Discovery Rate

Lucas Janson, Stanford Department of Statistics

WADAPT Workshop, NIPS, December 2016

Collaborators: Emmanuel Candès (Stanford), YingYing Fan, Jinchi Lv (USC)

Problem Statement

Controlled Variable Selection

Given:

- Y an outcome of interest (AKA response or dependent variable),
- $X_1, \dots, X_p$ a set of p potential explanatory variables (AKA covariates, features, or independent variables),

How can we select important explanatory variables with few mistakes?

Problem Statement

Controlled Variable Selection

Given:

- Y an outcome of interest (AKA response or dependent variable),
- $X_1, \ldots, X_p$ a set of p potential explanatory variables (AKA covariates, features, or independent variables),

How can we select important explanatory variables with few mistakes?

Applications to:

- Medicine/genetics/health care

Problem Statement

Controlled Variable Selection

Given:

- Y an outcome of interest (AKA response or dependent variable),
- $X_1, \ldots, X_p$ a set of p potential explanatory variables (AKA covariates, features, or independent variables),

How can we select important explanatory variables with few mistakes?

Applications to:

- Medicine/genetics/health care
- Economics/political science

Problem Statement

Controlled Variable Selection

Given:

- Y an outcome of interest (AKA response or dependent variable),
- $X_1, \ldots, X_p$ a set of p potential explanatory variables (AKA covariates, features, or independent variables),

How can we select important explanatory variables with few mistakes?

Applications to:

- Medicine/genetics/health care
- Economics/political science
- Industry/technology

Controlled Variable Selection

What is an important variable?

Controlled Variable Selection

What is an important variable?

We consider X_j to be unimportant if the conditional distribution of Y given $X_1, \ldots, X_p$ does not depend on X_j. Formally, X_j is unimportant if it is conditionally independent of Y given $X_{\text{-}j}$:

$$Y \perp\!\!\!\perp X_j \,|\, X_{\text{-}j}$$

Controlled Variable Selection

What is an important variable?

We consider X_j to be unimportant if the conditional distribution of Y given $X_1, \ldots, X_p$ does not depend on X_j. Formally, X_j is unimportant if it is conditionally independent of Y given $X_{\text{-}j}$:

$$Y \perp\!\!\!\perp X_j \,|\, X_{\text{-}j}$$

Markov Blanket of Y: smallest set S such that $Y \perp\!\!\!\perp X_{\text{-}S} \,|\, X_S$

Controlled Variable Selection

What is an important variable?

We consider X_j to be unimportant if the conditional distribution of Y given $X_1, \ldots, X_p$ does not depend on X_j. Formally, X_j is unimportant if it is conditionally independent of Y given $X_{\text{-}j}$:

$$Y \perp\!\!\!\perp X_j \,|\, X_{\text{-}j}$$

Markov Blanket of Y: smallest set S such that $Y \perp\!\!\!\perp X_{\text{-}S} \,|\, X_S$

To make sure we do not make too many mistakes, we seek to select a set $\hat{S}$ to control the **false discovery rate (FDR)**:

$$\mathrm{FDR}(\hat{S}) = \mathbb{E}\left(\frac{\#\{j \text{ in } \hat{S} \ : \ X_j \text{ unimportant}\}}{\#\{j \text{ in } \hat{S}\}}\right) \le q \ \ \text{(e.g. 10\%)}$$

"Here is a set of variables $\hat{S}$, 90% of which I expect to be important"

Sneak Peak

Model-free knockoffs solves the controlled variable selection problem

- Any model for Y and $X_1, \ldots, X_p$
- Any dimension (including $p > n$)
- Finite-sample control (non-asymptotic) of FDR
- Practical performance on real problems

Sneak Peak

Model-free knockoffs solves the controlled variable selection problem

- Any model for Y and $X_1, \dots, X_p$
- Any dimension (including $p > n$)
- Finite-sample control (non-asymptotic) of FDR
- Practical performance on real problems

Application: the Genetic Basis of Crohn's Disease (WTCCC, 2007)

- $\approx 5,000$ subjects ($\approx 40\%$ with Crohn's Disease)
- $\approx 375,000$ single nucleotide polymorphisms (SNPs) for each subject

Sneak Peak

Model-free knockoffs solves the controlled variable selection problem

- Any model for Y and $X_1, \ldots, X_p$
- Any dimension (including $p > n$)
- Finite-sample control (non-asymptotic) of FDR
- Practical performance on real problems

Application: the Genetic Basis of Crohn's Disease (WTCCC, 2007)

- $\approx 5,000$ subjects ($\approx 40\%$ with Crohn's Disease)
- $\approx 375,000$ single nucleotide polymorphisms (SNPs) for each subject

The original analysis of the data made **9** discoveries by running marginal tests of association on each SNP and applying a p-value cutoff corresponding (by a Bayesian argument, under assumptions) to a FDR of 10%

Sneak Peak

Model-free knockoffs solves the controlled variable selection problem

- Any model for Y and $X_1, \ldots, X_p$
- Any dimension (including $p > n$)
- Finite-sample control (non-asymptotic) of FDR
- Practical performance on real problems

Application: the Genetic Basis of Crohn's Disease (WTCCC, 2007)

- $\approx 5,000$ subjects ($\approx 40\%$ with Crohn's Disease)
- $\approx 375,000$ single nucleotide polymorphisms (SNPs) for each subject

The original analysis of the data made **9** discoveries by running marginal tests of association on each SNP and applying a p-value cutoff corresponding (by a Bayesian argument, under assumptions) to a FDR of 10%

Model-free knockoffs used the same FDR of 10% and made **18** discoveries, with many of the new discoveries confirmed by a larger meta-analysis

Methods for Controlled Variable Selection

What is required for valid inference?

	Low dimensions	Model for Y	Asymptopic regime	Sparsity	Random design
OLSp+BHq	Yes	Yes	No	No	No

Methods for Controlled Variable Selection

What is required for valid inference?

	Low dimensions	Model for Y	Asymptopic regime	Sparsity	Random design
OLSp+BHq	Yes	Yes	No	No	No
MLp+BHq	Yes	Yes	Yes	No	No

Methods for Controlled Variable Selection

What is required for valid inference?

