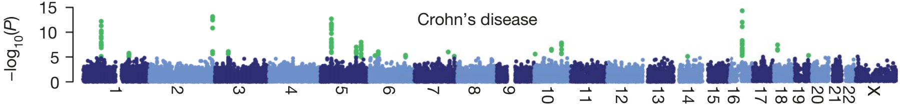
Sailing Through Data: Discoveries and Mirages

Emmanuel Candès, *Stanford University*

2018 Machine Learning Summer School, Buenos Aires, June 2018

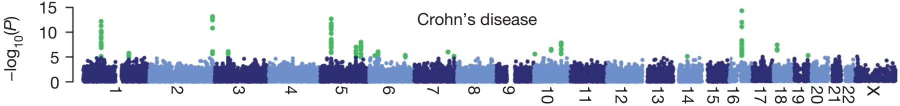
Controlled variable selection



- Response Y (e.g. disease status)
- Features $X_1, \ldots, X_p$ (e.g. SNPs)

Question: distribution of $Y \mid X$ depends on X through which variables?

Controlled variable selection



- Response Y (e.g. disease status)
- Features $X_1, \ldots, X_p$ (e.g. SNPs)

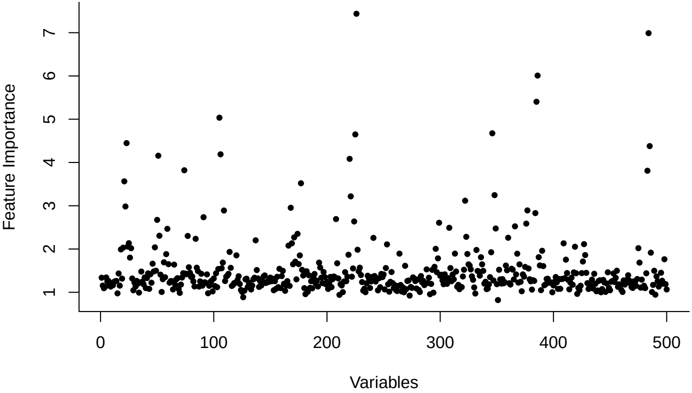
Question: distribution of $Y \mid X$ depends on X through which variables?

Goal: select set of features X_j that are likely to be relevant without too many false positives – do not run into the problem of irreproducibilty

$$\text{FDR} = \mathbb{E}\underbrace{\left[\frac{\#\text{ false positives}}{\#\text{ features selected}}\right]}_{\text{FDP}}$$

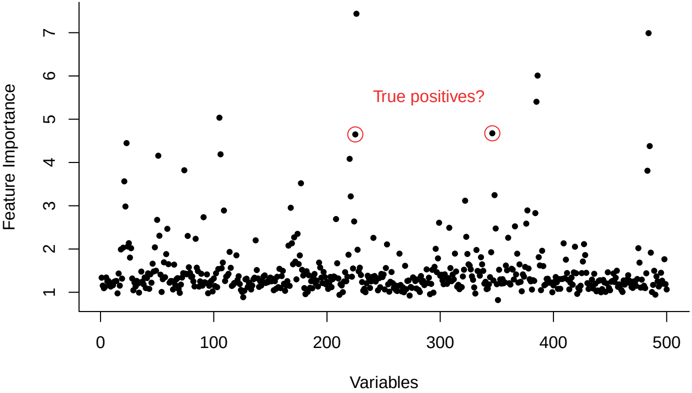
Which variables should we report?

Feature importance Z_j from random forests

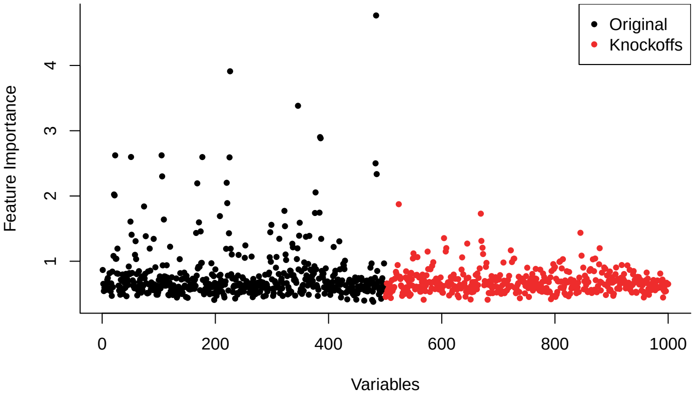


Which variables should we report?

Feature importance Z_j from random forests



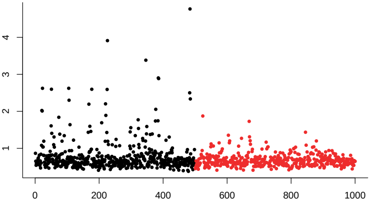
Knockoffs as negative controls



Exchangeability of feature importance statistics

Knockoff agnostic feature importance Z

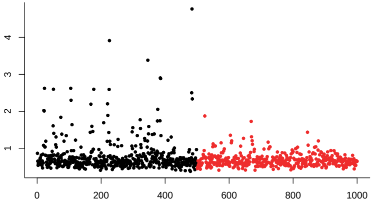
$$\underbrace{(Z_1, \ldots, Z_p}_{\text{originals}}, \underbrace{\tilde{Z}_1, \ldots, \tilde{Z}_p}_{\text{knockoffs}}) = z([\boldsymbol{X},\, \tilde{\boldsymbol{X}}],\, \boldsymbol{y})$$



