Controlling false discovery rate via knockoffs

Rina Foygel Barber & Emmanuel Candès

Jan 21 2015

Setting

An example:
Which mutations in the reverse transcriptase (RT) of HIV-1 determine susceptibility to reverse transcriptase inhibitors (RTIs)?

- $y_i \in \mathbb{R}$ = resistance of virus in sample $i$ to a RTI-type drug
- $X_{ij} \in \{0, 1\}$ indicates if mutation $j$ is present in virus sample $i$

How can we select mutations that determine drug resistance, in such a way that our answer will replicate in further trials?
Setting

Sparse linear model:

\[ y = X \cdot \beta + z, \text{ where } z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \]

- \( n \) observations, \( p \) features
- \( \beta \) is sparse
Goal: select a set of features $X_j$ that are likely to be relevant to the response $y$, without too many false positives.

One way to measure performance:

$$\text{FDR} = \mathbb{E} \left[ \frac{\text{# false positives}}{\text{total # of features selected}} \right] = \mathbb{E} \left[ \frac{|S \cap H_0|}{|S|} \right].$$

False discovery rate \quad False discovery proportion

\[ S = \text{set of selected features} \]

\[ H_0 = \text{“null hypotheses”} = \{j : \beta_j^* = 0\} \]
Sparse regression

Lasso: \( \beta_\lambda = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|y - X \cdot \beta\|^2_2 + \lambda \|\beta\|_1 \right\} \)

Asymptotically, Lasso will select the correct model (at a good \( \lambda \)).

In practice for a finite sample,

- True positives & false positives intermixed along the Lasso path
- How to pick \( \lambda \) to balance FDR vs power?
- Need to account for correlations between \( X_j \) & weak signals that may have been missed on the Lasso path.
Sparse regression

Simulated data with \( n = 1500, p = 500 \).

Lasso fitted model for \( \lambda = 1.75 \):
Sparse regression

Simulated data with $n = 1500, p = 500$.

Lasso fitted model for $\lambda = 1.75$:

![Graph showing fitted coefficients](image)
Sparse regression

Simulated data with $n = 1500, p = 500$.

Lasso fitted model for $\lambda = 1.75$:

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

Simulated data with $n = 1500, p = 500$.

Lasso fitted model for $\lambda = 1.75$:

\[
\text{FDP} = \frac{26}{55} = 47\%
\]

To estimate FDP, would need to calculate distribution of $\beta_j^\lambda$ for null $j$
(would need to know $\sigma^2, \beta^*, \ldots$).  (Donoho et al 2009)
Construct knockoffs

Main idea:

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

Setting:

- Require $n > p$ (ongoing work for high-dim. setting)
- Don’t need to know $\sigma^2$
- Don’t need any information about $\beta^*$
- Will get an exact, finite-sample guarantee for FDR
Construct knockoffs

Construction:

- The knockoffs replicate the correlation structure of $X$:
  \[ \tilde{X}_j \tilde{X}_k = X_j X_k \text{ for all } j, k \]

- Also preserve correlations between knockoffs & originals:
  \[ \tilde{X}_j X_k = X_j X_k \text{ for all } j \neq k \]

Augmented design matrix

\[
\begin{bmatrix}
X & \tilde{X}
\end{bmatrix} =
\begin{bmatrix}
X_1 & X_2 & \ldots & X_p & \tilde{X}_1 & \tilde{X}_2 & \ldots & \tilde{X}_p
\end{bmatrix} \in \mathbb{R}^{n \times 2p}
\]
Construct knockoffs

How?

Define $\tilde{X} = X \cdot (I_p - 2\xi\Sigma^{-1}) + U \cdot C$, where:

\[
\Sigma = X^\top X \succeq \xi I_p
\]

$U = n \times p$ orthonormal matrix orthogonal to $X$

$C^\top C = 4(\xi I_p - \xi^2 \Sigma^{-1})$ (Cholesky decomposition)

\[
\implies [X \ \tilde{X}]^\top [X \ \tilde{X}] = \begin{pmatrix}
\Sigma & \Sigma - 2\xi I_p \\
\Sigma - 2\xi I_p & \Sigma
\end{pmatrix}
\]
Construct knockoffs

Why?

