Knockoffs with Side Information

Zhimei Ren

Statistical Methods in Machine Learning
Bernoulli-IMS One World Symposium 2020
August 24th-August 28th, 2020

Collaborator

Emmanuel Candès

Variable Selection

Setting

Explanatory Variables		Response
$(X_1, X_2, \ldots, X_p)$	$\longrightarrow$	Y

Variable Selection

Setting

Explanatory Variables | Response

$$(X_1, X_2, \ldots, X_p) \longrightarrow Y$$

Goal

- Detect the important variables w.r.t. the response.

Variable Selection

Setting

Explanatory Variables — Response

$$(X_1, X_2, \ldots, X_p) \longrightarrow Y$$

Goal

- Detect the important variables w.r.t. the response.
- Control the proportion of the false discoveries.

Variable Selection

Setting

Explanatory Variables | Response

$$(X_1, X_2, \ldots, X_p) \longrightarrow Y$$

Goal

- Detect the important variables w.r.t. the response.
- Control the proportion of the false discoveries.

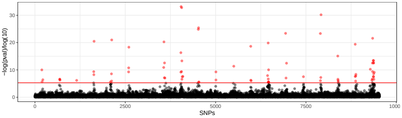


Figure: GWAS

Variable Selection

Setting

Explanatory Variables — Response

$$(X_1, \color{red}{X_2}, \dots, X_p) \longrightarrow Y$$

Goal

- Detect the important variables w.r.t. the response.
- Control the proportion of the false discoveries.

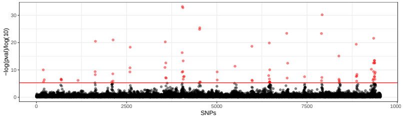


Figure: GWAS

Figure: MRI

Image from K. Pauly, G. Gold, RAD 220

Variable Selection with Side Information

Setting

Explanatory Variables		Response
$(X_1, X_2, \ldots, X_p)$	$\longrightarrow$	Y
$(U_1, U_2, \ldots, U_p)$		

Variable Selection with Side Information

Setting

Explanatory Variables		Response
$(X_1, X_2, \ldots, X_p)$	$\longrightarrow$	Y
$(U_1, U_2, \ldots, U_p)$		

Goal

- Detect the important variables w.r.t. the response with the help of side information.

Variable Selection with Side Information

Setting

Explanatory Variables | Response

$(X_1, X_2, \ldots, X_p) \longrightarrow Y$

$(U_1, U_2, \ldots, U_p)$

Goal

- Detect the important variables w.r.t. the response with the help of side information.
- Control the proportion of the false discoveries conditional on the side information.

Variable Selection with Side Information

Setting

Explanatory Variables — Response

$$(X_1, X_2, \ldots, X_p) \longrightarrow Y$$

$$(U_1, U_2, \ldots, U_p)$$

Goal

- Detect the important variables w.r.t. the response with the help of side information.
- Control the proportion of the false discoveries conditional on the side information.

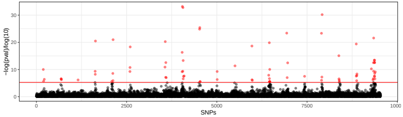


Figure: GWAS

Variable Selection with Side Information

Setting

Explanatory Variables		Response
$(X_1, {\color{red}X_2}, \ldots, X_p)$	$\longrightarrow$	Y
$(U_1, U_2, \ldots, U_p)$		

Goal

- Detect the important variables w.r.t. the response with the help of side information.
- Control the proportion of the false discoveries conditional on the side information.

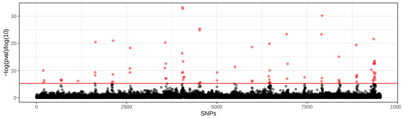


Figure: GWAS

Figure: MRI

Image from K. Pauly, G. Gold, RAD 2020

Problem Formulation

- A variable X_j defined as *null* if the following hypothesis is true:

$$\mathcal{H}_j : X_j \perp\!\!\!\perp Y \mid X_{-j}.$$

Problem Formulation

- A variable X_j defined as *null* if the following hypothesis is true:

$$\mathcal{H}_j : X_j \perp\!\!\!\perp Y \mid X_{-j}.$$

- R: the number of discoveries.

