High-Dimensional Variable Selection in Nonlinear Models that Controls the False Discovery Rate

Lucas Janson

Harvard University Department of Statistics

CMSA Big Data Conference, August 18, 2017

Collaborators: Emmanuel Candès (Stanford), Yingying Fan, Jinch Li (USC)
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?**
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
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
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 (cont’d)

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 $X -$ $j$: $Y \perp \perp X_j | X -$ $j$.

Markov Blanket of $Y$: smallest set $S$ such that $Y \perp \perp X -$ $S | X_S$.

- For GLMs with no stochastically redundant covariates, equivalent to $\{j: \beta_j = 0\}$.

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): $\text{FDR}(\hat{S}) = \mathbb{E}(\#\{j \text{ in } \hat{S}: X_j \text{ unimportant}\} / \#\{j \text{ in } \hat{S}\}) \leq q$ (e.g. 10%)

"Here is a set of variables $\hat{S}$, 90% of which I expect to be important" - Lucas Janson (Harvard Statistics)
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 $X_{-j}$:

$$Y \perp \!\!\!\perp X_j \mid X_{-j}$$
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 $X_{-j}$:

$$Y \perp \perp X_j \mid X_{-j}$$

**Markov Blanket** of $Y$: smallest set $S$ such that $Y \perp \perp X_S \mid X_S$
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 $X_{-j}$:

$$Y \perp\!
\perp X_j \mid X_{-j}$$

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

For GLMs with no stochastically redundant covariates, equivalent to $\{j : \beta_j = 0\}$
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 \( X_{-j} \):

\[
Y \perp \!\!\!\perp X_j \mid X_{-j}
\]

Markov Blanket of \( Y \): smallest set \( S \) such that \( Y \perp \!\!\!\perp X_{-S} \mid X_S \)

For GLMs with no stochastically redundant covariates, equivalent to \( \{ j : \beta_j = 0 \} \)

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):

\[
\text{FDR}(\hat{S}) = \mathbb{E} \left( \frac{\# \{ j \in \hat{S} : X_j \text{ unimportant} \}}{\# \{ j \in \hat{S} \}} \right) \leq q \quad \text{(e.g. 10%)}
\]

“Here is a set of variables \( \hat{S} \), 90% of which I expect to be important”
New interpretation of knockoffs solves the controlled variable selection problem

- Allows any model for $Y$ and $X_1, \ldots, X_p$
- Allows any dimension (including $p > n$)
- Finite-sample control (non-asymptotic) of FDR
- Practical performance on real problems
New interpretation of knockoffs solves the controlled variable selection problem

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

Analysis of 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
New interpretation of knockoffs solves the controlled variable selection problem

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

Analysis of 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

Original analysis of the data made 9 discoveries by running marginal tests and selecting p-values to target a FDR of 10%
New interpretation of knockoffs solves the controlled variable selection problem

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

Analysis of 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

Original analysis of the data made 9 discoveries by running marginal tests and selecting p-values to target 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
### Review of Methods for Controlled Variable Selection

#### What is required for valid inference?

<table>
<thead>
<tr>
<th>Method</th>
<th>Low dimensions</th>
<th>Model for $Y$</th>
<th>Asymptotic regime</th>
<th>Sparsity</th>
<th>Random design</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLSp+BHq</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Lucas Janson (Harvard Statistics)
### What is required for valid inference?

<table>
<thead>
<tr>
<th>Method</th>
<th>Low dimensions</th>
<th>Model for ( Y )</th>
<th>Asymptotic regime</th>
<th>Sparsity</th>
<th>Random design</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS(_p)+(BH_q)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ML(_p)+(BH_q)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
## Review of Methods for Controlled Variable Selection

### What is required for valid inference?

<table>
<thead>
<tr>
<th></th>
<th>Low dimensions</th>
<th>Model for $Y$</th>
<th>Asymptotic regime</th>
<th>Sparsity</th>
<th>Random design</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLSp+BHq</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MLp+BHq</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>HDp+BHq</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Lucas Janson (Harvard Statistics)  

Knockoffs for HD Controlled Variable Selection
### Review of Methods for Controlled Variable Selection

**What is required for valid inference?**

<table>
<thead>
<tr>
<th>Method</th>
<th>Low dimensions</th>
<th>Model for Y</th>
<th>Asymptotic regime</th>
<th>Sparsity</th>
<th>Random design</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLSp+BHq</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MLp+BHq</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>HDp+BHq</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Orig KnO</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Lucas Janson (Harvard Statistics)

Knockoffs for HD Controlled Variable Selection
### What is required for valid inference?

<table>
<thead>
<tr>
<th>Method</th>
<th>Low dimensions</th>
<th>Model for $Y$</th>
<th>Asymptotic regime</th>
<th>Sparsity</th>
<th>Random design</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS$p+BHQ$</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ML$p+BHQ$</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>HD$p+BHQ$</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Orig KnO</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>New KnO</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes*</td>
</tr>
</tbody>
</table>
The Knockoffs Idea
y and X_j are n × 1 column vectors of data: n draws from the random variables Y and X_j, respectively; design matrix \( X := [X_1 \cdots X_p] \)
Knockoffs (Barber and Candès, 2015)

$y$ and $X_j$ are $n \times 1$ column vectors of data: $n$ draws from the random variables $Y$ and $X_j$, respectively; design matrix $X := [X_1 \cdots X_p]$

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

$$
[X \tilde{X}]^\top [X \tilde{X}] = \begin{bmatrix}
X^\top X & X^\top X - \text{diag}\{s\} \\
X^\top X - \text{diag}\{s\} & X^\top X
\end{bmatrix}
$$
\( y \) and \( X_j \) are \( n \times 1 \) column vectors of data: \( n \) draws from the random variables \( Y \) and \( X_j \), respectively; design matrix \( X := [X_1 \cdots X_p] \)

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

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

(2) **Compute knockoff statistics**:
- Sufficiency: \( W_j \) only a function of \( [X \tilde{X}]^\top [X \tilde{X}] \) and \( [X \tilde{X}]^\top y \)
- Antisymmetry: swapping values of \( X_j \) and \( \tilde{X}_j \) flips sign of \( W_j \)
Knockoffs (Barber and Candès, 2015)

$y$ and $X_j$ are $n \times 1$ column vectors of data: $n$ draws from the random variables $Y$ and $X_j$, respectively; design matrix $X := [X_1 \cdots X_p]$

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

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

(2) **Compute knockoff statistics:**
- Sufficiency: $W_j$ only a function of $[X \tilde{X}]^\top [X \tilde{X}]$ and $[X \tilde{X}]^\top y$
- Antisymmetry: swapping values of $X_j$ and $\tilde{X}_j$ flips sign of $W_j$

(3) **Find the knockoff threshold:**
- Order the variables by decreasing $|W_j|$ and proceed down list
- Select only variables with positive $W_j$ until last time $\frac{\text{negatives}}{\text{positives}} \leq q$
Knockoffs (Barber and Candès, 2015)