	Low dimensions	Model for Y	Asymptopic regime	Sparsity	Random design
OLSp+BHq	Yes	Yes	No	No	No
MLp+BHq	Yes	Yes	Yes	No	No
HDp+BHq	No	Yes	Yes	Yes	Yes

Methods for Controlled Variable Selection

What is required for valid inference?

	Low dimensions	Model for Y	Asymptopic regime	Sparsity	Random design
OLSp+BHq	Yes	Yes	No	No	No
MLp+BHq	Yes	Yes	Yes	No	No
HDp+BHq	No	Yes	Yes	Yes	Yes
Orig KnO	Yes	Yes	No	No	No

Methods for Controlled Variable Selection

What is required for valid inference?

	Low dimensions	Model for Y	Asymptopic regime	Sparsity	Random design
OLSp+BHq	Yes	Yes	No	No	No
MLp+BHq	Yes	Yes	Yes	No	No
HDp+BHq	No	Yes	Yes	Yes	Yes
Orig KnO	Yes	Yes	No	No	No
MF KnO	No	No	No	No	Yes*

The Knockoffs Framework

The generic knockoffs procedure for controlling the FDR at level q:

(1) **Construct knockoffs**:

- Artificial versions ("knockoffs") of each variable
- Act as controls for assessing importance of original variables

The Knockoffs Framework

The generic knockoffs procedure for controlling the FDR at level q:

(1) **Construct knockoffs**:

- Artificial versions ("knockoffs") of each variable
- Act as controls for assessing importance of original variables

(2) **Compute knockoff statistics**:

- Scalar statistic W_j for each variable
- Measures how much more important a variable appears than its knockoff
- Positive W_j denotes original more important, strength measured by magnitude

The Knockoffs Framework

The generic knockoffs procedure for controlling the FDR at level q:

(1) **Construct knockoffs**:

- Artificial versions ("knockoffs") of each variable
- Act as controls for assessing importance of original variables

(2) **Compute knockoff statistics**:

- Scalar statistic W_j for each variable
- Measures how much more important a variable appears than its knockoff
- Positive W_j denotes original more important, strength measured by magnitude

(3) **Find the knockoff threshold**:

- Order the variables by decreasing $|W_j|$
- Going down the list, select variables with positive W_j
- Stop at last time the ratio of negatives to positives is below q

The Knockoffs Framework

The generic knockoffs procedure for controlling the FDR at level q:

(1) **Construct knockoffs**:
- Artificial versions ("knockoffs") of each variable
- Act as controls for assessing importance of original variables

(2) **Compute knockoff statistics**:
- Scalar statistic W_j for each variable
- Measures how much more important a variable appears than its knockoff
- Positive W_j denotes original more important, strength measured by magnitude

(3) **Find the knockoff threshold**:
- Order the variables by decreasing $|W_j|$
- Going down the list, select variables with positive W_j
- Stop at last time the ratio of negatives to positives is below q

Coin-flipping property: The key to the knockoffs procedure is that steps (1) and (2) are done specifically to ensure that, conditional on $|W_1|, \ldots, |W_p|$, the signs of the *unimportant/null* W_j are independently ± 1 with probability $1/2$

The Model-Free Knockoffs Procedure

The model-free knockoffs procedure for controlling the FDR at level q:

(1) **Construct knockoffs**: Exchangeability

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \overset{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

(requires joint distribution of $X_1, \ldots, X_p$ known)

The Model-Free Knockoffs Procedure

The model-free knockoffs procedure for controlling the FDR at level q:

(1) **Construct knockoffs**: Exchangeability

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \overset{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

(requires joint distribution of $X_1, \ldots, X_p$ known)

(2) **Compute knockoff statistics**:

- Variable importance measure Z
- Antisymmetric function $f_j : \mathbb{R}^2 \to \mathbb{R}$, i.e.,

$$f_j(z_1, z_2) = -f_j(z_2, z_1)$$

- $W_j = f_j(Z_j, \widetilde{Z}_j)$, where Z_j and $\widetilde{Z}_j$ are the variable importances of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$, respectively

The Model-Free Knockoffs Procedure

The model-free knockoffs procedure for controlling the FDR at level q:

(1) **Construct knockoffs**: Exchangeability

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \overset{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

(requires joint distribution of $X_1, \ldots, X_p$ known)

(2) **Compute knockoff statistics**:

- Variable importance measure Z
- Antisymmetric function $f_j : \mathbb{R}^2 \to \mathbb{R}$, i.e.,

$$f_j(z_1, z_2) = -f_j(z_2, z_1)$$

- $W_j = f_j(Z_j, \widetilde{Z}_j)$, where Z_j and $\widetilde{Z}_j$ are the variable importances of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$, respectively

(3) **Find the knockoff threshold**: just requires coin-flipping property

Known Covariate Distribution

Model-free knockoffs surprisingly **robust to overfitting**

Known Covariate Distribution

Model-free knockoffs surprisingly **robust to overfitting**

Reasonable approximation when:

1. Subjects sampled from a population, and

Known Covariate Distribution

Model-free knockoffs surprisingly **robust to overfitting**

Reasonable approximation when:

1. Subjects sampled from a population, and

2a. X_j highly structured, well-studied, or well-understood, OR

Known Covariate Distribution

Model-free knockoffs surprisingly **robust to overfitting**

Reasonable approximation when:

1. Subjects sampled from a population, and

2a. X_j highly structured, well-studied, or well-understood, OR

2b. Large set of unsupervised X data (without Y's)

Known Covariate Distribution

Model-free knockoffs surprisingly **robust to overfitting**

Reasonable approximation when:

1. Subjects sampled from a population, and

2a. X_j highly structured, well-studied, or well-understood, OR

2b. Large set of unsupervised X data (without Y's)

For instance, many **genome-wide association studies** satisfy all conditions:

1. Subjects sampled from a population (oversampling cases still valid)

Known Covariate Distribution

Model-free knockoffs surprisingly **robust to overfitting**

Reasonable approximation when:

