Derandomizing Knockoffs

— Stable Variable Selection with Error Control

Zhimei Ren

Stanford University

International Seminar on Selective Inference
October 15th, 2020

Collaborators

Yuting Wei
CMU Stat

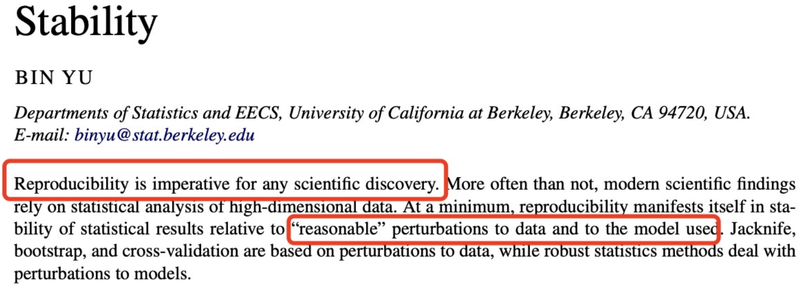
Emmanuel Candès
Stanford Stat & Math

Replication crisis

- Bayer Healthcare could replicate only 25% of 67 pre-clinical experiments [Prinz et al., 2011]
- Amgen could only confirm the findings in 6 out of 53 landmark cancer papers [Begley & Ellis, 2012]
- Social science papers in Science and Nature (2010 - 2015): only 13 out of 21 are consistent

The Economist
OCTOBER 19TH–25TH 2013
Economist.com
Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar
HOW SCIENCE GOES WRONG
99
Einsteinium

https://www.bbc.com/news/science-environment-39054778



Stability

BIN YU

Departments of Statistics and EECS, University of California at Berkeley, Berkeley, CA 94720, USA.
E-mail: binyu@stat.berkeley.edu

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to "reasonable" perturbations to data and to the model used. Jacknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models.

Variable selection

Explanatory Variables Response

$$(X_1, \textcolor{red}{X_2}, \ldots, X_p) \qquad \xrightarrow{\qquad\qquad} \qquad Y$$

Variable selection

Explanatory Variables Response

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

Detect the important variables that explain the response.

Variable selection

Explanatory Variables Response

$$(X_1, X_2, \dots, X_p) \xrightarrow{\hspace{4em}} Y$$

Detect the important variables that explain the response.

Stable and consistent selection

Variable selection

Explanatory Variables Response

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

Detect the important variables that explain the response.

Stable and consistent selection

- from perturbed datasets (error control);

Variable selection

Explanatory Variables $\qquad\qquad\qquad\qquad$ Response

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

Detect the important variables that explain the response.

Stable and consistent selection

- from perturbed datasets (error control);
- from different runs of one procedure.

Variable selection

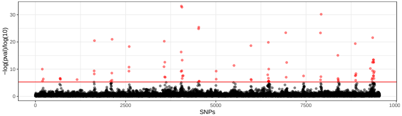
Explanatory Variables   Response

$$(X_1, X_2, \dots, X_p) \xrightarrow{\qquad\qquad\qquad} Y$$

Detect the important variables that explain the response.

Stable and consistent selection

- from perturbed datasets (error control);
- from different runs of one procedure.



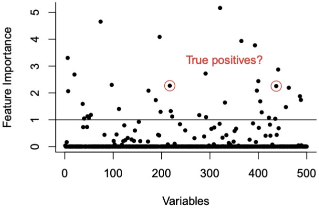
GWAS

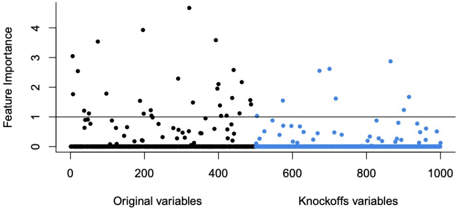
MRI

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

Model-X knockoffs framework

— [Barber and Candès, 2015; Candès et al., 2018]

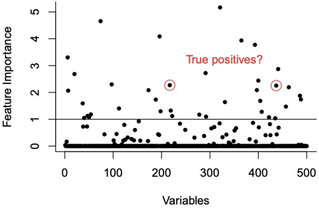


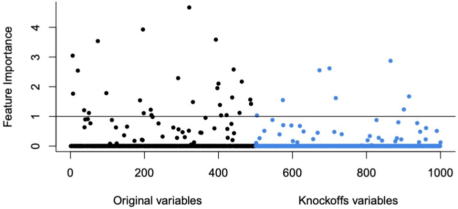


- Generate random "fake" copies.
- Controls the FDR.
- Another version [Janson and Su, 2016] controls the PFER and k-FWER.

Model-X knockoffs framework

— [Barber and Candès, 2015; Candès et al., 2018]

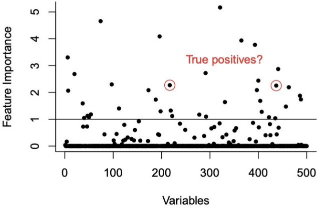


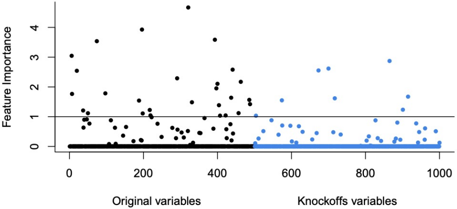


different runs ⇒ different selection sets

Model-X knockoffs framework

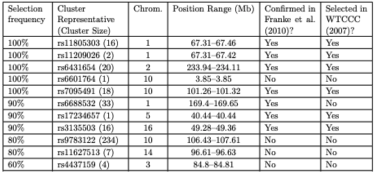
— [Barber and Candès, 2015; Candès et al., 2018]