$y$ and $X_j$ are $n \times 1$ column vectors of data: $n$ draws from the random variables $Y$ and $X_j$, respectively; design matrix $X := [X_1 \cdots X_p]$

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

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

2. **Compute knockoff statistics**:
   - Sufficiency: $W_j$ only a function of $[X \tilde{X}]^\top [X \tilde{X}]$ and $[X \tilde{X}]^\top y$
   - Antisymmetry: swapping values of $X_j$ and $\tilde{X}_j$ flips sign of $W_j$

3. **Find the knockoff threshold**:
   - Order the variables by decreasing $|W_j|$ and proceed down list
   - Select only variables with positive $W_j$ until last time $\frac{\text{negatives}}{\text{positives}} \leq q$

Comments:
- Finite-sample FDR control and leverages sparsity for power
- Requires data follow low-dimensional ($n \geq p$) Gaussian linear model
- Canonical approach: condition on $X$, rely heavily on model for $y$
(1) **Construct knockoffs:**
- Artificial versions ("knockoffs") of each variable
- Act as controls for assessing importance of original variables
(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
Generalizing the Knockoffs Procedure

(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:** (same as before)
- Order the variables by decreasing $|W_j|$ and proceed down list
- Select only variables with positive $W_j$ until last time $\frac{\text{negatives}}{\text{positives}} \leq q$
Generalizing the Knockoffs Procedure

(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:** (same as before)
- Order the variables by decreasing $|W_j|$ and proceed down list
- Select only variables with positive $W_j$ until last time $\frac{\text{negatives}}{\text{positives}} \leq q$

**Coin-flipping property:** The key to knockoffs 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 $\pm 1$ with probability $1/2$
New Interpretation of Knockoffs
Instead of modeling \( y \) and conditioning on \( X \), condition on \( y \) and model \( X \) (shifts the burden of knowledge from \( y \) onto \( X \))
Knockoffs Without a Model for $Y$ (Candès et al., 2016)

Instead of modeling $y$ and conditioning on $X$, condition on $y$ and model $X$ (shifts the burden of knowledge from $y$ onto $X$).

Explicitly,

$$\text{rows of } X = (X_{i,1}, \ldots, X_{i,p}) \overset{\text{iid}}{\sim} G$$

where $G$ can be arbitrary but is assumed known.
Knockoffs Without a Model for $Y$ (Candès et al., 2016)

Instead of modeling $y$ and conditioning on $X$, condition on $y$ and model $X$ (shifts the burden of knowledge from $y$ onto $X$)

Explicitly,

$$\text{rows of } X = (X_{i,1}, \ldots, X_{i,p}) \overset{\text{iid}}{\sim} G$$

where $G$ can be arbitrary but is assumed known

- As compared to original knockoffs, removes
  - Restriction on dimension
  - Linear model requirement for $Y | X_1, \ldots, X_p$
  - “Sufficiency” constraint for $W_j$
Knockoffs Without a Model for $Y$ (Candès et al., 2016)

Instead of modeling $y$ and conditioning on $X$, condition on $y$ and model $X$
(shifts the burden of knowledge from $y$ onto $X$)

Explicitly,

$$\text{rows of } X = (X_{i,1}, \ldots, X_{i,p}) \overset{iid}{\sim} G$$

where $G$ can be arbitrary but is assumed known

- As compared to original knockoffs, removes
  - Restriction on dimension
  - Linear model requirement for $Y \mid X_1, \ldots, X_p$
  - “Sufficiency” constraint for $W_j$
- The rows of $X$ must be i.i.d., not the columns (covariates)
Knockoffs Without a Model for $Y$ (Candès et al., 2016)

Instead of modeling $y$ and conditioning on $X$, condition on $y$ and model $X$ (shifts the burden of knowledge from $y$ onto $X$)

Explicitly,

\[ \text{rows of } X = (X_{i,1}, \ldots, X_{i,p}) \overset{\text{iid}}{\sim} G \]

where $G$ can be arbitrary but is assumed known

- As compared to original knockoffs, removes
  - Restriction on dimension
  - Linear model requirement for $Y \mid X_1, \ldots, X_p$
  - “Sufficiency” constraint for $W_j$

- The rows of $X$ must be i.i.d., not the columns (covariates)
- **Nothing** about $y$’s distribution is assumed or need be known
Knockoffs Without a Model for $Y$ (Candès et al., 2016)

Instead of modeling $y$ and conditioning on $X$, condition on $y$ and model $X$ (shifts the burden of knowledge from $y$ onto $X$)

Explicitly,

$$\text{rows of } X = (X_{i,1}, \ldots, X_{i,p}) \overset{iid}{\sim} G$$

where $G$ can be arbitrary but is assumed known

- As compared to original knockoffs, removes
  - Restriction on dimension
  - Linear model requirement for $Y \mid X_1, \ldots, X_p$
  - “Sufficiency” constraint for $W_j$

- The rows of $X$ must be i.i.d., not the columns (covariates)

- **Nothing** about $y$’s distribution is assumed or need be known

- **Robust** to overfitting $X$’s distribution in preliminary experiments
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.
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

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

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.
Figure: Covariates are \textbf{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.
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.
**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.
When is it appropriate?

1. Subjects sampled from a population, and

2a. $X_j$ highly structured, well-studied, or well-understood, OR
Shifting the Burden of Knowledge

When is it appropriate?

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)
Shifting the Burden of Knowledge

When is it appropriate?

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)
Shifting the Burden of Knowledge

When is it appropriate?

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
Shifting the Burden of Knowledge

When is it appropriate?

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
The New Knockoffs Procedure

(1) **Construct knockoffs:** Exchangeability

\[
\begin{bmatrix}
X_1 \cdots X_j \cdots X_p & \tilde{X}_1 \cdots \tilde{X}_j \cdots \tilde{X}_p
\end{bmatrix} \overset{\mathcal{D}}{\equiv} \begin{bmatrix}
X_1 \cdots \tilde{X}_j \cdots X_p & \tilde{X}_1 \cdots X_j \cdots \tilde{X}_p
\end{bmatrix}
\]
The New Knockoffs Procedure

(1) **Construct knockoffs:** Exchangeability

\[
[X_1 \cdots X_j \cdots X_p \, \tilde{X}_1 \cdots \tilde{X}_j \cdots \tilde{X}_p] \overset{D}{=} [X_1 \cdots \tilde{X}_j \cdots X_p \, \tilde{X}_1 \cdots X_j \cdots \tilde{X}_p]
\]

(2) **Compute knockoff statistics:**

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

\[
f_j(z_1, z_2) = -f_j(z_2, z_1)
\]

- $W_j = f_j(Z_j, \tilde{Z}_j)$, where $Z_j$ and $\tilde{Z}_j$ are the variable importances of $X_j$ and $\tilde{X}_j$, respectively
The New Knockoffs Procedure

