Searching for Consistent Associations with a Multi-Environment Knockoff Filter

Shuangning Li

Stanford University

Deep Learning for Genetics, Genomics and Metagenomics

Joint work with

M. Sesia

Y. Romano

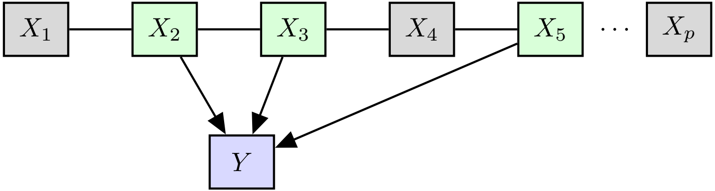
E. Candès

C. Sabatti

Li, Sesia, Romano, Candès & Sabatti. **Searching for Consistent Associations with a Multi-Environment Knockoff Filter.** *Biometrika*, 2021.

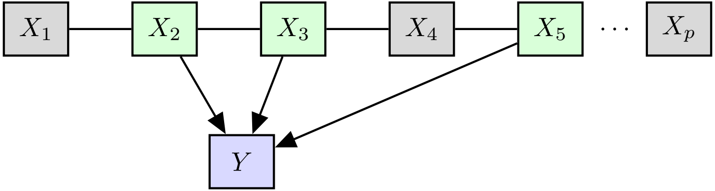
Introduction

- In Genome-Wide Association Studies (GWAS), geneticists have measured hundreds of thousands of genetic variants and wish to know which of these influence a trait. E.g. What are the genes that influence **height**?



- In standard analysis, focus variables X_j's that are associated with Y. The results are very far from identifying "causal" genetic variants.

Conditional Independence



- Better goal: test for conditional independence

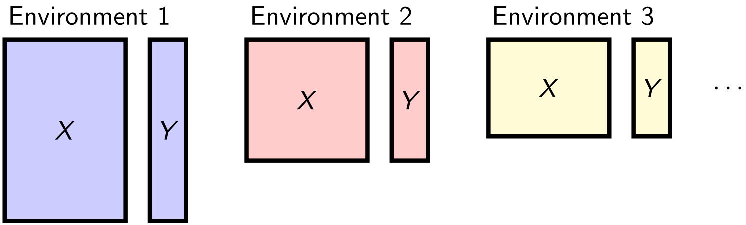
$$H_j : X_j \perp\!\!\!\perp Y \mid X_{-j}.$$

  If H_j is true, the j-th variable does not provide information about the response Y beyond what is already provided by all the other variables.

- Control the false discovery rate

$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections}}\right].$$

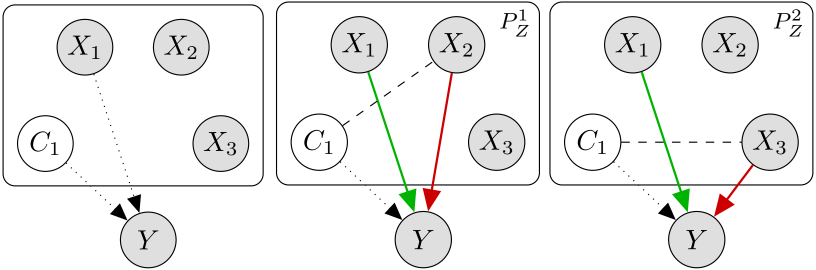
Consistence across environments



- We say a variable j is null in environment e if $X_j^e \perp\!\!\!\perp Y^e | X_{-j}^e$. We would like to find variables that are non-null in **all** environments.
- In other words, now a variable is null for "consistent independence hypothesis", if it is null in at least one environments.

Unobserved Confounder

- Shaded nodes are observed; “white” nodes are not.
- Dotted arrows represent true causal model connecting Z to Y
- Broken lines identify correlations across variables
- Filled arrows indicated detected conditional association

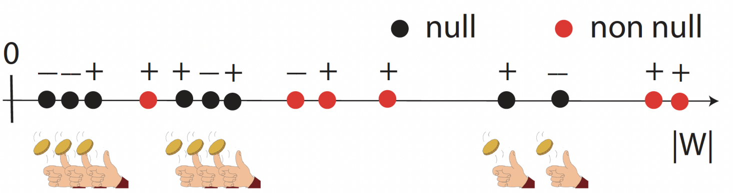


Knockoffs

- The method of **knockoffs** (Barber and Candès, 2015; Candès et al., 2018) allows one to test the conditional independence hypothesis and provably controls the FDR.
- Construct knockoffs $\Rightarrow$ Get important statistics $\Rightarrow$ Report selected set