[Candès et al., 2018]

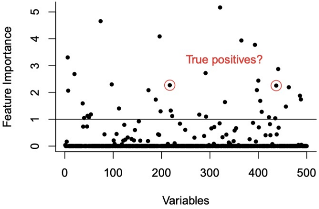
different runs ⇒ different selection sets

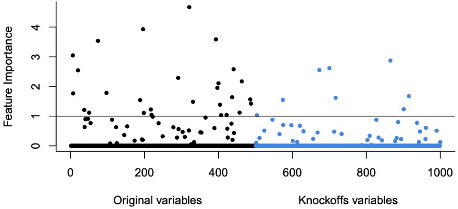


Selection frequency	Cluster Representative (Cluster Size)	Chrom.	Position Range (Mb)	Confirmed in Franke et al. (2010)?	Selected in WTCCC (2007)?
100%	rs11805303 (16)	1	67.31–67.46	Yes	Yes
100%	rs11209026 (2)	1	67.31–67.42	Yes	Yes
100%	rs6431654 (20)	2	233.94–234.11	Yes	Yes
100%	rs6601764 (1)	10	3.85–3.85	No	No
100%	rs7095491 (18)	10	101.26–101.32	Yes	Yes
90%	rs6688532 (33)	1	169.4–169.65	Yes	No
90%	rs17234657 (1)	5	40.44–40.44	Yes	Yes
90%	rs3135503 (16)	16	49.28–49.36	Yes	Yes
80%	rs9783122 (234)	10	106.43–107.61	No	No
80%	rs11627513 (7)	14	96.61–96.63	No	No
60%	rs4437159 (4)	3	84.8–84.81	No	No

Model-X knockoffs framework

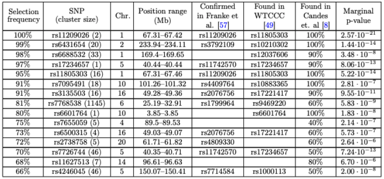
— [Barber and Candès, 2015; Candès et al., 2018]





[Sesia, Sabatti and Candès, 2019]

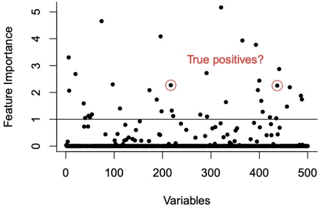
different runs ⇒ different selection sets

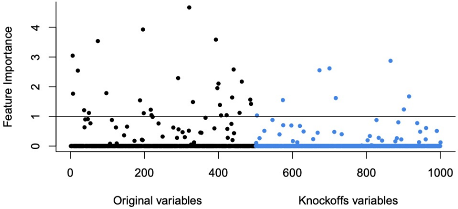


Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Franke et al. [57]	Found in WTCCC [49]	Found in Candes et. al [8]	Marginal p-value
100%	rs11209026 (2)	1	67.31–67.42	rs11209026	rs11805303	100%	$2.57{\cdot}10^{-21}$
99%	rs6431654 (20)	2	233.94–234.11	rs3792109	rs10210302	100%	$1.44{\cdot}10^{-14}$
98%	rs6688532 (33)	1	169.4–169.65		rs12037606	90%	$3.48 \cdot 10^{-8}$
97%	rs17234657 (1)	5	40.44–40.44	rs11742570	rs17234657	90%	$8.06{\cdot}10^{-13}$
95%	rs11805303 (16)	1	67.31–67.46	rs11209026	rs11805303	100%	$5.22{\cdot}10^{-14}$
91%	rs7095491 (18)	10	101.26–101.32	rs4409764	rs10883365	100%	$2.81 \cdot 10^{-7}$
91%	rs3135503 (16)	16	49.28–49.36	rs2076756	rs17221417	90%	$9.55{\cdot}10^{-11}$
81%	rs7768538 (1145)	6	25.19–32.91	rs1799964	rs9469220	60%	$5.83 \cdot 10^{-9}$
80%	rs6601764 (1)	10	3.85–3.85		rs6601764	100%	$1.83 \cdot 10^{-8}$
75%	rs7655059 (5)	4	89.5–89.53			40%	$2.14 \cdot 10^{-7}$
73%	rs6500315 (4)	16	49.03–49.07	rs2076756	rs17221417	60%	$5.73 \cdot 10^{-7}$
72%	rs2738758 (5)	20	61.71–61.82	rs4809330		60%	$2.64 \cdot 10^{-6}$
70%	rs7726744 (46)	5	40.35–40.71	rs11742570	rs17234657	50%	$7.24{\cdot}10^{-13}$
68%	rs11627513 (7)	14	96.61–96.63			80%	$6.70 \cdot 10^{-6}$
66%	rs4246045 (46)	5	150.07–150.41	rs7714584	rs1000113	50%	$2.00 \cdot 10^{-8}$

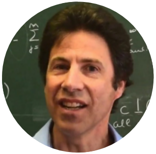
Model-X knockoffs framework

— [Barber and Candès, 2015; Candès et al., 2018]