For a null feature $X_j$,

$$X_j^\top y = X_j^\top X\beta^* + X_j^\top z \overset{D}{=} \tilde{X}_j^\top X\beta^* + \tilde{X}_j^\top z = \tilde{X}_j^\top y$$
Construct knockoffs

Why?

For a null feature $X_j$,

$$X_j^\top y = X_j^\top X\beta^* + X_j^\top z \overset{\mathcal{D}}{=} \tilde{X}_j^\top X\beta^* + \tilde{X}_j^\top z = \tilde{X}_j^\top y$$
Lemma 1: Pairwise exchangeability property.
For any $N \subset \mathcal{H}_0$,

$$
\left( \begin{bmatrix} X & \tilde{X} \end{bmatrix}_{\text{swap}(N)} \right) ^\top y \overset{\mathcal{D}}{=} \begin{bmatrix} X & \tilde{X} \end{bmatrix} ^\top y
$$

$\Rightarrow$ the knockoffs are a “control group” for the nulls
Knockoff method

Steps:

1. Construct knockoffs

2. Compute Lasso with augmented matrix:

\[
\beta_\lambda = \arg \min_{\beta \in \mathbb{R}^{2p}} \left\{ \frac{1}{2} \left\| y - [X \ 	ilde{X}] \cdot \beta \right\|^2_2 + \lambda \left\| \beta \right\|_1 \right\}
\]

3. Use \( \tilde{X}_j \) as a “control group” for \( X_j \)
Knockoff method

Fitted model for $\lambda = 1.75$ on the simulated dataset:

- Lasso selects 49 original features & 24 knockoff features
Knockoff method

Fitted model for $\lambda = 1.75$ on the simulated dataset:

- Lasso selects 49 original features & 24 knockoff features
- Pairwise exchangeability of the nulls
  $\implies$ probably $\approx 24$ false positives among the 49 original features
Knockoff method

Compute Lasso on the entire path $\lambda \in [0, \infty)$.

$$\lambda_j = \sup \left\{ \lambda : \beta_j^\lambda \neq 0 \right\} = \text{first time } X_j \text{ enters Lasso path}$$

$$\tilde{\lambda}_j = \sup \left\{ \lambda : \tilde{\beta}_j^\lambda \neq 0 \right\} = \text{first time } \tilde{X}_j \text{ enters Lasso path}$$

Then define statistics

$$W_j = \max\{\lambda_j, \tilde{\lambda}_j\} \cdot \text{sign}(\lambda_j - \tilde{\lambda}_j)$$
Knockoff method

\[ \lambda = 0 \]
variables enter late
(probably not significant)

\[ \lambda \rightarrow \infty \]
variables enter early
(likely significant)
Knockoff method

\( \lambda \) when \( X_j \) enters

\( \lambda \) when \( X_j \) enters

- Null
- Signal
Lemma 2: Pairwise exchangeability of the nulls \[ (W_1, W_2, \ldots, W_p) \overset{D}{=} (|W_1| \cdot \epsilon_1, |W_2| \cdot \epsilon_2, \ldots, |W_p| \cdot \epsilon_p) \]

where \( \epsilon_j = \text{sign}(W_j) \) for non-nulls and \( \epsilon_j \overset{iid}{\sim} \{\pm 1\} \) for nulls.
Knockoff method

Selected variables: \( S_\lambda = \{ j : W_j \geq +\lambda \} \)

Control group: \( \tilde{S}_\lambda = \{ j : W_j \leq -\lambda \} \)

\[ \hat{\text{FDP}}(S_\lambda) := \frac{|\tilde{S}_\lambda|}{|S_\lambda|} \]
Knockoff method