(1) **Construct knockoffs:** Exchangeability

\[
\begin{bmatrix}
X_1 & \ldots & X_j & \ldots & X_p & \tilde{X}_1 & \ldots & \tilde{X}_j & \ldots & \tilde{X}_p
\end{bmatrix}^D \equiv
\begin{bmatrix}
X_1 & \ldots & \tilde{X}_j & \ldots & X_p & \tilde{X}_1 & \ldots & X_j & \ldots & \tilde{X}_p
\end{bmatrix}
\]

(2) **Compute knockoff statistics:**
- Variable importance measure \( Z \)
- Antisymmetric function \( f_j : \mathbb{R}^2 \rightarrow \mathbb{R} \), i.e.,
  \[
  f_j(z_1, z_2) = -f_j(z_2, z_1)
  \]
- \( W_j = f_j(Z_j, \tilde{Z}_j) \), where \( Z_j \) and \( \tilde{Z}_j \) are the variable importances of \( X_j \) and \( \tilde{X}_j \), respectively

(3) **Find the knockoff threshold:** (same as before)
- Order the variables by decreasing \( |W_j| \) and proceed down list
- Select only variables with positive \( W_j \) until last time \( \frac{\text{negatives}}{\text{positives}} \leq q \)
Step (1): Construct Knockoffs
Proof that valid knockoff variables can be generated for any $X$ distribution.
Proof that valid knockoff variables can be generated for any $X$ distribution.

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

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

\[
\text{Cov}(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p) = \begin{bmatrix}
\Sigma & \Sigma - \text{diag}\{s\} \\
\Sigma - \text{diag}\{s\} & \Sigma
\end{bmatrix}
\]

For non-Gaussian $X$, still second-order-correct approximate knockoffs.
Knockoff Construction

Proof that valid knockoff variables can be generated for any $X$ distribution

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

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

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

For non-Gaussian $X$, still second-order-correct approximate knockoffs

- Linear algebra and semidefinite programming to find good $s$
- Recently: construction for Markov chains and HMMs (Sesia et al., 2017)
- Constructions also possible for grouped variables (Dai and Barber, 2016)
Step (2): Compute Knockoff Statistics
Recall $W_j$ an antisymmetric function $f_j$ of $Z_j$ and $\tilde{Z}_j$ (the variable importances of $X_j$ and $\tilde{X}_j$, respectively):

$$W_j = f_j(Z_j, \tilde{Z}_j) = -f_j(\tilde{Z}_j, Z_j)$$
Recall $W_j$, an antisymmetric function $f_j$ of $Z_j$ and $\tilde{Z}_j$ (the variable importances of $X_j$ and $\tilde{X}_j$, respectively):

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

For example,
- $Z$ is magnitude of fitted coefficient $\beta$ from a lasso regression of $y$ on $[X \, \tilde{X}]$
- $f_j(z_1, z_2) = z_1 - z_2$
Recall $W_j$ an antisymmetric function $f_j$ of $Z_j$ and $\tilde{Z}_j$ (the variable importances of $X_j$ and $\tilde{X}_j$, respectively):

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

For example,

- $Z$ is magnitude of fitted coefficient $\beta$ from a lasso regression of $y$ on $[X \ X]\$
- $f_j(z_1, z_2) = z_1 - z_2$

**Lasso Coefficient Difference (LCD) statistic:**

$$W_j = |\beta_j| - |\tilde{\beta}_j|$$
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any \( j \),

\[
\begin{bmatrix}
X_1 & \cdots & X_j & \cdots & X_p & \tilde{X}_1 & \cdots & \tilde{X}_j & \cdots & \tilde{X}_p
\end{bmatrix}
\]

\( \overset{\mathcal{D}}{=} \begin{bmatrix}
X_1 & \cdots & \tilde{X}_j & \cdots & X_p & \tilde{X}_1 & \cdots & X_j & \cdots & \tilde{X}_p
\end{bmatrix} \]
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any $j$,

$$\begin{bmatrix} X_1 & \cdots & X_j & \cdots & X_p & \tilde{X}_1 & \cdots & \tilde{X}_j & \cdots & \tilde{X}_p \end{bmatrix}$$

$$\mathcal{D} = \begin{bmatrix} X_1 & \cdots & \tilde{X}_j & \cdots & X_p & \tilde{X}_1 & \cdots & X_j & \cdots & \tilde{X}_p \end{bmatrix}$$

Coin-flipping property for $W_j$: 

Lucas Janson (Harvard Statistics)
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any $j$,

$$ [X_1 \cdots X_j \cdots X_p \tilde{X}_1 \cdots \tilde{X}_j \cdots \tilde{X}_p]$$

$$ \overset{\mathcal{D}}{=} [X_1 \cdots \tilde{X}_j \cdots X_p \tilde{X}_1 \cdots X_j \cdots \tilde{X}_p]$$

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

$$ (Z_j, \tilde{Z}_j) := \left( Z_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right), \; \tilde{Z}_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right) \right) $$
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any $j,$

$$[X_1 \ldots X_j \ldots X_p \tilde{X}_1 \ldots \tilde{X}_j \ldots \tilde{X}_p]$$

$$\overset{\mathcal{D}}{=} [X_1 \ldots \tilde{X}_j \ldots X_p \tilde{X}_1 \ldots X_j \ldots \tilde{X}_p]$$

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

$$\left( Z_j, \tilde{Z}_j \right) := \left( Z_j(y, [\ldots X_j \ldots \tilde{X}_j \ldots ]), \tilde{Z}_j(y, [\ldots X_j \ldots \tilde{X}_j \ldots ]) \right)$$

$$\overset{\mathcal{D}}{=} \left( Z_j(y, [\ldots \tilde{X}_j \ldots X_j \ldots ]), \tilde{Z}_j(y, [\ldots \tilde{X}_j \ldots X_j \ldots ]) \right)$$
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any $j$, 

$$[X_1 \cdots X_j \cdots X_p \tilde{X}_1 \cdots \tilde{X}_j \cdots \tilde{X}_p]$$

$$\mathcal{D} = [X_1 \cdots \tilde{X}_j \cdots X_p \tilde{X}_1 \cdots X_j \cdots \tilde{X}_p]$$

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

$$(Z_j, \tilde{Z}_j) := \left(Z_j\left(y, [\cdots X_j \cdots \tilde{X}_j \cdots]\right), \tilde{Z}_j\left(y, [\cdots X_j \cdots \tilde{X}_j \cdots]\right)\right)$$

$$\mathcal{D} \leftarrow \left(Z_j\left(y, [\cdots \tilde{X}_j \cdots X_j \cdots]\right), \tilde{Z}_j\left(y, [\cdots \tilde{X}_j \cdots X_j \cdots]\right)\right)$$

$$= \left(\tilde{Z}_j\left(y, [\cdots X_j \cdots \tilde{X}_j \cdots]\right), Z_j\left(y, [\cdots X_j \cdots \tilde{X}_j \cdots]\right)\right)$$
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any $j$,

$$\begin{bmatrix} X_1 & \cdots & X_j & \cdots & X_p & \tilde{X}_1 & \cdots & \tilde{X}_j & \cdots & \tilde{X}_p \end{bmatrix}$$

$$\equiv \begin{bmatrix} X_1 & \cdots & \tilde{X}_j & \cdots & X_p & \tilde{X}_1 & \cdots & X_j & \cdots & \tilde{X}_p \end{bmatrix}$$