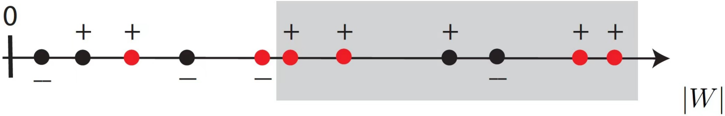
Knockoffs

- We compute importance statistics W from machine learning algorithms.
- Large W_j says that variable j appears important.
- Conditional on $|W|$, signs of null W_j's are i.i.d. coin flips — crucial for FDR control!



- FDR is still controlled if the signs are $\leq$ coin flips (More conservative!)

Knockoffs



- Let

$$\tau = \min\left\{ t : \widehat{\mathsf{FDP}}(t) = \frac{1 + |\mathcal{S}^-(t)|}{1 \vee |\mathcal{S}^+(t)|} \le q \right\}$$

- Report $\hat{\mathcal{S}} = \{W_j \ge \tau\}$.
- Then false discovery rate is controlled.

$$\mathbb{E}\left[\frac{\#\ \text{false positives}}{\#\ \text{selections}}\right] \le q$$

Multi-Environment Knockoff Filter (simple version!)

1. Compute importance statistics for each environment $\{W^e\}_{e=1}^E$ as usual.
2. Merge the statistics as follows

$$\operatorname{sign}(W_j) = \min_e \operatorname{sign}\left(W_j^e\right)$$

$$|W_j| = f\left(|W_j^1|, |W_j^2|, \dots, |W_j^E|\right)$$

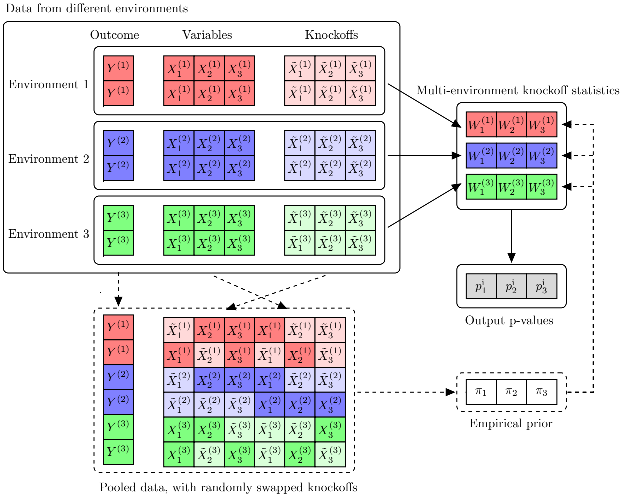
   For example,

$$|W_j| = \prod_{e=1}^{E} |W_j^e|$$

3. Report a selected set based on W.

- If a variable is "null", then it's null in at least one environment. Say it's null in envir 1. Then $\operatorname{sign}(W_j) = \min_e \operatorname{sign}\left(W_j^e\right) \leq \operatorname{sign}\left(W_j^1\right) \leq$ coin flips.

Multi-Environment Knockoff Filter (complicated version!)



Data from different environments

	Outcome	Variables			Knockoffs		
Environment 1	$Y^{(1)}$	$X_1^{(1)}$	$X_2^{(1)}$	$X_3^{(1)}$	$\tilde{X}_1^{(1)}$	$\tilde{X}_2^{(1)}$	$\tilde{X}_3^{(1)}$
	$Y^{(1)}$	$X_1^{(1)}$	$X_2^{(1)}$	$X_3^{(1)}$	$\tilde{X}_1^{(1)}$	$\tilde{X}_2^{(1)}$	$\tilde{X}_3^{(1)}$
Environment 2	$Y^{(2)}$	$X_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$	$\tilde{X}_1^{(2)}$	$\tilde{X}_2^{(2)}$	$\tilde{X}_3^{(2)}$
	$Y^{(2)}$	$X_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$	$\tilde{X}_1^{(2)}$	$\tilde{X}_2^{(2)}$	$\tilde{X}_3^{(2)}$
Environment 3	$Y^{(3)}$	$X_1^{(3)}$	$X_2^{(3)}$	$X_3^{(3)}$	$\tilde{X}_1^{(3)}$	$\tilde{X}_2^{(3)}$	$\tilde{X}_3^{(3)}$
	$Y^{(3)}$	$X_1^{(3)}$	$X_2^{(3)}$	$X_3^{(3)}$	$\tilde{X}_1^{(3)}$	$\tilde{X}_2^{(3)}$	$\tilde{X}_3^{(3)}$