Selected variables: \( S_{\lambda} = \{ j : W_j \geq +\lambda \} \)

Control group: \( \tilde{S}_{\lambda} = \{ j : W_j \leq -\lambda \} \)

\( \Rightarrow \quad \hat{\text{FDP}}(S_{\lambda}) := \frac{|\tilde{S}_{\lambda}|}{|S_{\lambda}|} \)

\[
\text{FDP}(S_{\lambda}) = \frac{|S_{\lambda} \cap \mathcal{H}_0|}{|S_{\lambda}|} \approx \frac{|\tilde{S}_{\lambda} \cap \mathcal{H}_0|}{|S_{\lambda}|} \leq \hat{\text{FDP}}(S_{\lambda})
\]
The knockoff filter: define

\[
\hat{\text{FDP}}(S_\lambda) := \frac{|\tilde{S}_\lambda|}{|S_\lambda|} = \frac{\# \{ j : W_j \leq -\lambda \}}{\# \{ j : W_j \geq +\lambda \}},
\]

then choose

\[
\Lambda = \min \left\{ \lambda : \hat{\text{FDP}}(S_\lambda) \leq q \right\} \quad \text{(or } \lambda = \infty \text{ if empty set)}
\]

and select the variable set

\[
S_\Lambda = \{ j : W_j \geq \Lambda \}.
\]
Theorem 1: For $S_\Lambda$ chosen by the knockoff filter,

$$\mathbb{E}[\text{mFDP}(S_\Lambda)] \leq q$$

where the modified FDP is given by

$$\text{mFDP}(S) = \frac{|S \cap \mathcal{H}_0|}{|S| + q^{-1}}.$$
The theoretical guarantees

The knockoff+ filter: define

$$\widehat{\text{FDP}}_+(S_\lambda) := \frac{|\tilde{S}_\lambda| + 1}{|S_\lambda|} = \frac{\# \{ j : W_j \leq -\lambda \} + 1}{\# \{ j : W_j \geq +\lambda \}},$$

then choose

$$\Lambda_+ = \min \left\{ \lambda : \widehat{\text{FDP}}_+(S_\lambda) \leq q \right\} \quad \text{(or } \lambda = \infty \text{ if empty set)}$$

and select the variable set

$$S_{\Lambda_+} = \{ j : W_j \geq \Lambda_+ \}.$$
Theoretical guarantees

Theorem 2: For $S_{\Lambda^+}$ chosen by the knockoff+ filter,

$$
\mathbb{E} \left[ \text{FDP} \left( S_{\Lambda^+} \right) \right] \leq q.
$$
Theoretical guarantees

Theorem 2: For $S_{\Lambda^+}$ chosen by the knockoff+ filter,

$$
\mathbb{E} \left[ \text{FDP}(S_{\Lambda^+}) \right] \leq q.
$$

Proof sketch:

$$
\text{FDP}(S_{\Lambda^+}) = \frac{|S_{\Lambda^+} \cap \mathcal{H}_0|}{|S_{\Lambda^+}|} = \frac{\tilde{S}_{\Lambda^+} \cap \mathcal{H}_0 + 1}{|S_{\Lambda^+}|} \cdot \frac{|S_{\Lambda^+} \cap \mathcal{H}_0|}{\tilde{S}_{\Lambda^+} \cap \mathcal{H}_0 + 1} \leq \text{FDP}_+(S_{\Lambda^+}) \leq q.
$$
Theoretical guarantees

Proof sketch cont’d:

\[
M(\lambda) = \frac{|S_\lambda \cap \mathcal{H}_0|}{|\tilde{S}_\lambda \cap \mathcal{H}_0| + 1}
\]

is a supermartingale w.r.t. increasing \(\lambda\),

and \(\Lambda_+\) is a stopping time.
Theoretical guarantees

Proof sketch cont’d:

\[ M(\lambda) = \frac{|S_{\lambda} \cap H_0|}{|\tilde{S}_{\lambda} \cap H_0| + 1} \]

is a supermartingale w.r.t. increasing \( \lambda \), and \( \Lambda_+ \) is a stopping time.
Theoretical guarantees