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

$$\left( Z_j, \tilde{Z}_j \right) := \left( Z_j(y, \cdots X_j \cdots \tilde{X}_j \cdots) \right), \quad \tilde{Z}_j(y, \cdots X_j \cdots \tilde{X}_j \cdots)$$

$$\equiv \left( Z_j(y, \cdots \tilde{X}_j \cdots X_j \cdots) \right), \quad \tilde{Z}_j(y, \cdots \tilde{X}_j \cdots X_j \cdots)$$

$$= \left( \tilde{Z}_j(y, \cdots X_j \cdots \tilde{X}_j \cdots) \right), \quad Z_j(y, \cdots X_j \cdots \tilde{X}_j \cdots)$$

$$= \left( \tilde{Z}_j, Z_j \right)$$
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any $j$,

$$[X_1 \cdots X_j \cdots X_p \tilde{X}_1 \cdots \tilde{X}_j \cdots \tilde{X}_p]$$

$$\overset{\mathcal{D}}{=} [X_1 \cdots \tilde{X}_j \cdots X_p \tilde{X}_1 \cdots X_j \cdots \tilde{X}_p]$$

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

$$(Z_j, \tilde{Z}_j) := \left( Z_j\left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right), \tilde{Z}_j\left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right) \right)$$

$$\overset{\mathcal{D}}{=} \left( Z_j\left( y, \left[ \cdots \tilde{X}_j \cdots X_j \cdots \right] \right), \tilde{Z}_j\left( y, \left[ \cdots \tilde{X}_j \cdots X_j \cdots \right] \right) \right)$$

$$= \left( \tilde{Z}_j\left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right), Z_j\left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right) \right)$$

$$= \left( \tilde{Z}_j, Z_j \right)$$

$$W_j = f_j(Z_j, \tilde{Z}_j) \overset{\mathcal{D}}{=} f_j(\tilde{Z}_j, Z_j)$$
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any $j$,

$$[X_1 \cdots X_j \cdots X_p \tilde{X}_1 \cdots \tilde{X}_j \cdots \tilde{X}_p]$$

$$\mathcal{D} = [X_1 \cdots \tilde{X}_j \cdots X_p \tilde{X}_1 \cdots X_j \cdots \tilde{X}_p]$$

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

$$(Z_j, \tilde{Z}_j) := \left( Z_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right), \ \tilde{Z}_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right) \right)$$

$$\mathcal{D} = \left( Z_j \left( y, \left[ \cdots \tilde{X}_j \cdots X_j \cdots \right] \right), \ \tilde{Z}_j \left( y, \left[ \cdots \tilde{X}_j \cdots X_j \cdots \right] \right) \right)$$

$$= \left( \tilde{Z}_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right), \ Z_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right) \right)$$

$$= \left( \tilde{Z}_j, Z_j \right)$$

$$W_j = f_j(Z_j, \tilde{Z}_j) \overset{\mathcal{D}}{=} f_j(\tilde{Z}_j, Z_j) = -f_j(Z_j, \tilde{Z}_j) = -W_j$$
Exchangeability Endows Coin-Flipping

Recall exchangeability property: for any \( j \),

\[
\begin{bmatrix}
X_1 & \cdots & X_j & \cdots & X_p \\
\tilde{X}_1 & \cdots & \tilde{X}_j & \cdots & \tilde{X}_p
\end{bmatrix}
\]

\[ \mathcal{D} = \begin{bmatrix}
X_1 & \cdots & \tilde{X}_j & \cdots & X_p \\
\tilde{X}_1 & \cdots & X_j & \cdots & \tilde{X}_p
\end{bmatrix} \]

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

\[
\left( Z_j, \tilde{Z}_j \right) := \left( Z_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right), \tilde{Z}_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right) \right)
\]
\[ \overset{\mathcal{D}}{=} \left( Z_j \left( y, \left[ \cdots \tilde{X}_j \cdots X_j \cdots \right] \right), \tilde{Z}_j \left( y, \left[ \cdots \tilde{X}_j \cdots X_j \cdots \right] \right) \right) \]
\[ = \left( \tilde{Z}_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right), Z_j \left( y, \left[ \cdots X_j \cdots \tilde{X}_j \cdots \right] \right) \right) \]
\[ = \left( \tilde{Z}_j, Z_j \right) \]

\[
W_j \overset{\mathcal{D}}{=} -W_j
\]
Adaptivity and Prior Information in $W_j$

Recall LCD: $W_j = |\beta_j| - |\tilde{\beta}_j|$, where $\beta_j$, $\tilde{\beta}_j$ come from $\ell_1$-penalized regression

Adaptivity

- Cross-validation (on $[X \ X\tilde{X}]$) to choose the penalty parameter in LCD
Recall LCD: \( W_j = |\beta_j| - |\tilde{\beta}_j| \), where \( \beta_j, \tilde{\beta}_j \) come from \( \ell_1 \)-penalized regression

Adaptivity

- Cross-validation (on \([X \tilde{X}]\)) to choose the penalty parameter in LCD
- Higher-level adaptivity: CV to choose best-fitting model for inference
Adaptivity and Prior Information in $W_j$

Recall LCD: $W_j = |\beta_j| - |\tilde{\beta}_j|$, where $\beta_j$, $\tilde{\beta}_j$ come from $\ell_1$-penalized regression

Adaptivity

- Cross-validation (on $[X \tilde{X}]$) to choose the penalty parameter in LCD
- Higher-level adaptivity: CV to choose best-fitting model for inference
  - E.g., fit random forest and $\ell_1$-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
Adaptivity and Prior Information in $W_j$

Recall LCD: $W_j = |\beta_j| - |\tilde{\beta}_j|$, where $\beta_j, \tilde{\beta}_j$ come from $\ell_1$-penalized regression

Adaptivity

- Cross-validation (on $[X \tilde{X}]$) to choose the penalty parameter in LCD
- Higher-level adaptivity: CV to choose best-fitting model for inference
  - E.g., fit random forest and $\ell_1$-penalized regression; derive feature importance from whichever has lower CV error—still strict FDR control
- Can even let analyst look at (masked version of) data to choose $Z$ function

Bayesian approach

- $Z_j$ 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
Adaptivity and Prior Information in $W_j$

Recall LCD: $W_j = |\beta_j| - |\tilde{\beta}_j|$, where $\beta_j$, $\tilde{\beta}_j$ come from $\ell_1$-penalized regression