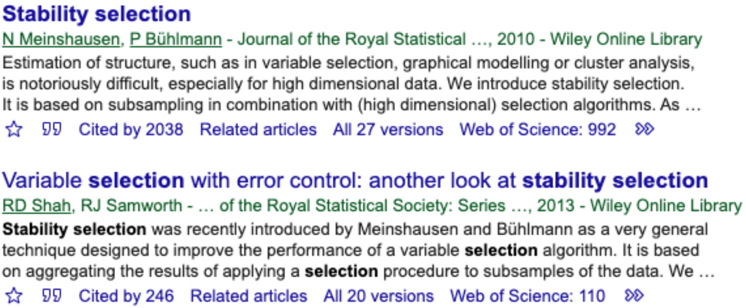


Stability?



different runs ⇒ different selection sets

Stablity selection



Stability selection

N Meinshausen, P Bühlmann - Journal of the Royal Statistical …, 2010 - Wiley Online Library

Estimation of structure, such as in variable selection, graphical modelling or cluster analysis, is notoriously difficult, especially for high dimensional data. We introduce stability selection. It is based on subsampling in combination with (high dimensional) selection algorithms. As …

☆ ” Cited by 2038 Related articles All 27 versions Web of Science: 992 »

Variable **selection** with error control: another look at **stability selection**

RD Shah, RJ Samworth - … of the Royal Statistical Society: Series …, 2013 - Wiley Online Library

Stability selection was recently introduced by Meinshausen and Bühlmann as a very general technique designed to improve the performance of a variable **selection** algorithm. It is based on aggregating the results of applying a **selection** procedure to subsamples of the data. We …

☆ ” Cited by 246 Related articles All 20 versions Web of Science: 110 »

Stablity selection (original form)

1. Start with the full dataset $\boldsymbol{Z}_{\mathsf{full}} = Z_1, \ldots, Z_n$.

Stablity selection (original form)

1. Start with the full dataset $\boldsymbol{Z}_{\mathsf{full}} = Z_1, \ldots, Z_n$.
2. For each $m = 1, \ldots, M$
   (i) Subsample without replacement to generate a smaller dataset of size $n/2$, denoted by $\boldsymbol{Z}_{(m)}$;
   (ii) Run the selection algorithm on $\boldsymbol{Z}_{(m)}$ to obtain a selection set $\widehat{S}^m$.

Stablity selection (original form)

1. Start with the full dataset $\boldsymbol{Z}_{\textsf{full}} = Z_1, \ldots, Z_n$.
2. For each $m = 1, \ldots, M$
   (i) Subsample without replacement to generate a smaller dataset of size $n/2$, denoted by $\boldsymbol{Z}_{(m)}$;
   (ii) Run the selection algorithm on $\boldsymbol{Z}_{(m)}$ to obtain a selection set $\widehat{S}^m$.
3. Calculate the selection fequency

$$\Pi_j = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{j \in \widehat{S}^m\}.$$

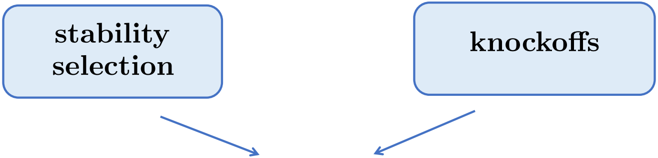
Stablity selection (original form)

1. Start with the full dataset $\boldsymbol{Z}_{\text{full}} = Z_1, \ldots, Z_n$.
2. For each $m = 1, \ldots, M$
   (i) Subsample without replacement to generate a smaller dataset of size $n/2$, denoted by $\boldsymbol{Z}_{(m)}$;
   (ii) Run the selection algorithm on $\boldsymbol{Z}_{(m)}$ to obtain a selection set $\widehat{S}^m$.
3. Calculate the selection fequency

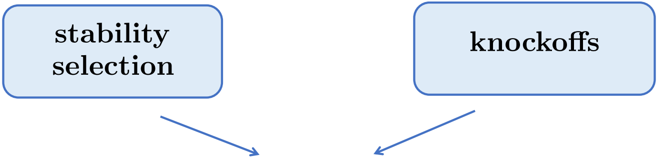
$$\Pi_j = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{j \in \widehat{S}^m\}.$$

4. Given a threshold $\eta > 0$, return the final selection set

$$\widehat{S} = \{j \in [p] : \Pi_j \geq \eta\}.$$



This work: derandomizing knockoffs



stability selection

knockoffs

This work: derandomizing knockoffs

stability, error guarantee and enhanced power

A brief review of the knockoffs framework

Variable selection: mathematical formulation

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

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

Variable selection: mathematical 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.

Variable selection: mathematical 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.

Variable selection: mathematical 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*

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

Variable selection: mathematical 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*

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

  - *Per Family Error Rate*

$$\mathsf{PFER} \triangleq \mathbb{E}\left[V\right].$$

  - *k family-wise error rate*

$$k\text{-}\mathsf{FWER} \triangleq \mathbb{P}\left(V \geq k\right).$$

Variable selection: mathematical 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*

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

  - *Per Family Error Rate*

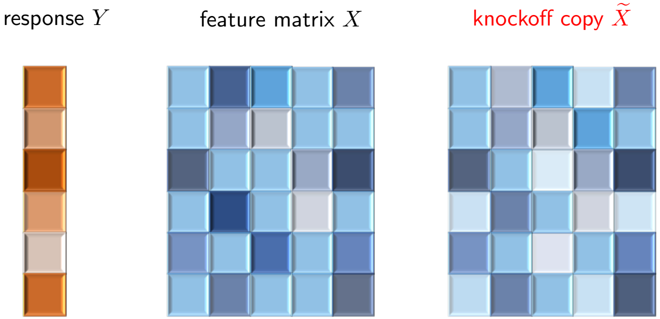
$$\mathsf{PFER} \triangleq \mathbb{E}\left[V\right].$$