Proof sketch cont’d:

\[ M(\lambda) = \frac{|S_{\lambda} \cap \mathcal{H}_0|}{|\tilde{S}_{\lambda} \cap \mathcal{H}_0| + 1} \]

is a supermartingale w.r.t. increasing \( \lambda \), and \( \Lambda_+ \) is a stopping time.

\[ |W| \]

\begin{align*}
+ & + & + & + & + & + & + \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
- & - & - & - & - & - \\
\end{align*}
Theoretical guarantees

Proof sketch cont’d:

\[ M(\lambda) = \frac{|S_\lambda \cap \mathcal{H}_0|}{|\tilde{S}_\lambda \cap \mathcal{H}_0| + 1} \]

is a supermartingale w.r.t. increasing \( \lambda \),

and \( \Lambda_+ \) is a stopping time.

\[
\mathbb{E}[M(\Lambda_+)] \leq \mathbb{E}[M(0)] = \mathbb{E}[C|H_0| - C + 1] \leq 1,
\]

for \( C \) = number of + coin flips ∼ Bin\(|H_0|, 0.5\).

Jan 21 2015 Controlling false discovery rate via knockoffs 26/36
Theoretical guarantees

Proof sketch cont’d:

\[
M(\lambda) = \frac{|S_\lambda \cap \mathcal{H}_0|}{|\tilde{S}_\lambda \cap \mathcal{H}_0| + 1}
\]

is a supermartingale w.r.t. increasing \( \lambda \), and \( \Lambda_+ \) is a stopping time.
Theoretical guarantees

Proof sketch cont’d:

\[ M(\lambda) = \frac{|S_{\lambda} \cap \mathcal{H}_0|}{|\tilde{S}_{\lambda} \cap \mathcal{H}_0| + 1} \]

is a supermartingale w.r.t. increasing \( \lambda \), and \( \Lambda_+ \) is a stopping time.

\[ \mathbb{E} [M(\Lambda_+)] \leq \mathbb{E} [M(0)] = \mathbb{E} \left[ \frac{C}{|\mathcal{H}_0| - C + 1} \right] \leq 1, \]

for \( C = \# \text{ of + coin flips} \sim \text{Bin}(|\mathcal{H}_0|, 0.5) \).
Simulations

Setup:

- \( n = 3000, \ p = 1000, \) sparsity level \( k \)
- Features \( X_j \) are random unit vectors with correlation level \( \rho \)
- For signals \( j, \ \beta^*_j \overset{iid}{\sim} \{\pm A\} \) for amplitude level \( A \)
- \( y = X\beta + N(0, I_n) \)

Compare knockoff, knockoff+, & Benjamini-Hochberg (BH).
Simulations

- Fix amplitude $A = 3.5$ & sparsity level $k = 30$
- Vary feature correlation $\rho$ from 0 to 0.9 (set $\mathbb{E}[X_j^\top X_k] = \rho|j-k|$)
HIV data

Which mutations in the RT or protease of HIV-1 determine susceptibility to RT inhibitors or protease inhibitors?

Data:
Available at [hivdb.stanford.edu](http://hivdb.stanford.edu) (Stanford HIV Drug Resistance Database)

- Each drug analysed separately
- Response $y = \text{resistance to the drug}$
- Features $X = \text{which mutations are present in the RT or in the protease}$
The data set:

<table>
<thead>
<tr>
<th>Drug type</th>
<th># drugs</th>
<th>Sample size</th>
<th># protease or RT positions genotyped</th>
<th># mutations appearing ≥ 3 times in sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>6</td>
<td>848</td>
<td>99</td>
<td>209</td>
</tr>
<tr>
<td>NRTI</td>
<td>6</td>
<td>639</td>
<td>240</td>
<td>294</td>
</tr>
<tr>
<td>NNRTI</td>
<td>3</td>
<td>747</td>
<td>240</td>
<td>319</td>
</tr>
</tbody>
</table>
# HIV data