Adaptivity

- Cross-validation (on $[X \tilde{X}]$) to choose the penalty parameter in LCD
- Higher-level adaptivity: CV to choose best-fitting model for inference
  - E.g., fit random forest and $\ell_1$-penalized regression; derive feature importance from whichever has lower CV error—still strict FDR control
- Can even let analyst look at (masked version of) data to choose $Z$ function

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 $W_j$

Recall LCD: $W_j = |\beta_j| - |\tilde{\beta}_j|$, where $\beta_j$, $\tilde{\beta}_j$ come from $\ell_1$-penalized regression

Adaptivity

- Cross-validation (on $[X \ X\tilde{X}]$) to choose the penalty parameter in LCD
- Higher-level adaptivity: CV to choose best-fitting model for inference
  - E.g., fit random forest and $\ell_1$-penalized regression; derive feature importance from whichever has lower CV error—still strict FDR control
- Can even let analyst look at (masked version of) data to choose $Z$ function

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
Step (3): Find the Knockoff Threshold
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$: 

$W_1 \mid W_2 \mid W_3 \mid W_4 \mid W_5 \mid W_6 \mid W_7 \mid W_8 \mid W_9 \mid W_{10}$

$q = 20\% \ \{\text{negative } W_j \} \ \{\text{positive } W_j \}$

$S^\hat{\tau} = \{1, 4, 5, 6, 7\}$
Find the Knockoff Threshold

Example with \( p = 10 \) and \( q = 20\% = 1/5 \):

\[
\hat{\tau} = \{1, 4, 5, 6, 7\}
\]
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$|W_9| \quad |W_2| \quad |W_7| \quad |W_{10}| \quad |W_6|$

$|W_8| \quad |W_3| \quad |W_1| \quad |W_4| \quad |W_5|$
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

\[
\begin{array}{cccc}
|W_9| & |W_2| & |W_7| & |W_{10}| \\
|W_8| & |W_3| & |W_1| & |W_4| \\
\end{array}
\]

\[
S = \{1, 4, 5, 6, 7\}
\]

Lucas Janson (Harvard Statistics)

Knockoffs for HD Controlled Variable Selection
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$$\hat{\tau} = \{1, 4, 5, 6, 7\}$$

Lucas Janson (Harvard Statistics)
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$$\hat{\tau}_S = \{1, 4, 5, 6, 7\}$$

Lucas Janson (Harvard Statistics)
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$\hat{\tau}_S = \{1, 4, 5, 6, 7\}$

$\frac{\# \{\text{negative } W_j\}}{\# \{\text{positive } W_j\}}$

$q = 20\%$
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$\hat{\tau}_S = \{1, 4, 5, 6, 7\}$
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

\[
\begin{array}{cccccc}
|W_1| & |W_2| & |W_3| & |W_4| & |W_5| & |W_6| \\
\hline
0 & 1/4 & 1/3 & 0 & 0 & 2/5
\end{array}
\]

\[
\frac{\# \{\text{negative } W_j} \} \}{\# \{\text{positive } W_j} \} = q = 20\%
\]
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$\|W_9\| \quad \|W_2\| \quad \|W_7\| \quad \|W_{10}\| \quad \|W_6\|$

$\|W_8\| \quad \|W_3\| \quad \|W_1\| \quad \|W_4\| \quad \|W_5\|$

$\frac{1}{5} \quad \frac{1}{4} \quad \frac{1}{3} \quad \frac{0}{3} \quad \frac{0}{2} \quad \frac{0}{1}$

$\frac{\#\{\text{negative } W_j\}}{\#\{\text{positive } W_j\}}$

$q = 20\%$
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

\[
|W_9| \quad |W_2| \quad |W_7| \quad |W_{10}| \quad |W_6|
\]

\[
\begin{array}{cccc}
|W_8| & |W_3| & |W_1| & |W_4| & |W_5| \\
\frac{2}{5} & \frac{1}{5} & \frac{1}{4} & \frac{1}{3} & \frac{0}{3} \\
0 & 0 & 2 & 0 & 1
\end{array}
\]

\[
\frac{\#\{\text{negative } W_j\}}{\#\{\text{positive } W_j\}}
\]

$q = 20\%$
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

\[
\begin{array}{cccccc}
|W_9| & |W_2| & |W_7| & |W_{10}| & |W_6| \\
\frac{3}{5} & \frac{2}{5} & \frac{1}{5} & & \\
|W_8| & |W_3| & |W_1| & |W_4| & |W_5| \\
\frac{1}{4} & \frac{1}{3} & \frac{0}{3} & & \frac{0}{2} & \frac{0}{1}
\end{array}
\]

$\hat{\tau}_S = \{1, 4, 5, 6, 7\}$

$\frac{\#\{\text{negative } W_j\}}{\#\{\text{positive } W_j\}}$

$q = 20\%$
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$|W_9| \quad |W_2| \quad |W_7| \quad |W_{10}| \quad |W_6|$

$|W_8| \quad |W_3| \quad |W_1| \quad |W_4| \quad |W_5|$

$\frac{3}{6} \quad \frac{3}{5} \quad \frac{2}{5} \quad \frac{1}{5} \quad \frac{1}{4} \quad \frac{1}{3} \quad \frac{0}{3} \quad \frac{0}{2} \quad \frac{0}{1}$

$q = 20\%$
Find the Knockoff Threshold

Example with \( p = 10 \) and \( q = 20\% = 1/5 \):

\[ S = \{1, 4, 5, 6, 7\} \]

Lucas Janson (Harvard Statistics)
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$|W_9| \quad |W_2| \quad |W_7| \quad |W_{10}| \quad |W_6|

0

$|W_8| \quad |W_3| \quad |W_1| \quad |W_4| \quad |W_5|

$\frac{3}{7} \quad \frac{3}{6} \quad \frac{3}{5} \quad \frac{2}{5} \quad \frac{1}{5}

$\frac{1}{4} \quad \frac{1}{3} \quad \frac{0}{3}

$\frac{0}{2} \quad \frac{0}{1}$

$q = 20\%$
Find the Knockoff Threshold

Example with $p = 10$ and $q = 20\% = 1/5$:

$S = \{1, 4, 5, 6, 7\}$
FDR = \mathbb{E} \left( \frac{\# \{ \text{null } X_j \text{ selected} \}}{\# \{ \text{total } X_j \text{ selected} \}} \right)
Intuition for FDR Control

FDR = \[\mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{\#\{\text{total } X_j \text{ selected}\}} \right)\]