  - *k family-wise error rate*

$$k\text{-}\mathsf{FWER} \triangleq \mathbb{P}\left(V \geq k\right).$$

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

Construct knockoffs



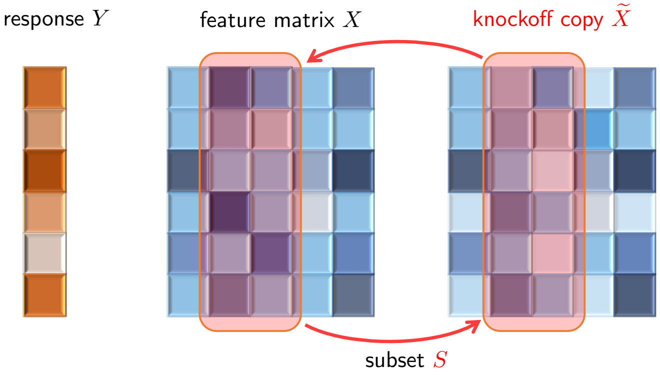
response Y

feature matrix X

knockoff copy $\widetilde{X}$

- $\widetilde{X} \perp\!\!\!\perp Y \mid X$

Construct knockoffs



- $\widetilde{X} \perp\!\!\!\perp Y \mid X$
- for any subset $S \subset \{1, 2, \ldots p\}$: distribution $(X, \widetilde{X})_{\text{swap}(S)} \overset{\text{d}}{=} (X, \widetilde{X})$

A simple example

Suppose $X \sim \mathcal{N}(0, \Sigma)$, how to construct $\widetilde{X}$?

A simple example

Suppose $X \sim \mathcal{N}(0, \Sigma)$, how to construct $\widetilde{X}$?

$$(X, \widetilde{X}) \sim \mathcal{N}(0, G) \quad \text{where} \quad G = \begin{bmatrix} \Sigma & \Sigma - \mathsf{diag}(s) \\ \Sigma - \mathsf{diag}(s) & \Sigma \end{bmatrix}.$$

A simple example

Suppose $X \sim \mathcal{N}(0, \Sigma)$, how to construct $\widetilde{X}$?

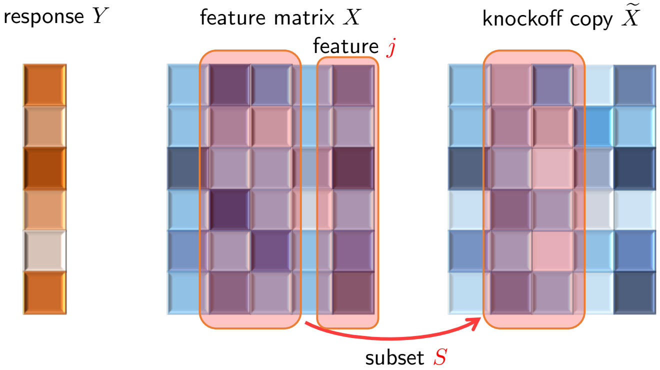
$$(X, \widetilde{X}) \sim \mathcal{N}(0, G) \quad \text{where} \quad G = \begin{bmatrix} \Sigma & \Sigma - \mathsf{diag}(s) \\ \Sigma - \mathsf{diag}(s) & \Sigma \end{bmatrix}.$$

$$\widetilde{X} \mid X \sim \mathcal{N}(\mu, V)$$

where

$$\begin{aligned} \mu &= X - X\Sigma^{-1}\mathsf{diag}(s) \\ V &= 2\mathsf{diag}(s) - \mathsf{diag}(s)\Sigma^{-1}\mathsf{diag}(s) \end{aligned}$$

Feature statistics $w_j([X, \widetilde{X}], y)$



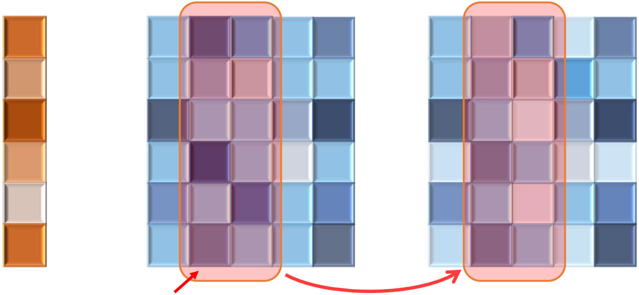
$$w_j([X, \widetilde{X}]_{\text{swap}(S)}, y) = w_j([X, \widetilde{X}], y) \qquad j \notin S$$

Feature statistics $w_j([X, \widetilde{X}], y)$

response Y

feature matrix X

knockoff copy $\widetilde{X}$



feature j

subset S

$$w_j([X, \widetilde{X}]_{\mathsf{swap}(S)}, y) = -w_j([X, \widetilde{X}], y) \qquad j \in S$$

A simple example: Lasso coefficient difference

Run Lasso

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{2p}} \quad \frac{1}{2}\|y-[X,\widetilde{X}]\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$$

Lasso coefficient difference statistics (LCD):

$$W_j = |\hat{\boldsymbol{\beta}}_j(\lambda)| - |\hat{\boldsymbol{\beta}}_{j+p}(\lambda)|$$

A simple example: Lasso coefficient difference

Run Lasso

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{2p}} \quad \frac{1}{2}\|y - [X, \widetilde{X}]\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Lasso coefficient difference statistics (LCD):

$$W_j = |\hat{\boldsymbol{\beta}}_j(\lambda)| - |\hat{\boldsymbol{\beta}}_{j+p}(\lambda)|$$

Key properties

- null W_j's are symmetrically distributed.
- conditional on $|W_j|$, signs of null W_j's are i.i.d. coin flips.