1. Subjects sampled from a population, and

2a. X_j highly structured, well-studied, or well-understood, OR

2b. Large set of unsupervised X data (without Y's)

For instance, many **genome-wide association studies** satisfy all conditions:

1. Subjects sampled from a population (oversampling cases still valid)

2a. Strong spatial structure: linkage disequilibrium models, e.g., Markov chains, are well-studied and work well

Known Covariate Distribution

Model-free knockoffs surprisingly **robust to overfitting**

Reasonable approximation when:

1. Subjects sampled from a population, and

2a. X_j highly structured, well-studied, or well-understood, OR

2b. Large set of unsupervised X data (without Y's)

For instance, many **genome-wide association studies** satisfy all conditions:

1. Subjects sampled from a population (oversampling cases still valid)

2a. Strong spatial structure: linkage disequilibrium models, e.g., Markov chains, are well-studied and work well

2b. Other studies have collected same or similar SNP arrays on different subjects

Knockoff Construction

Valid model-free knockoff variables can always be generated:

Algorithm 1 Sequential Conditional Independent Pairs

for $j = \{1, \ldots, p\}$ **do**

  Sample $\tilde{X}_j$ from $\mathcal{L}(X_j \,|\, X_{\text{-}j}, \tilde{X}_{1:j-1})$

end

Knockoff Construction

Valid model-free knockoff variables can always be generated:

Algorithm 1 Sequential Conditional Independent Pairs

for $j = \{1, \ldots, p\}$ **do**
  Sample $\tilde{X}_j$ from $\mathcal{L}(X_j \,|\, X_{\text{-}j}, \tilde{X}_{1:j-1})$
end

If $(X_1, \ldots, X_p)$ multivariate Gaussian, exchangeability reduces to matching first and second moments when X_j, $\tilde{X}_j$ swapped

For $\mathrm{Cov}(X_1, \ldots, X_p) = \boldsymbol{\Sigma}$:

$$\mathrm{Cov}(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p) = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

In non-Gaussian case, can be thought of as second-order-correct model-free knockoffs

Exchangeability Endows Coin-Flipping

Recall exchangeability property:

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \stackrel{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

for any j

Exchangeability Endows Coin-Flipping

Recall exchangeability property:

$$
\begin{aligned}
& [\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \\
\overset{\mathcal{D}}{=} \, & [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]
\end{aligned}
$$

for any j

Coin-flipping property for W_j:

Exchangeability Endows Coin-Flipping

Recall exchangeability property:

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \overset{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

for any j

Coin-flipping property for W_j: for any *unimportant* variable j,

$$\Big(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big) \overset{\mathcal{D}}{=} \Big(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big)$$

Exchangeability Endows Coin-Flipping

Recall exchangeability property:

$$
\begin{aligned}
&[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \\
\overset{\mathcal{D}}{=} &[\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]
\end{aligned}
$$

for any j

Coin-flipping property for W_j: for any *unimportant* variable j,

$$
\begin{aligned}
&\Big(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big) \\
\overset{\mathcal{D}}{=} &\Big(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big) \\
= &\Big(\widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big)
\end{aligned}
$$

Exchangeability Endows Coin-Flipping

Recall exchangeability property:

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \overset{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

for any j

Coin-flipping property for W_j: for any *unimportant* variable j,

$$\begin{aligned}
&\left(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\right) \\
\overset{\mathcal{D}}{=} &\left(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\right) \\
= &\left(\widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\right)
\end{aligned}$$

$$W_j = f_j(Z_j, \widetilde{Z}_j) \overset{\mathcal{D}}{=} f_j(\widetilde{Z}_j, Z_j)$$

Exchangeability Endows Coin-Flipping

Recall exchangeability property:

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p]$$
$$\stackrel{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

for any j

Coin-flipping property for W_j: for any *unimportant* variable j,

$$\begin{aligned}
&\Big(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big) \\
\stackrel{\mathcal{D}}{=} &\Big(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big) \\
= &\Big(\widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\Big)
\end{aligned}$$

$$W_j = f_j(Z_j, \widetilde{Z}_j) \stackrel{\mathcal{D}}{=} f_j(\widetilde{Z}_j, Z_j) = -f_j(Z_j, \widetilde{Z}_j) = -W_j$$

Exchangeability Endows Coin-Flipping

Recall exchangeability property:

$$[\boldsymbol{X}_1 \cdots \boldsymbol{X}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \tilde{\boldsymbol{X}}_p] \stackrel{\mathcal{D}}{=} [\boldsymbol{X}_1 \cdots \tilde{\boldsymbol{X}}_j \cdots \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdots \boldsymbol{X}_j \cdots \tilde{\boldsymbol{X}}_p]$$

for any j

Coin-flipping property for W_j: for any *unimportant* variable j,

$$\begin{aligned}
&\left(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\right) \\
\stackrel{\mathcal{D}}{=} &\left(Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), \widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\right) \\
= &\left(\widetilde{Z}_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p]), Z_j(\boldsymbol{y}, [\boldsymbol{X}_1 \cdot\cdot \boldsymbol{X}_j \cdot\cdot \boldsymbol{X}_p \, \tilde{\boldsymbol{X}}_1 \cdot\cdot \tilde{\boldsymbol{X}}_j \cdot\cdot \tilde{\boldsymbol{X}}_p])\right)
\end{aligned}$$

$$W_j \stackrel{\mathcal{D}}{=} -W_j$$

Adaptivity and Prior Information in W_j

Lasso Coefficient Difference (LCD): ℓ_1-penalized regression of $\boldsymbol{y}$ on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$

$$W_j = |\beta_j| - |\tilde{\beta}_j|$$

Adaptivity and Prior Information in W_j

Lasso Coefficient Difference (LCD): ℓ_1-penalized regression of $\boldsymbol{y}$ on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$

$$W_j = |\beta_j| - |\tilde{\beta}_j|$$

Adaptivity

- Cross-validation (on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$) to choose the penalty parameter in the lasso