= \[\mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right)\]
Intuition for FDR Control

\[
\text{FDR} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected} \}}{\#\{\text{total } X_j \text{ selected} \}} \right) \\
= \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau} \}}{\#\{\text{positive } |W_j| > \hat{\tau} \}} \right) \\
\approx \mathbb{E} \left( \frac{\#\{\text{null negative } |W_j| > \hat{\tau} \}}{\#\{\text{positive } |W_j| > \hat{\tau} \}} \right)
\]
Intuition for FDR Control

\[
\text{FDR} = \mathbb{E} \left( \frac{\# \{ \text{null } X_j \text { selected} \}}{\# \{ \text{total } X_j \text { selected} \}} \right) \\
= \mathbb{E} \left( \frac{\# \{ \text{null positive } |W_j| > \hat{\tau} \}}{\# \{ \text{positive } |W_j| > \hat{\tau} \}} \right) \\
\approx \mathbb{E} \left( \frac{\# \{ \text{null negative } |W_j| > \hat{\tau} \}}{\# \{ \text{positive } |W_j| > \hat{\tau} \}} \right) \\
\leq \mathbb{E} \left( \frac{\# \{ \text{negative } |W_j| > \hat{\tau} \}}{\# \{ \text{positive } |W_j| > \hat{\tau} \}} \right)
\]
GWAS Application
2007 case-control study by WTCCC

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

- $n \approx 5,000$, $p \approx 375,000$; preprocessing mirrored original analysis

- **Strong spatial structure**: second-order knockoffs generated using genetic covariance estimate (Wen and Stephens, 2010)
2007 case-control study by WTCCC

- $n \approx 5,000$, $p \approx 375,000$; preprocessing mirrored original analysis

- **Strong spatial structure**: second-order knockoffs generated using genetic covariance estimate (Wen and Stephens, 2010)

- Entire analysis took 6 hours of serial computation time; **1 hour** in parallel
Genetic Analysis of Crohn’s Disease

2007 case-control study by WTCCC

- \( n \approx 5,000 \), \( p \approx 375,000 \); preprocessing mirrored original analysis

- **Strong spatial structure**: second-order knockoffs generated using genetic covariance estimate (Wen and Stephens, 2010)

- Entire analysis took 6 hours of serial computation time; **1 hour** in parallel

- Knockoffs made **twice as many discoveries** as original analysis
2007 case-control study by WTCCC

- $n \approx 5,000$, $p \approx 375,000$; preprocessing mirrored original analysis

- **Strong spatial structure**: second-order knockoffs generated using genetic covariance estimate (Wen and Stephens, 2010)

- Entire analysis took 6 hours of serial computation time; 1 hour in parallel

- Knockoffs made **twice as many discoveries** as original analysis
  - Some new discoveries **confirmed** in larger study
2007 case-control study by WTCCC

- $n \approx 5,000$, $p \approx 375,000$; preprocessing mirrored original analysis

- **Strong spatial structure**: second-order knockoffs generated using genetic covariance estimate (Wen and Stephens, 2010)

- Entire analysis took 6 hours of serial computation time; 1 hour in parallel

- Knockoffs made **twice as many discoveries** as original analysis
  - Some new discoveries **confirmed** in larger study
  - Some corroborated by work on nearby genes: **promising candidates**
Genetic Analysis of Crohn’s Disease

2007 case-control study by WTCCC

- $n \approx 5,000$, $p \approx 375,000$; preprocessing mirrored original analysis
- **Strong spatial structure**: second-order knockoffs generated using genetic covariance estimate (Wen and Stephens, 2010)
- Entire analysis took 6 hours of serial computation time; 1 hour in parallel
- Knockoffs made **twice as many discoveries** as original analysis
  - Some new discoveries confirmed in larger study
  - Some corroborated by work on nearby genes: promising candidates
  - Similar result when HMM knockoffs applied to same data (Sesia et al., 2017)
Discussion
By conditioning on $Y$ and modeling $X$, knockoffs can be applied to high-dimensional and nonlinear problems, where it is powerful, flexible, and appears robust.

Some future directions for research:

Theoretical: rigorous guarantees on robustness

Methodological: develop knockoff constructions for new $X$ distributions

Applied: team up with domain experts who know/control their $X$, e.g., gene knockout/knockdown, climate change modeling

Thank you!
Summary and Next Steps

By conditioning on \( Y \) and modeling \( X \), knockoffs can be applied to high-dimensional and nonlinear problems, where it is powerful, flexible, and appears robust.

Some future directions for research:

- **Theoretical**: rigorous guarantees on robustness
Summary and Next Steps

By conditioning on $Y$ and modeling $X$, knockoffs can be applied to high-dimensional and nonlinear problems, where it is powerful, flexible, and appears robust.

Some future directions for research:

- **Theoretical**: rigorous guarantees on robustness
- **Methodological**: develop knockoff constructions for new $X$ distributions

Thank you!

Lucas Janson (Harvard Statistics)
Summary and Next Steps

By conditioning on $Y$ and modeling $X$, knockoffs can be applied to high-dimensional and nonlinear problems, where it is powerful, flexible, and appears robust.

Some future directions for research:

- **Theoretical**: rigorous guarantees on robustness
- **Methodological**: develop knockoff constructions for new $X$ distributions
- **Applied**: team up with domain experts who know/control their $X$, e.g., gene knockout/knockdown, climate change modeling
By conditioning on $Y$ and modeling $X$, knockoffs can be applied to high-dimensional and nonlinear problems, where it is powerful, flexible, and appears robust.

Some future directions for research:

- **Theoretical**: rigorous guarantees on robustness
- **Methodological**: develop knockoff constructions for new $X$ distributions
- **Applied**: team up with domain experts who know/control their $X$, e.g., gene knockout/knockdown, climate change modeling

Thank you!
Appendix


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.
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.
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.
Checking Sensitivity to Misspecification Error

Concern about misspecification

\[
\begin{array}{c|c|c}
Y \mid X & X \\
\hline
\text{Yes} & \text{No} \\
\text{No} & \text{Yes} \\
\end{array}
\]

Canonical (model $Y$, not $X$)

model $X$, not $Y$
Checking Sensitivity to Misspecification Error

<table>
<thead>
<tr>
<th>Concern about misspecification</th>
<th>( Y \mid X )</th>
<th>( X )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical (model ( Y ), not ( X ))</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>model ( X ), not ( Y )</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Misspecification replicated in simulation?  
- No
- Yes
Checking Sensitivity to Misspecification Error

Concern about misspecification

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>X</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Canonical (model Y, not X)

model X, not Y

Misspecification replicated in simulation?