Define selection set

Model-X v-knockoffs [Janson and Su, 2016]

- Order the features according to the magnitudes of W_j's:

$$|W_{\pi_1}| \geq |W_{\pi_2}| \geq \ldots |W_{\pi_p}|.$$

- Define

$$T := \inf_{k \in [p]} \Big\{ \sum_{j=1}^{k} \mathbf{1}_{\{W_{\pi_j} < 0\}} \geq v \Big\}.$$

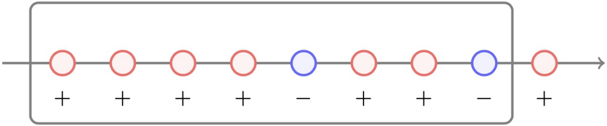
- Reject π_j such that $j \leq T$ and $W_{\pi_j} > 0$.

Define selection set

- Reject π_j such that $j \leq T$ and $W_{\pi_j} > 0$

$$T := \inf_{k \in [p]} \left\{ \sum_{j=1}^{k} \mathbf{1}_{\{W_{\pi_j} < 0\}} \geq v \right\}.$$

- If $v = 2$, stop the procedure the first time seeing 2 "$-$"s.



Property

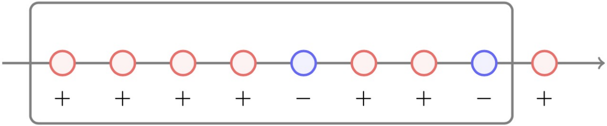
Conditional on $|W_j|$, signs of null W_j's are i.i.d. coin flips

Define selection set

- Reject π_j such that $j \leq T$ and $W_{\pi_j} > 0$

$$T := \inf_{k \in [p]} \left\{ \sum_{j=1}^{k} \mathbf{1}_{\{W_{\pi_j} < 0\}} \geq v \right\}.$$

- If $v = 2$, stop the procedure the first time seeing 2 "$-$"s.



Lemma (Janson and Su, 2016)

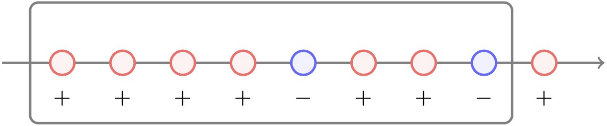
The number of false discoveries V is stochastically dominated by NB$(v, 1/2)$.

Define selection set

- Reject π_j such that $j \le T$ and $W_{\pi_j} > 0$

$$T := \inf_{k \in [p]} \left\{ \sum_{j=1}^{k} \mathbf{1}_{\{W_{\pi_j} < 0\}} \ge v \right\}.$$

- If $v = 2$, stop the procedure the first time seeing 2 "$-$"s.



Lemma (Janson and Su, 2016)

The number of false discoveries V is stochastically dominated by NB$(v, 1/2)$.

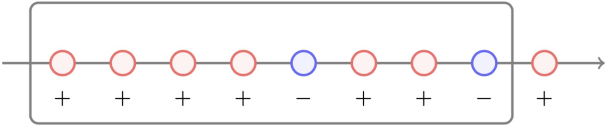
- $Z \sim \mathrm{NB}(m, q)$ negative binomial random variable

Define selection set

- Reject π_j such that $j \leq T$ and $W_{\pi_j} > 0$

$$T := \inf_{k \in [p]} \left\{ \sum_{j=1}^{k} \mathbf{1}_{\{W_{\pi_j} < 0\}} \geq v \right\}.$$

- If $v = 2$, stop the procedure the first time seeing 2 "$-$"s.



Lemma (Janson and Su, 2016)

The number of false discoveries V is stochastically dominated by NB$(v, 1/2)$.

- $\mathbb{E}[V] \leq v$.

Knockoffs framework (summary)

Three-step procedure:

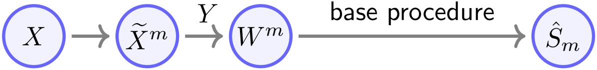
- Construct knockoff feature matrix $\widetilde{X} \in^{n \times p}$.
- Define feature statistics $w_j([X, \widetilde{X}, y])$ for each $j \in \{1, 2, \ldots, 2p\}$.
- Decide selection set $\widehat{S}$ (v-knockoffs).

Derandomizing knockoffs

- Given (X, Y), generate $m = 1, \ldots, M$ realizations of knockoffs.

Derandomizing knockoffs

- Given (X, Y), generate $m = 1, \ldots, M$ realizations of knockoffs.
- For each realization of knockoff m:



$$\begin{tikzcd} X \arrow[r] & \widetilde{X}^m \arrow[r, "Y"] & W^m \arrow[r, "\text{base procedure}"] & \hat{S}_m \end{tikzcd}$$

Derandomizing knockoffs

- Given (X, Y), generate $m = 1, \dots, M$ realizations of knockoffs.
- For each realization of knockoff m:

$$\begin{tikzcd} X \arrow[r] & \widetilde{X}^m \arrow[r, "Y"] & W^m \arrow[r, "\text{base procedure}"] & \hat{S}_m \end{tikzcd}$$

- For each feature j, define selection probability

$$\Pi_j := \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{j \in \hat{S}_m\}.$$