Adaptivity and Prior Information in W_j

Lasso Coefficient Difference (LCD): ℓ_1-penalized regression of $\boldsymbol{y}$ on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$

$$W_j = |\beta_j| - |\tilde{\beta}_j|$$

Adaptivity

- Cross-validation (on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$) to choose the penalty parameter in the lasso
- Higher-level adaptivity: CV to choose best-fitting model for inference

Adaptivity and Prior Information in W_j

Lasso Coefficient Difference (LCD): ℓ_1-penalized regression of $\boldsymbol{y}$ on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$

$$W_j = |\beta_j| - |\tilde{\beta}_j|$$

Adaptivity

- Cross-validation (on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$) to choose the penalty parameter in the lasso
- Higher-level adaptivity: CV to choose best-fitting model for inference
- Fit random forest and ℓ_1-penalized regression; derive feature importance from whichever has lower CV error—**still strict FDR control**

Adaptivity and Prior Information in W_j

Lasso Coefficient Difference (LCD): ℓ_1-penalized regression of $\boldsymbol{y}$ on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$

$$W_j = |\beta_j| - |\tilde{\beta}_j|$$

Adaptivity

- Cross-validation (on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$) to choose the penalty parameter in the lasso
- Higher-level adaptivity: CV to choose best-fitting model for inference
- Fit random forest and ℓ_1-penalized regression; derive feature importance from whichever has lower CV error—**still strict FDR control**

Prior information

- **Bayesian approach**: choose prior and model, and Z_j could be the posterior probability that X_j contributes to the model

Adaptivity and Prior Information in W_j

Lasso Coefficient Difference (LCD): ℓ_1-penalized regression of $\boldsymbol{y}$ on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$

$$W_j = |\beta_j| - |\tilde{\beta}_j|$$

Adaptivity

- Cross-validation (on $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$) to choose the penalty parameter in the lasso
- Higher-level adaptivity: CV to choose best-fitting model for inference
- Fit random forest and ℓ_1-penalized regression; derive feature importance from whichever has lower CV error—**still strict FDR control**

Prior information

- **Bayesian approach**: choose prior and model, and Z_j could be the posterior probability that X_j contributes to the model
- Still strict FDR control, **even if wrong prior or MCMC has not converged**

Summary and Next Steps

Summary

- The **controlled variable selection** problem arises in many important modern statistical applications, but remained unsolved in all but the simplest settings

Summary and Next Steps

Summary

- The **controlled variable selection** problem arises in many important modern statistical applications, but remained unsolved in all but the simplest settings
- **Model-free knockoffs** is a powerful, adaptive, and robust solution whenever there is considerable outside information on the covariate distribution, which includes some of the most pressing applications such as GWAS

Summary and Next Steps

Summary

- The **controlled variable selection** problem arises in many important modern statistical applications, but remained unsolved in all but the simplest settings
- **Model-free knockoffs** is a powerful, adaptive, and robust solution whenever there is considerable outside information on the covariate distribution, which includes some of the most pressing applications such as GWAS

Next steps

- *Theoretical*: rigorous results on robustness

Summary and Next Steps

Summary

- The **controlled variable selection** problem arises in many important modern statistical applications, but remained unsolved in all but the simplest settings
- **Model-free knockoffs** is a powerful, adaptive, and robust solution whenever there is considerable outside information on the covariate distribution, which includes some of the most pressing applications such as GWAS

Next steps

- *Theoretical*: rigorous results on robustness
- *Applied*: domain-specific knockoff constructions and knockoff statistics for interesting applications, e.g., gene knockout/knockdown

Summary and Next Steps

Summary

- The **controlled variable selection** problem arises in many important modern statistical applications, but remained unsolved in all but the simplest settings
- **Model-free knockoffs** is a powerful, adaptive, and robust solution whenever there is considerable outside information on the covariate distribution, which includes some of the most pressing applications such as GWAS

Next steps

- *Theoretical*: rigorous results on robustness
- *Applied*: domain-specific knockoff constructions and knockoff statistics for interesting applications, e.g., gene knockout/knockdown

Thank you!

Appendix

References

Athey, S., Imbens, G. W., and Wager, S. (2016). Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2016). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202.

Wen, X. and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.*, 4(3):1158–1182.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Original Knockoffs (Barber and Candès, 2015)

$\boldsymbol{y}$ and $\boldsymbol{X}_j$ are $n \times 1$ column vectors of data: n draws from the random variables Y and X_j, respectively; design matrix $\boldsymbol{X} := [\boldsymbol{X}_1 \cdots \boldsymbol{X}_p]$

Original Knockoffs (Barber and Candès, 2015)

$\boldsymbol{y}$ and $\boldsymbol{X}_j$ are $n \times 1$ column vectors of data: n draws from the random variables Y and X_j, respectively; design matrix $\boldsymbol{X} := [\boldsymbol{X}_1 \cdots \boldsymbol{X}_p]$

(1) **Construct knockoffs**: Knockoffs $\tilde{\boldsymbol{X}}_j$ must satisfy, ($\tilde{\boldsymbol{X}} := [\tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_p]$)

$$[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^\top [\boldsymbol{X}\ \tilde{\boldsymbol{X}}] = \left[\begin{array}{cc} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{X} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{X}^\top \boldsymbol{X} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{X}^\top \boldsymbol{X} \end{array} \right]$$

Original Knockoffs (Barber and Candès, 2015)

$\boldsymbol{y}$ and $\boldsymbol{X}_j$ are $n \times 1$ column vectors of data: n draws from the random variables Y and X_j, respectively; design matrix $\boldsymbol{X} := [\boldsymbol{X}_1 \cdots \boldsymbol{X}_p]$

(1) **Construct knockoffs**: Knockoffs $\tilde{\boldsymbol{X}}_j$ must satisfy, ($\tilde{\boldsymbol{X}} := [\tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_p]$)

$$[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^\top [\boldsymbol{X}\ \tilde{\boldsymbol{X}}] = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{X} - \operatorname{diag}\{\boldsymbol{s}\} \\ \boldsymbol{X}^\top \boldsymbol{X} - \operatorname{diag}\{\boldsymbol{s}\} & \boldsymbol{X}^\top \boldsymbol{X} \end{bmatrix}$$