Problem Formulation

- A variable X_j defined as *null* if the following hypothesis is true:

$$\mathcal{H}_j : X_j \perp\!\!\!\perp Y \mid X_{-j}.$$

- R: the number of discoveries.
- V: the number of false discoveries.

Problem Formulation

- A variable X_j defined as *null* if the following hypothesis is true:

$$\mathcal{H}_j : X_j \perp\!\!\!\perp Y \mid X_{-j}.$$

- R: the number of discoveries.
- V: the number of false discoveries.
- Error criterion: *False Discovery Rate*

$$\text{FDR} \triangleq \mathbb{E}\left[\frac{V}{\max(R, 1)}\right].$$

Problem Formulation

- A variable X_j defined as *null* if the following hypothesis is true:

$$\mathcal{H}_j : X_j \perp\!\!\!\perp Y \mid X_{-j}.$$

- R: the number of discoveries.
- V: the number of false discoveries.
- Error criterion: *False Discovery Rate*

$$\text{FDR} \triangleq \mathbb{E}\left[\frac{V}{\max(R, 1)}\right].$$

- Goal: detect as many non-null variables as possible while controlling the FDR below level α.

Variable Selection Procedure: Knockoffs (Review)

- Scientific paradigm:

n i.i.d. observations | Selection set

$$(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i) \xrightarrow{\text{Blackbox algorithm}} S$$

Variable Selection Procedure: Knockoffs (Review)

- Scientific paradigm:

n i.i.d. observations		Selection set
	Blackbox algorithm	
$(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i)$	$\xrightarrow{\hspace{4cm}}$	S

- Knockoffs (Barber et al., 2015; Candès et al., 2018):

Variable Selection Procedure: Knockoffs (Review)

- Scientific paradigm:

  n i.i.d. observations → Selection set

  $$(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i) \xrightarrow{\text{Blackbox algorithm}} S$$

- Knockoffs (Barber et al., 2015; Candès et al., 2018):
  - a wrapper around the blackbox algorithms;

Variable Selection Procedure: Knockoffs (Review)

- Scientific paradigm:

n i.i.d. observations — Blackbox algorithm → Selection set

$$(X_{i1}, X_{i2}, \dots, X_{ip}, Y_i) \xrightarrow{\text{Blackbox algorithm}} S$$

- Knockoffs (Barber et al., 2015; Candès et al., 2018):
  - a wrapper around the blackbox algorithms;
  - produces a seletion set with FDR controlled below the target α.

Variable Selection Procedure: Knockoffs (Review)

- Scientific paradigm:

n i.i.d. observations	Blackbox algorithm	Selection set
$(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i)$	$\longrightarrow$	S

- Knockoffs (Barber et al., 2015; Candès et al., 2018):
  - a wrapper around the blackbox algorithms;
  - produces a seletion set with FDR controlled below the target α.
- New paradigm:

Variable Selection Procedure: Knockoffs (Review)

- Scientific paradigm:

  n i.i.d. observations   Blackbox algorithm   Selection set

  $$(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i) \xrightarrow{\text{Blackbox algorithm}} S$$

- Knockoffs (Barber et al., 2015; Candès et al., 2018):
  - a wrapper around the blackbox algorithms;
  - produces a seletion set with FDR controlled below the target α.

- New paradigm:

  n i.i.d. observations   Knockoffs + Blackbox algorithm   Selection set

  $$(X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i) \xrightarrow{\text{Knockoffs + Blackbox algorithm}} S$$

Variable Selection Procedure: Knockoffs (Review)

- For each X_j, create a knockoff copy $\tilde{X}_j$ that serves as a "control".

Variable Selection Procedure: Knockoffs (Review)

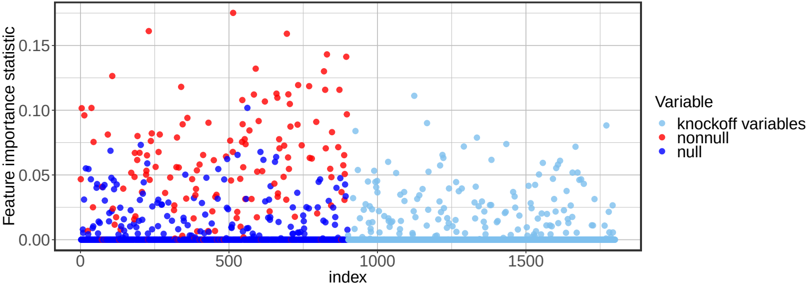
- For each X_j, create a knockoff copy $\tilde{X}_j$ that serves as a "control".
- Use the blackbox algorithm to assess the effect of X_j on Y and the effect of $\tilde{X}_j$ on Y.