Derandomizing knockoffs

- Given (X, Y), generate $m = 1, \ldots, M$ realizations of knockoffs.
- For each realization of knockoff m:

$$\begin{tikzcd} X \arrow[r] & \widetilde{X}^m \arrow[r, "Y"] & W^m \arrow[rr, "\text{base procedure}"] & & \hat{S}_m \end{tikzcd}$$

- For each feature j, define selection probability

$$\Pi_j := \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{j \in \hat{S}_m\}.$$

- For a threshold η, the final selection set S is

$$\hat{S} := \{j \in [p] : \Pi_j \geq \eta\}.$$

Theoretical guarantees

Theorem (R., Wei and Candès ('20))

Suppose the the base procedure is the v-knockoffs. If for every $j \in \mathcal{H}_0$,

$$\mathbb{P}(\Pi_j \geq \eta) \ \leq \ \gamma \mathbb{E}[\Pi_j], \tag{1}$$

then the PFER can be controlled as

$$\mathbb{E}[V] \ \leq \ \gamma v.$$

- Per family error rate (PFER): $\mathbb{E}[V]$ (V number of false discoveries)

Theoretical guarantees

Theorem (R., Wei and Candès ('20))

Suppose the the base procedure is the v-knockoffs. If for every $j \in \mathcal{H}_0$,

$$\mathbb{P}(\Pi_j \geq \eta) \;\leq\; \gamma \mathbb{E}[\Pi_j], \tag{1}$$

then the PFER can be controlled as

$$\mathbb{E}[V] \;\leq\; \gamma v.$$

- Per family error rate (PFER): $\mathbb{E}[V]$ (V number of false discoveries)

$$\mathbb{E}[V] = \mathbb{E}\left[\sum_{j \in \mathcal{H}_0} \mathbb{1}\{\Pi_j \geq \eta\}\right] = \sum_{j \in \mathcal{H}_0} \mathbb{P}\left(\Pi_j \geq \eta\right)$$

$$\leq \sum_{j \in \mathcal{H}_0} \gamma \mathbb{E}[\Pi_j] = \gamma \mathbb{E}[V_1] \leq \gamma v$$

Theoretical guarantees

Theorem (R., Wei and Candès ('20))

Suppose the the base procedure is the v-knockoffs. If for every $j \in \mathcal{H}_0$,

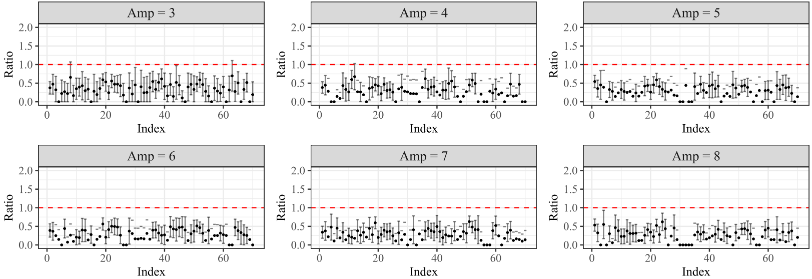
$$\mathbb{P}(\Pi_j \geq \eta) \; \leq \; \gamma \mathbb{E}[\Pi_j], \tag{1}$$

then the PFER can be controlled as

$$\mathbb{E}[V] \; \leq \; \gamma v.$$

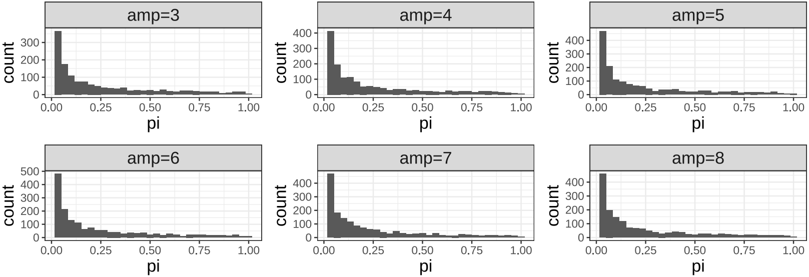
- Per family error rate (PFER): $\mathbb{E}[V]$ (V number of false discoveries)
- With $\eta = 1/2$, Markov's inequality gives $\gamma = 2$

Plotting the ratio for $\eta = 1/2$



Realized ratio of $\mathbb{P}(\Pi_j \geq 1/2)/\mathbb{E}[\Pi_j]$ with the 95% confidence interval, estimated from $1,000$ repetitions.

How to tighten γ ? An observation...



Pooled histogram of all nonzero null Π_j's.

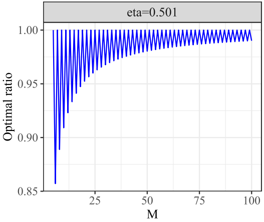
A sharper guarantee

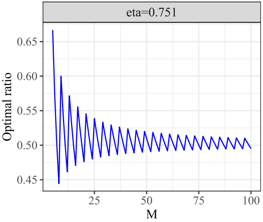
- If the pmf of Π_j is monotonically non-increasing for each $j \in \mathcal{H}_0$

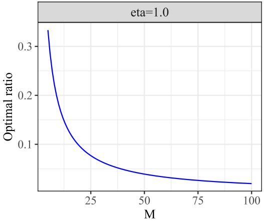
$$\gamma = \max \sum_{m \geq M\eta} y_m,$$

$$s.t. \quad y_m \geq 0, \quad y_{m-1} \geq y_m, \; m \in [M],$$

$$\sum_{m=0}^{M} y_m \cdot \frac{m}{M} = 1.$$







Theoretical guarantees

- k family-wise error rate (k-FWER): $\mathbb{P}(V \geq k)$.
- $Z \sim \mathrm{NB}(m, q)$ negative binomial random variable.