(2) **Compute knockoff statistics**:

- Sufficiency: W_j only a function of $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^\top [\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$ and $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^\top \boldsymbol{y}$
- Antisymmetry: swapping values of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ flips sign of W_j

Original Knockoffs (Barber and Candès, 2015)

$\boldsymbol{y}$ and $\boldsymbol{X}_j$ are $n \times 1$ column vectors of data: n draws from the random variables Y and X_j, respectively; design matrix $\boldsymbol{X} := [\boldsymbol{X}_1 \cdots \boldsymbol{X}_p]$

(1) **Construct knockoffs**: Knockoffs $\tilde{\boldsymbol{X}}_j$ must satisfy, ($\tilde{\boldsymbol{X}} := [\tilde{\boldsymbol{X}}_1 \cdots \tilde{\boldsymbol{X}}_p]$)

$$[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^\top [\boldsymbol{X}\ \tilde{\boldsymbol{X}}] = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{X} - \mathrm{diag}\{\boldsymbol{s}\} \\ \boldsymbol{X}^\top \boldsymbol{X} - \mathrm{diag}\{\boldsymbol{s}\} & \boldsymbol{X}^\top \boldsymbol{X} \end{bmatrix}$$

(2) **Compute knockoff statistics**:

- Sufficiency: W_j only a function of $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^\top [\boldsymbol{X}\ \tilde{\boldsymbol{X}}]$ and $[\boldsymbol{X}\ \tilde{\boldsymbol{X}}]^\top \boldsymbol{y}$
- Antisymmetry: swapping values of $\boldsymbol{X}_j$ and $\tilde{\boldsymbol{X}}_j$ flips sign of W_j

Comments:

- Finite-sample FDR control (non-asymptotic)
- Sparsity-based W_j for greater power than OLS+BHq
- Requires data follow Gaussian linear model
- Can only be run in low dimensions ($n \geq p$)
- Sufficiency requirement restricts choice of W_j, limiting power/adaptivity

Robustness Simulations

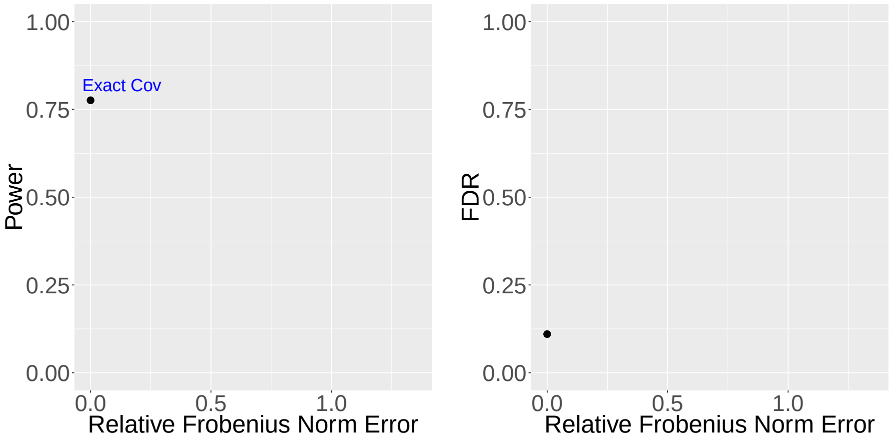

Figure: Covariates are **AR(1) with autocorrelation coefficient 0.3**. $n = 800$, $p = 1500$, and target FDR is 10%. Y comes from a binomial linear model with logit link function with 50 nonzero entries.

Robustness Simulations

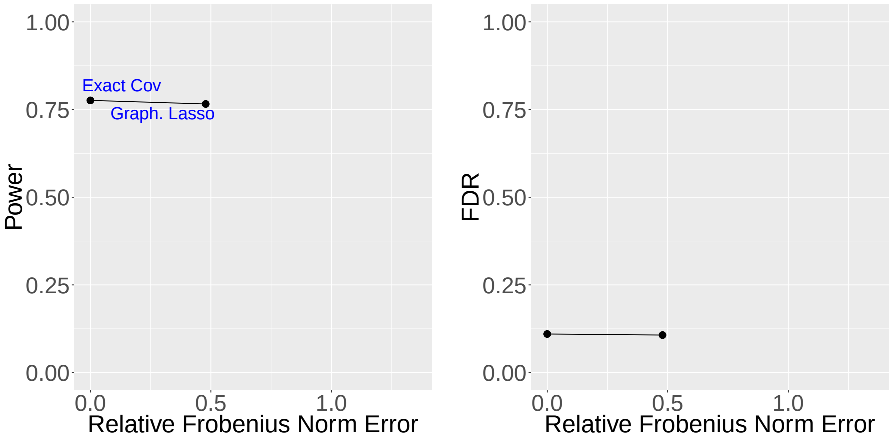

Figure: Covariates are **AR(1) with autocorrelation coefficient 0.3**. $n = 800$, $p = 1500$, and target FDR is 10%. Y comes from a binomial linear model with logit link function with 50 nonzero entries.

Robustness Simulations

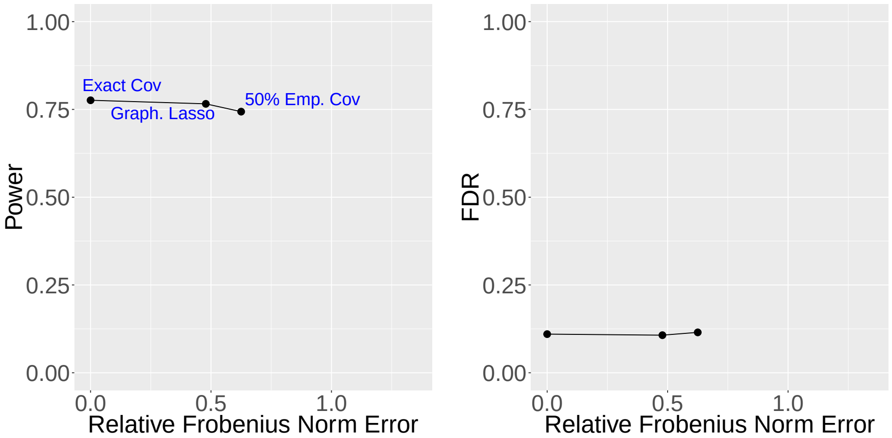

Figure: Covariates are **AR(1) with autocorrelation coefficient 0.3**. $n = 800$, $p = 1500$, and target FDR is 10%. Y comes from a binomial linear model with logit link function with 50 nonzero entries.