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Can actually check sensitivity to misspecification error!
Figure: Power and FDR (target is 10%) for model-free knockoffs applied to subsamples of a chromosome 1 of real genetic design matrix; $n \approx 1,400$. 
Computation of Second-Order Knockoffs

\[ \text{Cov}(X_1, \ldots, X_p) = \Sigma, \text{ need:} \]

\[ \text{Cov}(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p) = \begin{bmatrix} \Sigma & \Sigma - \text{diag}\{s\} \\ \Sigma - \text{diag}\{s\} & \Sigma \end{bmatrix} \]
Computation of Second-Order Knockoffs

\[ \text{Cov}(X_1, \ldots, X_p) = \Sigma, \text{ need:} \]

\[
\text{Cov}(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p) = \begin{bmatrix}
\Sigma & \Sigma - \text{diag}\{s\} \\
\Sigma - \text{diag}\{s\} & \Sigma
\end{bmatrix}
\]

- **Equicorrelated (EQ)** (fast, less powerful): \( s_{j}^{\text{EQ}} = 2\lambda_{\text{min}}(\Sigma) \wedge 1 \) for all \( j \)
Computation of Second-Order Knockoffs

\[ \text{Cov}(X_1, \ldots, X_p) = \Sigma, \text{ need:} \]

\[ \text{Cov}(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p) = \begin{bmatrix} \Sigma & \Sigma - \text{diag}\{s\} \\ \Sigma - \text{diag}\{s\} & \Sigma \end{bmatrix} \]

- **Equicorrelated (EQ)** *(fast, less powerful)*: \( s_{\text{EQ}}^j = 2\lambda_{\text{min}}(\Sigma) \wedge 1 \) for all \( j \)

- **Semidefinite program (SDP)** *(slower, more powerful)*:

  \[
  \begin{align*}
  \text{minimize} & \quad \sum_j |1 - s_{\text{SDP}}^j| \\
  \text{subject to} & \quad s_{\text{SDP}}^j \geq 0 \\
  & \quad \text{diag}\{s_{\text{SDP}}\} \preceq 2\Sigma,
  \end{align*}
\]
Computation of Second-Order Knockoffs

\[
\text{Cov}(X_1, \ldots, X_p) = \Sigma, \text{ need:}
\]

\[
\text{Cov}(X_1, \ldots, X_p, \tilde{X}_1, \ldots, \tilde{X}_p) = \begin{bmatrix}
\Sigma & \Sigma - \text{diag}\{s\} \\
\Sigma - \text{diag}\{s\} & \Sigma
\end{bmatrix}
\]

- **Equicorrelated (EQ)** (fast, less powerful): \( s_j^{\text{EQ}} = 2\lambda_{\text{min}}(\Sigma) \wedge 1 \) for all \( j \)

- **Semidefinite program (SDP)** (slower, more powerful):

  minimize \( \sum_j |1 - s_j^{\text{SDP}}| \)
  
  subject to \( s_j^{\text{SDP}} \geq 0 \)
  
  \( \text{diag}\{s^{\text{SDP}}\} \preceq 2\Sigma \),

- **(New) Approximate SDP**: 
  - Approximate \( \Sigma \) as block diagonal so that SDP separates
  - Bisection search scalar multiplier of solution to account for approximation
  - faster than SDP, more powerful than EQ, and easily parallelizable
Algorithm 1 Sequential Conditional Independent Pairs

\textbf{for} \ j = \{1, \ldots, p\} \ \textbf{do} \\
\hspace{1em} \text{Sample } \tilde{X}_j \text{ from } \mathcal{L}(X_j | X_{-j}, \tilde{X}_{1:j-1}) \text{ conditionally independently of } X_j \\
\textbf{end}

Proof sketch (discrete case): Denote PMF of \((X_1:p, \tilde{X}_1:j-1)\) by \(\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})\). Conditional PMF of \(\tilde{X}_j | X_1:p, \tilde{X}_{1:j-1}\) is \(\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})/\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})\). Joint PMF of \((X_1:p, \tilde{X}_1:j)\) is \(\mathcal{L}(X_j | X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1}) \mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1}) / \sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})\).
Algorithm 1 Sequential Conditional Independent Pairs

for $j = \{1, \ldots, p\}$ do
    Sample $\tilde{X}_j$ from $\mathcal{L}(X_j \mid X_{-j}, \tilde{X}_{1:j-1})$ conditionally independently of $X_j$
end

Proof sketch (discrete case):
- Denote PMF of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})$
Algorithm 1 Sequential Conditional Independent Pairs

for $j = \{1, \ldots, p\}$ do
  Sample $\tilde{X}_j$ from $\mathcal{L}(X_j \mid X_{-j}, \tilde{X}_{1:j-1})$ conditionally independently of $X_j$
end

Proof sketch (discrete case):

- Denote PMF of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})$
- Conditional PMF of $\tilde{X}_j \mid X_{1:p}, \tilde{X}_{1:j-1}$ is

$$\frac{\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}.$$
Algorithm 1 Sequential Conditional Independent Pairs

for $j = \{1, \ldots, p\}$ do
  Sample $\tilde{X}_j$ from $\mathcal{L}(X_j \mid X_{-j}, \tilde{X}_{1:j-1})$ conditionally independently of $X_j$
end

Proof sketch (discrete case):

- Denote PMF of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})$
- Conditional PMF of $\tilde{X}_j \mid X_{1:p}, \tilde{X}_{1:j-1}$ is
  \[
  \frac{\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}.
  \]
- Joint PMF of $(X_{1:p}, \tilde{X}_{1:j})$ is
  \[
  \frac{\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1}) \mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}.
  \]
Algorithm 1 Sequential Conditional Independent Pairs

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

Proof sketch (discrete case):

- Denote PMF of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $L(X_{-j}, X_j, \tilde{X}_{1:j-1})$
- Conditional PMF of $\tilde{X}_j | X_{1:p}, \tilde{X}_{1:j-1}$ is
  \[
  \frac{L(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u L(X_{-j}, u, \tilde{X}_{1:j-1})}.
  \]
- Joint PMF of $(X_{1:p}, \tilde{X}_{1:j})$ is
  \[
  \frac{L(X_{-j}, X_j, \tilde{X}_{1:j-1})L(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u L(X_{-j}, u, \tilde{X}_{1:j-1})}.
  \]
Algorithm 1 Sequential Conditional Independent Pairs

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

Proof sketch (discrete case):

- Denote PMF of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})$
- Conditional PMF of $\tilde{X}_j | X_{1:p}, \tilde{X}_{1:j-1}$ is
  \[
  \frac{\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}.
  \]
- Joint PMF of $(X_{1:p}, \tilde{X}_{1:j})$ is
  \[
  \frac{\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1}) \mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}.
  \]
Algorithm 1 Sequential Conditional Independent Pairs

\begin{algorithm}
\begin{algorithmic}
  \FOR{$j = \{1, \ldots, p\}$}
  \STATE Sample $\tilde{X}_j$ from $\mathcal{L}(X_j \mid X_{-j}, \tilde{X}_{1:j-1})$ conditionally independently of $X_j$
  \ENDFOR
\end{algorithmic}
\end{algorithm}

Proof sketch (discrete case):