Theorem (R., Wei and Candès (20'))

Suppose condition (1) *holds with* γ. *For* $k \geq 2$, *suppose that*

$$\sum_{u=1}^{k-1} \mathbb{P}(V \in [k-u, k)) \geq \sum_{u=1}^{k} \mathbb{P}(V \in [k, k+u)),$$

then the k*-FWER can be controlled as*

$$\mathbb{P}(V \geq k) \leq \frac{\gamma v}{2k}.$$

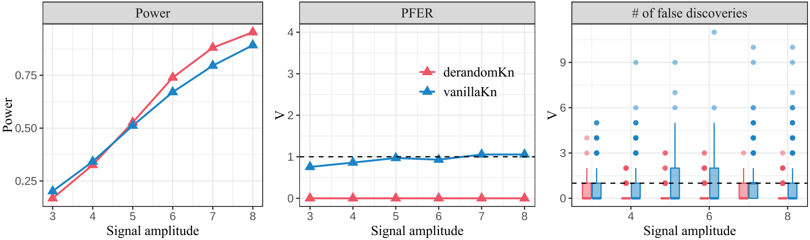
Theoretical guarantees (extensions)

- Under similar partial sum conditions, we can derive similar bounds using other types of inequality.

$$\mathbb{P}(V \geq k) \;\leq\; \min\left\{\frac{v}{2k},\; \frac{\mathbb{E}[(2Z)^{\alpha}]}{2k^{\alpha}},\; \frac{\mathbb{E}[\exp(\lambda(2Z))]}{2\exp(\lambda k)}\right\}$$

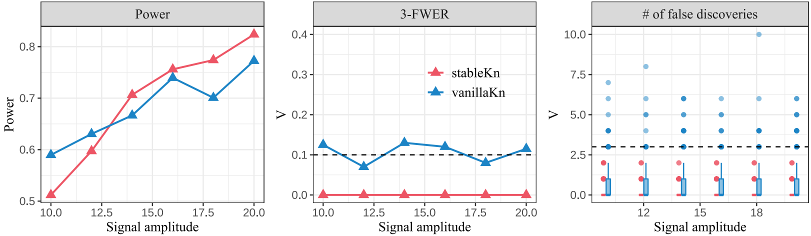
- The minimum is also taken over $\alpha \in [k-1], \lambda \in (0,1)$.

Simulation studies: PFER control



Settings: $n = 200$, $p = 100$, $X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\Sigma_{ij} = 0.6^{|i-j|}$, and $Y \mid X \sim$ a linear model with 30 non-zero coefficients. Each nonzero coefficient β_j takes value $A/\sqrt{n}$ where A ranges in $\{3, 4, \ldots, 8\}$ and the sign is determined by i.i.d. coin flips. The locations of the non-zero signal are randomly chosen from $[p]$. We show the averaged results over 200 trials.

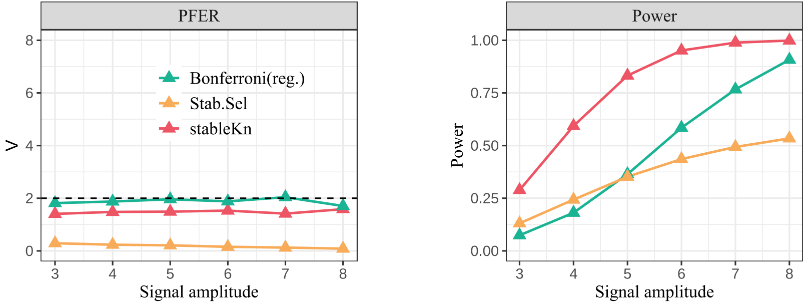
Simulation studies: k-FWER control



Settings: $n = 300$ and $p = 50$, $X \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.1^{|i-j|}$. $Y \mid X \sim$ logistic model with 20 non-zero entries in $\boldsymbol{\beta}$. These nonzero entries take values $A/\sqrt{n}$ where A ranges in $\{10, 12, \ldots, 20\}$ and the sign is determined by i.i.d. coin flips. Parameters $\eta = 0.5$ and $v = 0.6$.

Simulation studies: more comparisons

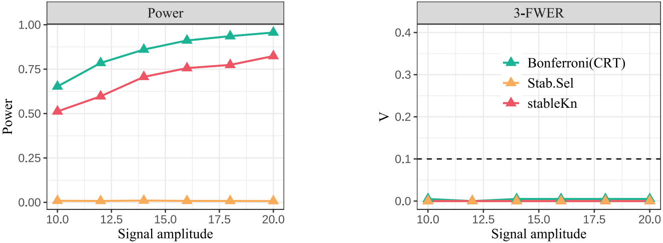
Compared to other methods: PFER



Settings: $n = 2000$, $p = 1000$ and $\Sigma_{ij} = 0.5^{|i-j|}$. $Y \mid X \sim$ a linear model with 60 non-zero coefficients. Target PFER level is $v = 2$.