Variable Selection Procedure: Knockoffs (Review)

- For each X_j, create a knockoff copy $\tilde{X}_j$ that serves as a "control".
- Use the blackbox algorithm to assess the effect of X_j on Y and the effect of $\tilde{X}_j$ on Y.
- Compare the effects.

Variable Selection Procedure: Knockoffs (Review)

- For each X_j, create a knockoff copy $\tilde{X}_j$ that serves as a "control".
- Use the blackbox algorithm to assess the effect of X_j on Y and the effect of $\tilde{X}_j$ on Y.
- Compare the effects.



Knockoffs with Side Information: Adaptive Knockoffs

- A variable selection procedure that utilizes the side information.
- Controls the finite-sample FDR conditional on the side information.
- Improves the statistical power in simulations and real applications.

Knockoffs with Side Information: Adaptive Knockoffs

Input: (X, Y)

- Construct a knockoff copy $\tilde{X}$.

Knockoffs with Side Information: Adaptive Knockoffs

Input: (X, Y)

- Construct a knockoff copy $\tilde{X}$.
- Apply our favorite blackbox algorithm $\mathcal{A}$ to $(X, \tilde{X}, Y)$ to generate a feature importance score Z_j and $\tilde{Z}_j$ for each X_j and $\tilde{X}_j$:

$$(Z_1, \ldots, Z_j, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_j, \ldots, \tilde{Z}_p) = \mathcal{A}([X, \tilde{X}], Y).$$

Knockoffs with Side Information: Adaptive Knockoffs

Input: (X, Y)

- Construct a knockoff copy $\tilde{X}$.
- Apply our favorite blackbox algorithm $\mathcal{A}$ to $(X, \tilde{X}, Y)$ to generate a feature importance score Z_j and $\tilde{Z}_j$ for each X_j and $\tilde{X}_j$:

$$(Z_1, \ldots, Z_j, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_j, \ldots, \tilde{Z}_p) = \mathcal{A}([X, \tilde{X}], Y).$$

- Contruct for each j a feature importace statistic that contrasts Z_j and $\tilde{Z}_j$:

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

Knockoffs with Side Information: Adaptive Knockoffs

Input: (X, Y)

- Construct a knockoff copy $\tilde{X}$.
- Apply our favorite blackbox algorithm $\mathcal{A}$ to $(X, \tilde{X}, Y)$ to generate a feature importance score Z_j and $\tilde{Z}_j$ for each X_j and $\tilde{X}_j$:

$$(Z_1, \ldots, Z_j, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_j, \ldots, \tilde{Z}_p) = \mathcal{A}([X, \tilde{X}], Y).$$

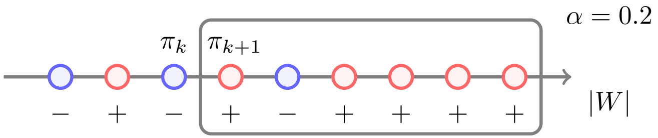
- Contruct for each j a feature importace statistic that contrasts Z_j and $\tilde{Z}_j$:

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

Lemma (Candès et al., '18)

Conditional on $(|W_1|, \ldots, |W_p|)$*, the signs of the null* W_j*'s,* $j \in \mathcal{H}_0$*, are i.i.d. coin flips.*

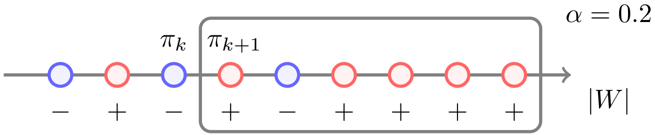
Knockoffs with Side Information: Adaptive Knockoffs



- Knockoffs
  Sequentially examines the hypotheses in an ordering determined by W_j.