Robustness Simulations

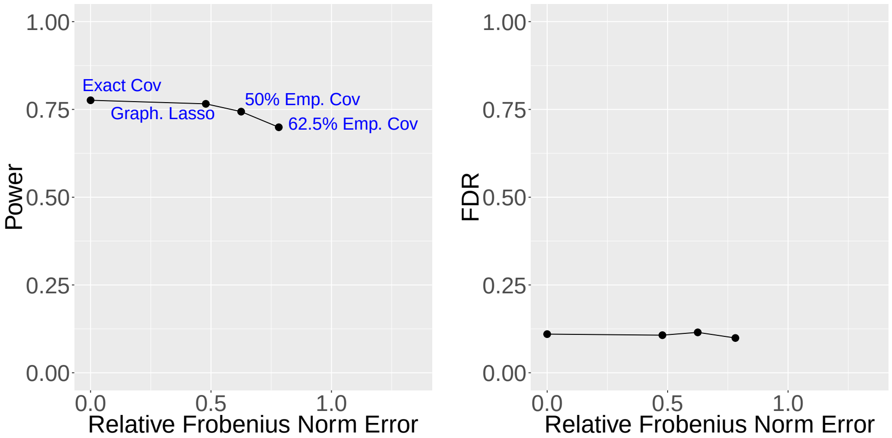

Figure: Covariates are **AR(1) with autocorrelation coefficient 0.3**. $n = 800$, $p = 1500$, and target FDR is 10%. Y comes from a binomial linear model with logit link function with 50 nonzero entries.

Robustness Simulations

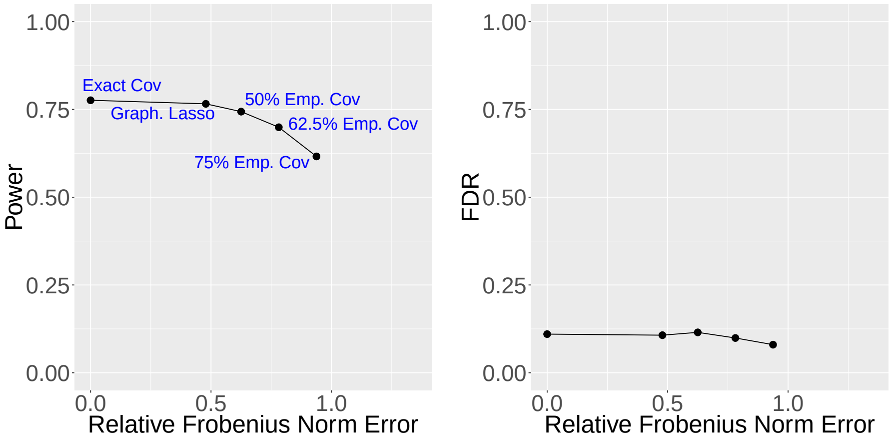

Figure: Covariates are **AR(1) with autocorrelation coefficient 0.3**. $n = 800$, $p = 1500$, and target FDR is 10%. Y comes from a binomial linear model with logit link function with 50 nonzero entries.

Robustness Simulations

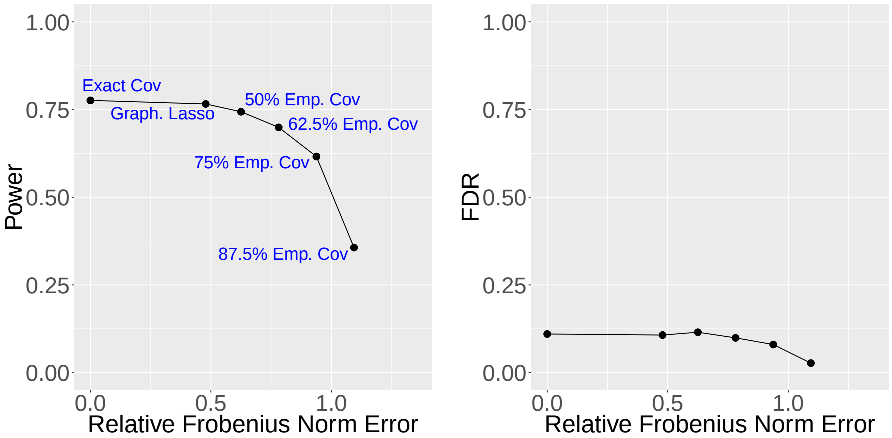

Figure: Covariates are **AR(1) with autocorrelation coefficient 0.3**. $n = 800$, $p = 1500$, and target FDR is 10%. Y comes from a binomial linear model with logit link function with 50 nonzero entries.

Robustness Simulations

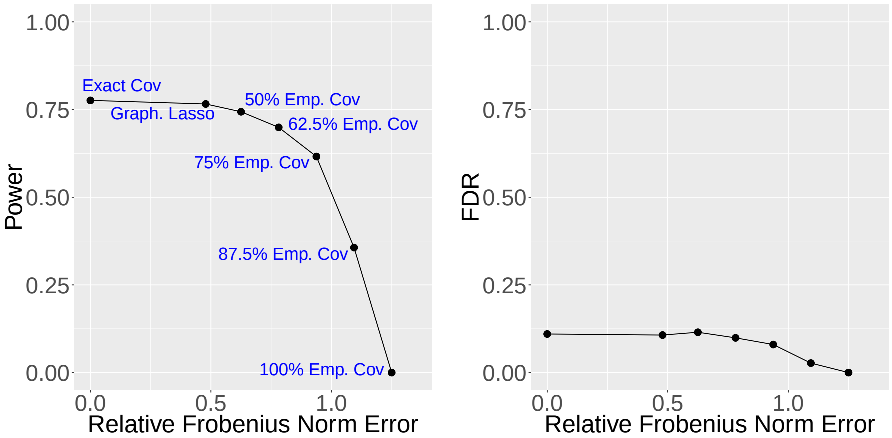

Figure: Covariates are **AR(1) with autocorrelation coefficient 0.3**. $n = 800$, $p = 1500$, and target FDR is 10%. Y comes from a binomial linear model with logit link function with 50 nonzero entries.