Multi-environment knockoff statistics

$W_1^{(1)}$	$W_2^{(1)}$	$W_3^{(1)}$
$W_1^{(2)}$	$W_2^{(2)}$	$W_3^{(2)}$
$W_1^{(3)}$	$W_2^{(3)}$	$W_3^{(3)}$

p_1^{i}	p_2^{i}	p_3^{i}

Output p-values

$Y^{(1)}$	$\tilde{X}_1^{(1)}$	$X_2^{(1)}$	$X_3^{(1)}$	$X_1^{(1)}$	$\tilde{X}_2^{(1)}$	$\tilde{X}_3^{(1)}$
$Y^{(1)}$	$X_1^{(1)}$	$\tilde{X}_2^{(1)}$	$X_3^{(1)}$	$\tilde{X}_1^{(1)}$	$X_2^{(1)}$	$\tilde{X}_3^{(1)}$
$Y^{(2)}$	$\tilde{X}_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$	$X_1^{(2)}$	$\tilde{X}_2^{(2)}$	$\tilde{X}_3^{(2)}$
$Y^{(2)}$	$\tilde{X}_1^{(2)}$	$\tilde{X}_2^{(2)}$	$\tilde{X}_3^{(2)}$	$X_1^{(2)}$	$X_2^{(2)}$	$X_3^{(2)}$
$Y^{(3)}$	$X_1^{(3)}$	$X_2^{(3)}$	$\tilde{X}_3^{(3)}$	$\tilde{X}_1^{(3)}$	$\tilde{X}_2^{(3)}$	$X_3^{(3)}$
$Y^{(3)}$	$X_1^{(3)}$	$\tilde{X}_2^{(3)}$	$X_3^{(3)}$	$\tilde{X}_1^{(3)}$	$X_2^{(3)}$	$\tilde{X}_3^{(3)}$

Pooled data, with randomly swapped knockoffs

π_1	π_2	π_3

Empirical prior

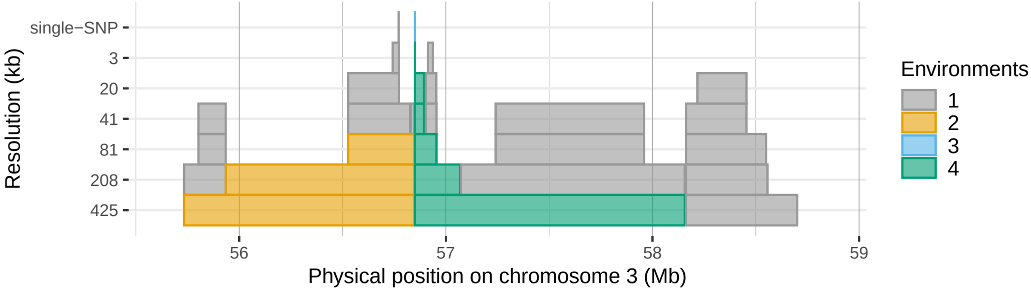
Partial Conjunction

- If some of the environments have small sample size, then power can be low.
- Weaker goal: find variables that are non-null in $\geq r$ environments

UK biobank data analysis

Environment	Sample size	Self-reported ancestries
African	7,623	"African", "Caribbean", "Any other black background", "Black or Black British"
Asian	3,284	"Asian or Asian British", "Chinese", "Any other Asian background"
British	429,934	"British"
European	28,994	"Any other white background", "Irish", "White"
Indian	7,628	"Indian", "Pakistani", "Bangladeshi"

UK biobank data analysis



References

Barber, R. F. and Candès, E. (2015). Controlling the false discovery rate via knockoffs. *Ann. Stat.*, 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. *J. R. Stat. Soc. B*, 80(3):551–577.

Thank you!