- Adaptive Knockoffs
  Sequentially examines the hypotheses in an ordering determined by (W_j, U_j).

Knockoffs with Side Information: Adaptive Knockoffs



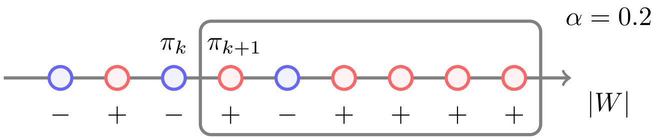
Adaptive Knockoffs Algorithm: for steps $k = 0, 1, 2 \ldots$

- Compute the estimated FDP among the unexamined hypotheses:

$$\widehat{\mathrm{FDP}}(k) = \frac{1 + \#\{j > k : W_{\pi_j} < 0\}}{\#\{j > k : W_{\pi_j} > 0\}}.$$

If $\widehat{\mathrm{FDP}}(k) \leq \alpha$, stop the procedure; otherwise, proceed.

Knockoffs with Side Information: Adaptive Knockoffs



Adaptive Knockoffs Algorithm: for steps $k = 0, 1, 2 \ldots$

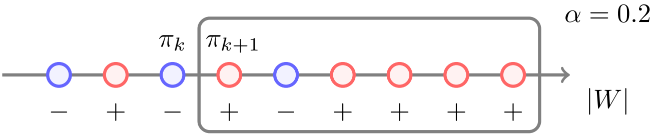
- Compute the estimated FDP among the unexamined hypotheses:

$$\widehat{\mathrm{FDP}}(k) = \frac{1 + \#\{j > k : W_{\pi_j} < 0\}}{\#\{j > k : W_{\pi_j} > 0\}}.$$

  If $\widehat{\mathrm{FDP}}(k) \leq \alpha$, stop the procedure; otherwise, proceed.
- Use a *filter* ϕ_{k+1} to determine the next hypothesis to be examined:

$$\pi_{k+1} = \phi_{k+1}.$$

Knockoffs with Side Information: Adaptive Knockoffs



Adaptive Knockoffs Algorithm: for steps $k = 0, 1, 2 \ldots$

- Compute the estimated FDP among the unexamined hypotheses:

$$\widehat{\mathrm{FDP}}(k) = \frac{1 + \#\{j > k : W_{\pi_j} < 0\}}{\#\{j > k : W_{\pi_j} > 0\}}.$$

  If $\widehat{\mathrm{FDP}}(k) \le \alpha$, stop the procedure; otherwise, proceed.
- Use a *filter* ϕ_{k+1} to determine the next hypothesis to be examined:

$$\pi_{k+1} = \phi_{k+1}.$$

- Output the unexamined features with positive feature importance statistics.

Adaptive knockoffs: FDR control

Requirement

At step k, the filter ϕ_{k+1} is measurable w.r.t. the σ-field (denoted by $\mathcal{F}_k$) generated by the "available information":

Adaptive knockoffs: FDR control

Requirement

At step k, the filter ϕ_{k+1} is measurable w.r.t. the σ-field (denoted by $\mathcal{F}_k$) generated by the "available information":

- Magnitude of all W_j's: $|W_j|$ for $j \in [p]$.

Adaptive knockoffs: FDR control

Requirement

At step k, the filter ϕ_{k+1} is measurable w.r.t. the σ-field (denoted by $\mathcal{F}_k$) generated by the "available information":

- Magnitude of all W_j's: $|W_j|$ for $j \in [p]$.
- Signs of the W_j's that have been examined: $\mathrm{sign}(W_{\pi_j})$ for $j \leq k$.

Adaptive knockoffs: FDR control

Requirement

At step k, the filter ϕ_{k+1} is measurable w.r.t. the σ-field (denoted by $\mathcal{F}_k$) generated by the "available information":

- Magnitude of all W_j's: $|W_j|$ for $j \in [p]$.
- Signs of the W_j's that have been examined: $\mathrm{sign}(W_{\pi_j})$ for $j \leq k$.
- Side information: U_j for $j \in [p]$.