Robustness on Real Data

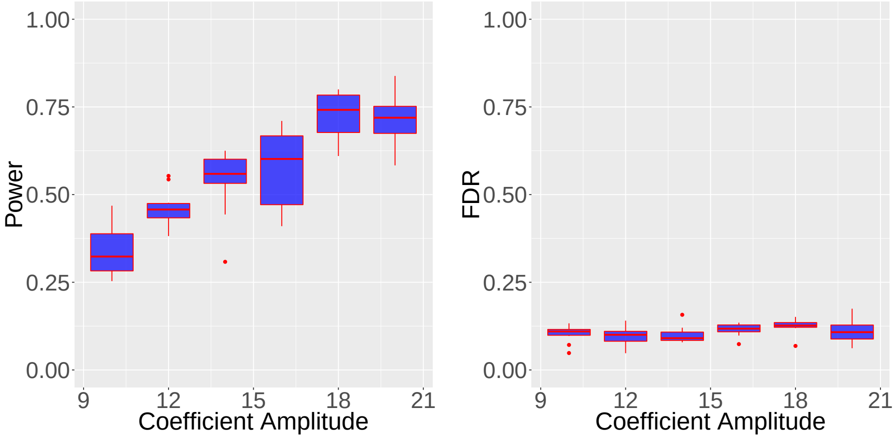

Figure: Power and FDR (target is 10%) for model-free knockoffs applied to subsamples of a real genetic design matrix. $n \approx 1,400$, $p \approx 70,000$, and each boxplot represents 10 different logistic regression models with 60 nonzero coefficients, while each sample in each boxplot is an average over 10 design matrices drawn from actual SNP data.

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis
- **Strong spatial structure**: second-order approximate SDP knockoffs on covariance estimate of Wen and Stephens (2010) which shrinks off-diagonal entries of empirical covariance using HapMap spatial structure

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis
- **Strong spatial structure**: second-order approximate SDP knockoffs on covariance estimate of Wen and Stephens (2010) which shrinks off-diagonal entries of empirical covariance using HapMap spatial structure
- Nearby SNPs had very high correlations: affects power

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis
- **Strong spatial structure**: second-order approximate SDP knockoffs on covariance estimate of Wen and Stephens (2010) which shrinks off-diagonal entries of empirical covariance using HapMap spatial structure
- Nearby SNPs had very high correlations: affects power
- SNPs clustered into groups of average size ≈ 5; each group represented by a single SNP chosen by t-test on a held-out subset of data: $p \longrightarrow 70,000$

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis
- **Strong spatial structure**: second-order approximate SDP knockoffs on covariance estimate of Wen and Stephens (2010) which shrinks off-diagonal entries of empirical covariance using HapMap spatial structure
- Nearby SNPs had very high correlations: affects power
- SNPs clustered into groups of average size ≈ 5; each group represented by a single SNP chosen by t-test on a held-out subset of data: $p \longrightarrow 70,000$
- Checked robustness by running entire procedure on repeated subsamples of larger design matrix, with simulated response

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis
- **Strong spatial structure**: second-order approximate SDP knockoffs on covariance estimate of Wen and Stephens (2010) which shrinks off-diagonal entries of empirical covariance using HapMap spatial structure
- Nearby SNPs had very high correlations: affects power
- SNPs clustered into groups of average size ≈ 5; each group represented by a single SNP chosen by t-test on a held-out subset of data: $p \longrightarrow 70,000$
- Checked robustness by running entire procedure on repeated subsamples of larger design matrix, with simulated response
- Model-free knockoffs makes **twice as many discoveries** as original analysis

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis
- **Strong spatial structure**: second-order approximate SDP knockoffs on covariance estimate of Wen and Stephens (2010) which shrinks off-diagonal entries of empirical covariance using HapMap spatial structure
- Nearby SNPs had very high correlations: affects power
- SNPs clustered into groups of average size ≈ 5; each group represented by a single SNP chosen by t-test on a held-out subset of data: $p \longrightarrow 70,000$
- Checked robustness by running entire procedure on repeated subsamples of larger design matrix, with simulated response
- Model-free knockoffs makes **twice as many discoveries** as original analysis
    - Some new discoveries confirmed in larger study

Genetic Analysis of Crohn's Disease

- 2007 case-control study of Crohn's disease by WTCCC; $n \approx 5,000$, $p \approx 375,000$, preprocessing mirrored original analysis
- **Strong spatial structure**: second-order approximate SDP knockoffs on covariance estimate of Wen and Stephens (2010) which shrinks off-diagonal entries of empirical covariance using HapMap spatial structure
- Nearby SNPs had very high correlations: affects power
- SNPs clustered into groups of average size ≈ 5; each group represented by a single SNP chosen by t-test on a held-out subset of data: $p \longrightarrow 70,000$
- Checked robustness by running entire procedure on repeated subsamples of larger design matrix, with simulated response
- Model-free knockoffs makes **twice as many discoveries** as original analysis
  - Some new discoveries confirmed in larger study
  - Some corroborated by work on nearby genes: promising candidates