- Denote PMF of $(X_{1:p}, \tilde{X}_{1:j-1})$ by $\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})$
- Conditional PMF of $\tilde{X}_j \mid X_{1:p}, \tilde{X}_{1:j-1}$ is
  \[ \frac{\mathcal{L}(X_{-j}, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}. \]
- Joint PMF of $(X_{1:p}, \tilde{X}_{1:j})$ is
  \[ \frac{\mathcal{L}(X_{-j}, X_j, \tilde{X}_{1:j-1})\mathcal{L}(X_j, \tilde{X}_j, \tilde{X}_{1:j-1})}{\sum_u \mathcal{L}(X_{-j}, u, \tilde{X}_{1:j-1})}. \]
Proof of Control

\[ \text{FDR} = \mathbb{E} \left( \frac{\# \{ \text{null } X_j \text{ selected} \}}{\# \{ \text{total } X_j \text{ selected} \}} \right) \]
Proof of Control

\[
\text{FDR} = \mathbb{E}\left( \frac{\#\{\text{null } X_j \text{ selected}\}}{\#\{\text{total } X_j \text{ selected}\}} \right)
\]

\[
= \mathbb{E}\left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right)
\]

More precisely:

\[
\hat{m}_{\text{FDR}} = \mathbb{E}\left( \frac{\#\{\text{null } X_j \text{ selected}\}}{q - 1} + \#\{\text{total } X_j \text{ selected}\} \right)
\]

\[
= \mathbb{E}\left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{q - 1} + \#\{\text{positive } |W_j| > \hat{\tau}\} \right)
\]

\[
\approx \mathbb{E}\left( \frac{\#\{\text{null negative } |W_j| > \hat{\tau}\}}{q - 1} + \#\{\text{positive } |W_j| > \hat{\tau}\} \right)
\]

\[
\leq \mathbb{E}\left( \frac{\#\{\text{negative } |W_j| > \hat{\tau}\}}{q - 1} + \#\{\text{positive } |W_j| > \hat{\tau}\} \right)
\]

\[
\text{Supermartingale} \leq 1 \text{ with } \hat{\tau} \text{ a stopping time}
\]
Proof of Control

\[ \text{FDR} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{\#\{\text{total } X_j \text{ selected}\}} \right) \]

\[ = \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

\[ \approx \mathbb{E} \left( \frac{\#\{\text{null negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

More precisely:

\[ m_{\text{FDR}} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{q - 1} + \#\{\text{total } X_j \text{ selected}\} \right) \]

\[ = \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{q - 1} + \#\{\text{positive } |W_j| > \hat{\tau}\} \right) \]

\[ \approx \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{1} + \#\{\text{null negative } |W_j| > \hat{\tau}\} \right) \]

\[ \leq 1 \]

with \( \hat{\tau} \) a stopping time.
Proof of Control

\[
\text{FDR} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{\#\{\text{total } X_j \text{ selected}\}} \right)
\]

\[
= \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right)
\]

\[
\approx \mathbb{E} \left( \frac{\#\{\text{null negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right)
\]

\[
\leq \mathbb{E} \left( \frac{\#\{\text{negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right)
\]
Proof of Control

\[ \text{FDR} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{\#\{\text{total } X_j \text{ selected}\}} \right) \]

\[ = \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

\[ \approx \mathbb{E} \left( \frac{\#\{\text{null negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

\[ \leq \mathbb{E} \left( \frac{\#\{\text{negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

More precisely:

\[ \text{mFDR} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{q^{-1} + \#\{\text{total } X_j \text{ selected}\}} \right) = \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{q^{-1} + \#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]
Proof of Control

\[
\text{FDR} = \mathbb{E} \left( \frac{\# \{ \text{null } X_j \text{ selected} \}}{\# \{ \text{total } X_j \text{ selected} \}} \right) \\
= \mathbb{E} \left( \frac{\# \{ \text{null positive } |W_j| > \hat{\tau} \}}{\# \{ \text{positive } |W_j| > \hat{\tau} \}} \right) \\
\approx \mathbb{E} \left( \frac{\# \{ \text{null negative } |W_j| > \hat{\tau} \}}{\# \{ \text{positive } |W_j| > \hat{\tau} \}} \right) \\
\leq \mathbb{E} \left( \frac{\# \{ \text{negative } |W_j| > \hat{\tau} \}}{\# \{ \text{positive } |W_j| > \hat{\tau} \}} \right)
\]

More precisely:

\[
\text{mFDR} = \mathbb{E} \left( \frac{\# \{ \text{null } X_j \text{ selected} \}}{q^{-1} + \# \{ \text{total } X_j \text{ selected} \}} \right) = \mathbb{E} \left( \frac{\# \{ \text{null positive } |W_j| > \hat{\tau} \}}{q^{-1} + \# \{ \text{positive } |W_j| > \hat{\tau} \}} \right) \\
= \mathbb{E} \left( \frac{\# \{ \text{null positive } |W_j| > \hat{\tau} \}}{1 + \# \{ \text{null negative } |W_j| > \hat{\tau} \}} \cdot \frac{1 + \# \{ \text{null negative } |W_j| > \hat{\tau} \}}{q^{-1} + \# \{ \text{positive } |W_j| > \hat{\tau} \}} \right)
\]
Proof of Control

FDR = \[ E \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{\#\{\text{total } X_j \text{ selected}\}} \right) \]

= \[ E \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

\approx \[ E \left( \frac{\#\{\text{null negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

\leq \[ E \left( \frac{\#\{\text{negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

More precisely:

mFDR = \[ E \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{q^{-1} + \#\{\text{total } X_j \text{ selected}\}} \right) \]

= \[ E \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{q^{-1} + \#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

= \[ E \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{1 + \#\{\text{null negative } |W_j| > \hat{\tau}\}} \cdot \frac{1 + \#\{\text{null negative } |W_j| > \hat{\tau}\}}{q^{-1} + \#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \]

\leq q \text{ by definition of } \hat{\tau}
Proof of Control

\[
\text{FDR} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{\#\{\text{total } X_j \text{ selected}\}} \right) \\
= \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \\
\cong \mathbb{E} \left( \frac{\#\{\text{null negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \\
\leq \mathbb{E} \left( \frac{\#\{\text{negative } |W_j| > \hat{\tau}\}}{\#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \\
\leq \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{q^{-1} + \#\{\text{total } X_j \text{ selected}\}} \right) \\
= \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{q^{-1} + \#\{\text{positive } |W_j| > \hat{\tau}\}} \right) \\
\leq q \text{ by definition of } \hat{\tau}
\]

More precisely:

\[
\text{mFDR} = \mathbb{E} \left( \frac{\#\{\text{null } X_j \text{ selected}\}}{q^{-1} + \#\{\text{total } X_j \text{ selected}\}} \right) \\
= \mathbb{E} \left( \frac{\#\{\text{null positive } |W_j| > \hat{\tau}\}}{1 + \#\{\text{null negative } |W_j| > \hat{\tau}\}} \cdot \frac{1 + \#\{\text{null negative } |W_j| > \hat{\tau}\}}{q^{-1} + \#\{\text{positive } |W_j| > \hat{\tau}\}} \right)
\]

Supermartingale \leq 1 with \( \hat{\tau} \) a stopping time

Lucas Janson (Harvard Statistics)
Knockoffs for HD Controlled Variable Selection
18 / 18