Adaptive knockoffs: FDR control

Requirement

At step k, the filter ϕ_{k+1} is measurable w.r.t. the σ-field (denoted by $\mathcal{F}_k$) generated by the "available information":

- Magnitude of all W_j's: $|W_j|$ for $j \in [p]$.
- Signs of the W_j's that have been examined: $\mathrm{sign}(W_{\pi_j})$ for $j \leq k$.
- Side information: U_j for $j \in [p]$.
- The signs of the non-null W_j's: $\{\mathrm{sign}(W_j)\}_{j \text{ non-null}}$.

Adaptive knockoffs: FDR control

Requirement

At step k, the filter ϕ_{k+1} is measurable w.r.t. the σ-field (denoted by $\mathcal{F}_k$) generated by the "available information":

- Magnitude of all W_j's: $|W_j|$ for $j \in [p]$.
- Signs of the W_j's that have been examined: $\mathrm{sign}(W_{\pi_j})$ for $j \leq k$.
- Side information: U_j for $j \in [p]$.
- The signs of the non-null W_j's: $\{\mathrm{sign}(W_j)\}_{j \text{ non-null}}$.
- The number of positive and negative null W_j's in the unexamined hypotheses.

Adaptive knockoffs: FDR control

Requirement

At step k, the filter ϕ_{k+1} is measurable w.r.t. the σ-field (denoted by $\mathcal{F}_k$) generated by the "available information":

- Magnitude of all W_j's: $|W_j|$ for $j \in [p]$.
- Signs of the W_j's that have been examined: $\mathrm{sign}(W_{\pi_j})$ for $j \leq k$.
- Side information: U_j for $j \in [p]$.
- The signs of the non-null W_j's: $\{\mathrm{sign}(W_j)\}_{j \text{ non-null}}$.
- The number of positive and negative null W_j's in the unexamined hypotheses.

Theorem (R. and Candès, '20+)

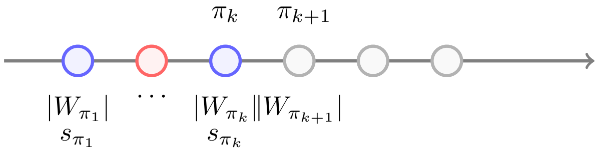
Given X, Y, U, if $\tilde{X}$ is valid knockoff copy of X conditional on U, and if the filter ϕ_{k+1} is measurable w.r.t. $\mathcal{F}_k$ for $k = 0, \ldots, p-1$, adaptive knockoffs controls the FDR below nominal level α (conditional on U).

Adaptive knockoffs: choices of filters

At step k, how should we use the available information ($\mathcal{F}_k$) to construct the filter ϕ_{k+1}?

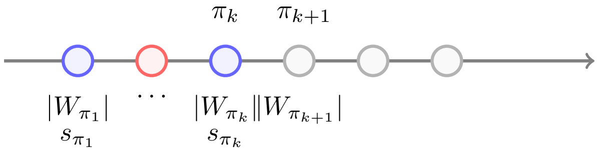
Adaptive knockoffs: choices of filters

At step k, how should we use the available information ($\mathcal{F}_k$) to construct the filter ϕ_{k+1}?



Adaptive knockoffs: choices of filters

At step k, how should we use the available information ($\mathcal{F}_k$) to construct the filter ϕ_{k+1}?



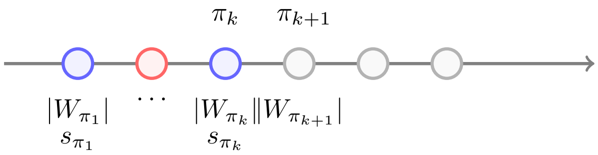
Example: GLM filter

- Model the probability of having a negative W_j via GLM:

$$\mathbb{P}(\mathrm{sign}(W_j) = -1 \mid |W_j|, U_j) = \frac{e^{\beta_0 + \beta_1 |W_j| + \beta_2 U_j}}{1 + e^{\beta_0 + \beta_1 |W_j| + \beta_2 U_j}}.$$

Adaptive knockoffs: choices of filters

At step k, how should we use the available information ($\mathcal{F}_k$) to construct the filter ϕ_{k+1}?