## The data set:

<table>
<thead>
<tr>
<th>Drug type</th>
<th># drugs</th>
<th>Sample size</th>
<th># protease or RT positions genotyped</th>
<th># mutations appearing $\geq 3$ times in sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>6</td>
<td>848</td>
<td>99</td>
<td>209</td>
</tr>
<tr>
<td>NRTI</td>
<td>6</td>
<td>639</td>
<td>240</td>
<td>294</td>
</tr>
<tr>
<td>NNRTI</td>
<td>3</td>
<td>747</td>
<td>240</td>
<td>319</td>
</tr>
</tbody>
</table>

To validate results:

- **Treatment-selected mutation (TSM) panel:**
  
  A separate study identifies mutations frequently present in patients who have been treated with each type of drug
HIV data

Results for PI type drugs

Resistance to APV

Resistance to ATV

Resistance to IDV

Resistance to LPV

Resistance to NFV

Resistance to SQV

Data set size: n=768, p=201

Data set size: n=329, p=147

Data set size: n=826, p=208

Data set size: n=516, p=184

Data set size: n=843, p=209

Data set size: n=825, p=208

# HIV−1 protease positions selected

Appear in TSM list
Not in TSM list

Jan 21 2015 Controlling false discovery rate via knockoffs 31/36
HIV data

Results for NRTI type drugs

- Resistance to X3TC
  - Knockoff BHq
  - Data set size: n=633, p=292
  - # HIV−1 RT positions selected
  - Resistance to ABC
  - Knockoff BHq
  - Data set size: n=628, p=294
  - Resistance to AZT
  - Knockoff BHq
  - Data set size: n=630, p=292
  - Resistance to D4T
  - Knockoff BHq
  - Data set size: n=630, p=293
  - Resistance to DDI
  - Knockoff BHq
  - Data set size: n=353, p=218
  - Resistance to TDF

Results for NNRTI type drugs

- Resistance to DLV
  - Knockoff BHq
  - Data set size: n=732, p=311
  - Resistance to EFV
  - Knockoff BHq
  - Data set size: n=734, p=318
  - Resistance to NVP
  - Knockoff BHq
  - Data set size: n=746, p=319

Jan 21 2015 Controlling false discovery rate via knockoffs
Can knockoffs be replaced by permutations?

Let $X^\pi = X$ with rows randomly permuted. Then

$$[X \ X^\pi]^\top [X \ X^\pi] \approx \begin{pmatrix} \Sigma & 0 \\ 0 & \Sigma \end{pmatrix}$$
Can knockoffs be replaced by permutations?

Let $X^\pi = X$ with rows randomly permuted. Then

$$[X \ X^\pi]^\top [X \ X^\pi] \approx \begin{pmatrix} \Sigma & 0 \\ 0 & \Sigma \end{pmatrix}$$

<table>
<thead>
<tr>
<th></th>
<th>FDR (target level $q = 20%$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knockoff method</td>
<td>12.29%</td>
</tr>
<tr>
<td>Permutation method</td>
<td>45.61%</td>
</tr>
</tbody>
</table>
The knockoff filter for inference in a sparse linear model:

- Creates a “control group” for any type of statistic
- Handles any type of feature correlation
- Unknown noise level & sparsity level
- Finite-sample FDR guarantees
Summary

Future work:

1. How to move to high-dimensional setting?
2. Extend to GLMs or other regression models?
3. Similar principles for other problems, e.g. graphical models?
Thank you!


- Joint work with Emmanuel Candès @ Stanford
- R. F. B. was partially supported by NSF award DMS-1203762. E. C. is partially supported by AFOSR under grant FA9550-09-1-0643, by NSF via grant CCF-0963835 and by the Math + X Award from the Simons Foundation.