Simulations in Low-Dimensional Linear Model

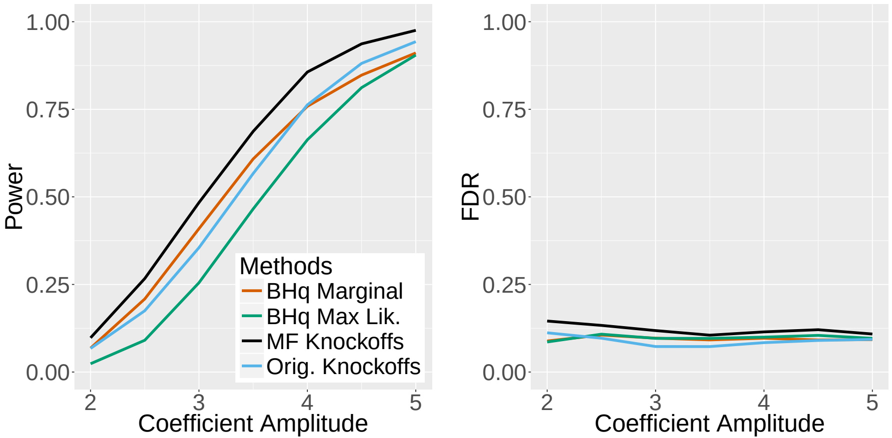

Figure: Power and FDR (target is 10%) for MF knockoffs and alternative procedures. The design matrix is i.i.d. $\mathcal{N}(0, 1/n)$, $n = 3000$, $p = 1000$, and y comes from a Gaussian linear model with 60 nonzero regression coefficients having equal magnitudes and random signs. The noise variance is 1.

Simulations in Low-Dimensional Nonlinear Model

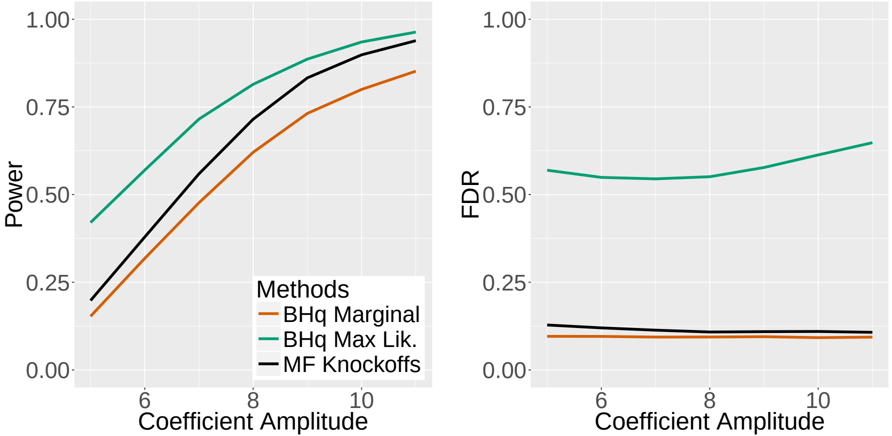

Figure: Power and FDR (target is 10%) for MF knockoffs and alternative procedures. The design matrix is i.i.d. $\mathcal{N}(0, 1/n)$, $n = 3000$, $p = 1000$, and y comes from a binomial linear model with logit link function, and 60 nonzero regression coefficients having equal magnitudes and random signs.

Simulations in High Dimensions

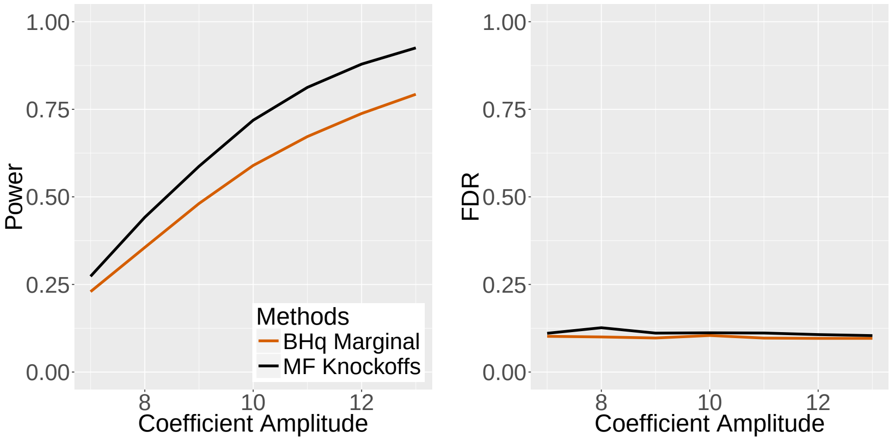

Figure: Power and FDR (target is 10%) for MF knockoffs and alternative procedures. The design matrix is i.i.d. $\mathcal{N}(0, 1/n)$, $n = 3000$, $p = 6000$, and y comes from a binomial linear model with logit link function, and 60 nonzero regression coefficients having equal magnitudes and random signs.

Simulations in High Dimensions with Dependence

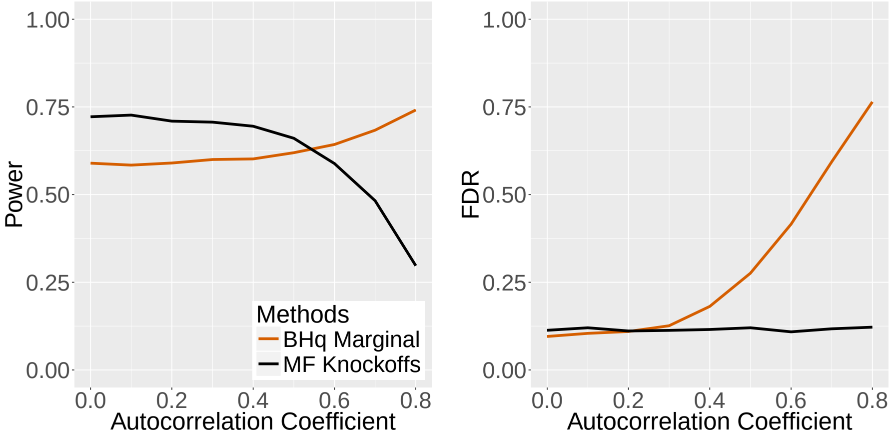

Figure: Power and FDR (target is 10%) for MF knockoffs and alternative procedures. The design matrix has AR(1) columns, and marginally each $X_j \sim \mathcal{N}(0, 1/n)$. $n = 3000$, $p = 6000$, and y follows a binomial linear model with logit link function, and 60 nonzero coefficients with random signs and randomly selected locations.