Example: GLM filter

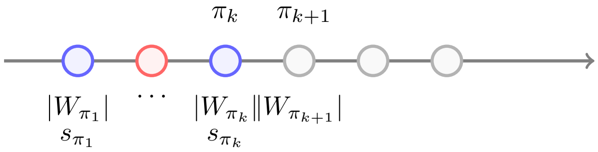
- Model the probability of having a negative W_j via GLM:

$$\mathbb{P}(\text{sign}(W_j) = -1 \,|\, |W_j|, U_j) = \frac{e^{\beta_0 + \beta_1 |W_j| + \beta_2 U_j}}{1 + e^{\beta_0 + \beta_1 |W_j| + \beta_2 U_j}}.$$

- Fit the model using available data.

Adaptive knockoffs: choices of filters

At step k, how should we use the available information ($\mathcal{F}_k$) to construct the filter ϕ_{k+1}?



Example: GLM filter

- Model the probability of having a negative W_j via GLM:

$$\mathbb{P}(\text{sign}(W_j) = -1 \,|\, |W_j|, U_j) = \frac{e^{\beta_0 + \beta_1 |W_j| + \beta_2 U_j}}{1 + e^{\beta_0 + \beta_1 |W_j| + \beta_2 U_j}}.$$

- Fit the model using available data.
- Pick the hypothesis with the highest probability of having a negative W_j among the unexamined hypothesis, i.e.

$$\phi_{k+1} = \underset{j \in [p] \setminus \{\pi_1, \ldots, \pi_k\}}{\text{argmax}} \mathbb{P}(\text{sign}(W_j) = -1 \,|\, |W_j|, U_j)$$

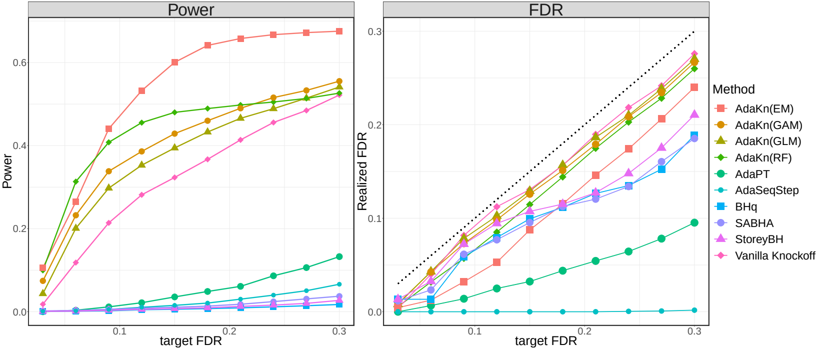
Adaptive Knockoffs: choices of filters

- Alternative models: GAM, Random Forest, Neural Network...

Adaptive Knockoffs: choices of filters

- Alternative models: GAM, Random Forest, Neural Network...
- The correctness of the model does not affect the FDR control (but may affect the power).

Numerical Simulations



Application

We apply adaptive knockoffs to the WTCCC Crohn's disease dataset.

Application

We apply adaptive knockoffs to the WTCCC Crohn's disease dataset.

- Inferential goal: discover which genetic variants are significant w.r.t. Crohn's disease among the British population.

Application

We apply adaptive knockoffs to the WTCCC Crohn's disease dataset.

- Inferential goal: discover which genetic variants are significant w.r.t. Crohn's disease among the British population.
- Side information: the summary statistics (p-values or z-values corresponding to SNPs) reported by previous GWAS in Crohn's disease among other populations.

Application

We apply adaptive knockoffs to the WTCCC Crohn's disease dataset.

- Inferential goal: discover which genetic variants are significant w.r.t. Crohn's disease among the British population.
- Side information: the summary statistics (p-values or z-values corresponding to SNPs) reported by previous GWAS in Crohn's disease among other populations.
- Obtain summary statistics from GWAS in East Asia, Iran, Belgium, Germany and the US.

GWAS in Crohn's disease

Study/Method	Number of SNPs discovered
WTCCC. (2007)	9
Candès et al. (2018)	18
Sesia et al. (2018)	22.8
Adaptive knockoffs	33.3

Table: Number of SNPs discovered to be associated with Crohn's disease by different methods.

References I

Barber, R. F., Candès, E. J., et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection series b statistical methodology.

Sesia, M., Sabatti, C., and Candès, E. (2018). Gene hunting with hidden markov model knockoffs. *Biometrika*.

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

Knockoffs with side information

(https://arxiv.org/abs/2001.07835)