Simulation studies: more comparisons

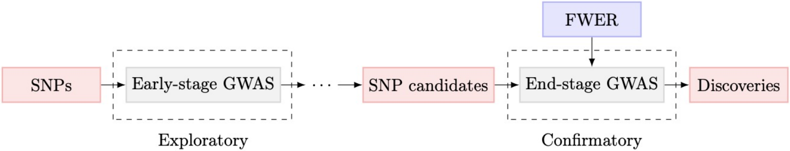
Compared to other methods: k-FWER



Settings: $n = 300$ and $p = 50$, $X \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.1^{|i-j|}$. $Y \mid X \sim$ logistic model with 20 non-zero entries in $\boldsymbol{\beta}$. These nonzero entries take values $A/\sqrt{n}$ where A ranges in $\{10, 12, \ldots, 20\}$ and the sign is determined by i.i.d. coin flips. Parameters $\eta = 0.5$ and $v = 0.6$.

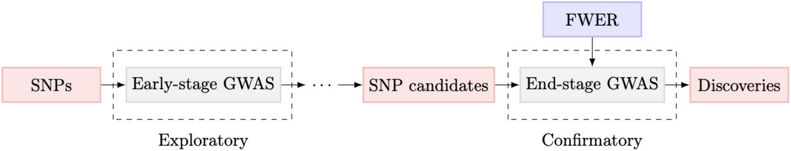
Genome-Wide Association Study (GWAS)

A typical workflow of multi-stage GWAS:



Genome-Wide Association Study (GWAS)

A typical workflow of multi-stage GWAS:

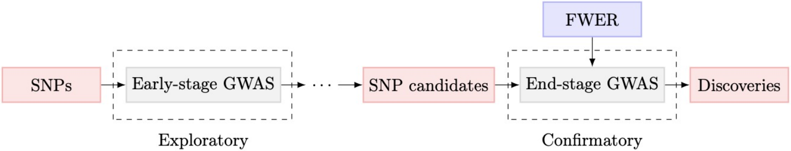


Potential problem:

- Suppose a subset of candidate SNPs $\mathcal{C}$ is selected in stage one.
- Conduct data analysis on Y and $X_{\mathcal{C}}$.
- Answering question about $Y \mid X_{\mathcal{C}}$ instead of $Y \mid X$?

Genome-Wide Association Study (GWAS)

A typical workflow of multi-stage GWAS:



Conditional knockoffs:

- suppose a subset of candidate SNPs $\mathcal{C}$ is selected in stage one
- construct a conditional knockoff copy *only* for $X_{\mathcal{C}}$

$$(X_{\mathcal{C}}, \tilde{X}_{\mathcal{C}})_{\mathsf{swap}(g)} \mid X_{-\mathcal{C}} \overset{\mathrm{d}}{=} (X_{\mathcal{C}}, \tilde{X}_{\mathcal{C}}) \mid X_{-\mathcal{C}}$$

A real data example

Procedures

- Data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)

Procedures

- Data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)
- Early-stage: selecting p-values from [schumacher et al., 2018] below 10^{-3} gives 4072 pre-selected SNPs

Procedures

- Data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)
- Early-stage: selecting p-values from [schumacher et al., 2018] below 10^{-3} gives 4072 pre-selected SNPs
- Partition the SNPs into clusters at a level of resolution 2% and the resulting average length of the clusters is 0.226 Mb.

Procedures

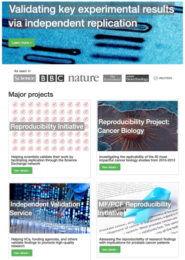
- Data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)
- Early-stage: selecting p-values from [schumacher et al., 2018] below 10^{-3} gives 4072 pre-selected SNPs
- Partition the SNPs into clusters at a level of resolution 2% and the resulting average length of the clusters is 0.226 Mb.
- Apply derandomized knockoffs with target FWER level 0.1 (ten runs of conditional group HMM knockoffs)

Results

Lead SNP	Chromosome	Position range (Mb)	Size	Confirmed by?
rs12621278	2	173.28-173.58	68	[Wang et al. (2015)]
rs1512268	8	23.39-23.55	48	[Wang et al. (2015)]
rs1016343	8	128.07-128.24	45	[Hui et al. (2014)]
rs6983267	8	128.40-128.47	37	[Wang et al. (2015)]
rs7121039	11	2.18-2.31	40	[Wang et al. (2015)]*
rs10896449	11	68.80-69.02	62	[Wang et al. (2015)]
rs7501939	17	36.05-36.18	55	[Elliott et al. (2010)]
rs1859962	17	69.07-69.24	40	[Wang et al. (2015)]

Discoveries at 2% resolution and the target FWER level set to 0.1 and $\eta = 1$ and $M = 10$.

Concluding remarks



Validating key experimental results via independent replication

Major projects

Reproducibility Initiative

Helping scientists validate their work by facilitating replication through the Science Exchange network

Reproducibility Project: Cancer Biology

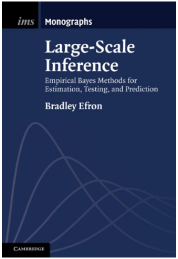
Investigating the replicability of the 50 most impactful cancer biology studies from 2010-2012

Independent Validation Service

Helping VCs, funding agencies, and others validate findings to promote high-quality research

MF/PCF Reproducibility Initiative

Assessing the reproducibility of research findings with implications for prostate cancer patients



ims Monographs

Large-Scale Inference

Empirical Bayes Methods for Estimation, Testing, and Prediction

Bradley Efron

CAMBRIDGE

Future directions

- Adapt to other base procedures.
- Characterize the power.
- False discovery rate or false discovery exceedence.

— “Derandomizing Knockoffs,” Zhimei Ren, Yuting Wei, and Emmanuel Candès, in preparation, 2020