Exchangeability of feature importance statistics

Knockoff agnostic feature importance Z

$$(\underbrace{Z_1,\dots,Z_p}_{\text{originals}},\underbrace{\tilde Z_1,\dots,\tilde Z_p}_{\text{knockoffs}}) = z([\boldsymbol{X},\,\tilde{\boldsymbol{X}}],\,\boldsymbol{y})$$

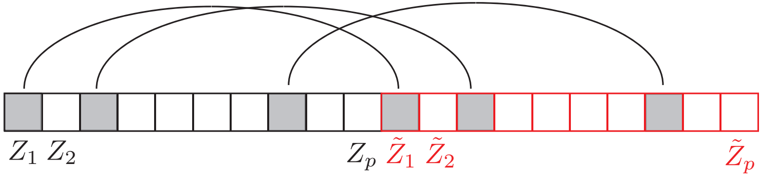


This lecture

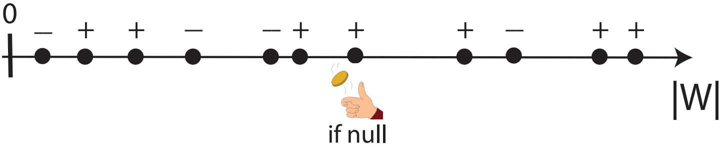
Can construct knockoff features such that

$$j \text{ null} \quad\Longrightarrow\quad (Z_j,\tilde Z_j) \overset{d}{=} (\tilde Z_j, Z_j)$$

$$\text{more generally}\quad \mathcal{T} \text{ subset of nulls} \quad\Longrightarrow\quad (Z,\tilde Z)_{\text{swap}(\mathcal{T})} \overset{d}{=} (Z,\tilde Z)$$



Knockoffs-adjusted scores



Ordering of variables + 1-bit p-values

Adjusted scores W_j with flip-sign property

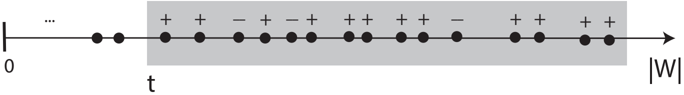
Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

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

e.g. $$W_j = Z_j - \tilde{Z}_j \qquad W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1 & Z_j > \tilde{Z}_j \\ -1 & Z_j \leq \tilde{Z}_j \end{cases}$$

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

Selection by sequential testing



Select $\mathcal{S}^+(t)$ $\Longrightarrow$ $\widehat{\mathrm{FDP}}(t) = \frac{1 + |\mathcal{S}^-(t)|}{1 \vee |\mathcal{S}^+(t)|}$

$$\mathcal{S}^+(t) = \{j : W_j \geq t\}$$
$$\mathcal{S}^-(t) = \{j : W_j \leq -t\}$$

Theorem (Barber and C. ('15))

Select $\mathcal{S}^+(\tau)$, $\tau = \min\{t : \widehat{\mathrm{FDP}}(t) \leq q\}$

- *Knockoff*

$$\mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections} + q^{-1}}\right] \leq q$$

- *Knockoff+*

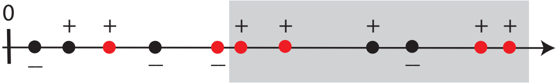
$$\mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections}}\right] \leq q$$

Some Pretty Math... (I Think)

Proof Sketch of FDR Control

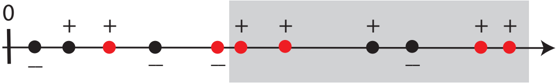
Why does all this work?

$$\tau = \min \left\{ t : \frac{1 + |\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q \right\} \qquad \begin{array}{l} \mathcal{S}^+(t) = \{j : W_j \geq t\} \\ \mathcal{S}^-(t) = \{j : W_j \leq -t\} \end{array}$$



Why does all this work?

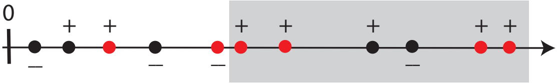
$$\tau = \text{min}\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \le q\right\} \qquad \begin{array}{l}\mathcal{S}^+(t) = \{j : W_j \ge t\} \\ \mathcal{S}^-(t) = \{j : W_j \le -t\}\end{array}$$



$$\text{FDP}(\tau) = \frac{\#\{j \text{ null} : j \in \mathcal{S}^+(\tau)\}}{\#\{j : j \in \mathcal{S}^+(\tau)\} \vee 1}$$

Why does all this work?

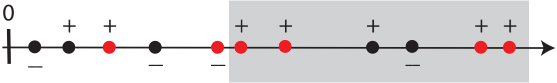
$$\tau = \text{min}\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\} \qquad \begin{array}{l} \mathcal{S}^+(t) = \{j : W_j \geq t\} \\ \mathcal{S}^-(t) = \{j : W_j \leq -t\} \end{array}$$



$$\text{FDP}(\tau) = \frac{\#\{j \text{ null} : j \in \mathcal{S}^+(\tau))\}}{\#\{j : j \in \mathcal{S}^+(\tau)\} \vee 1} \cdot \frac{1 + \#\{j \text{ null} : j \in \mathcal{S}^-(\tau)\}}{1 + \#\{j \text{ null} : j \in \mathcal{S}^-(\tau)\}}$$

Why does all this work?

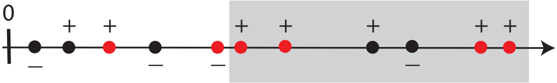
$$\tau = \mathsf{min}\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\} \qquad \begin{array}{l}\mathcal{S}^+(t) = \{j : W_j \geq t\}\\ \mathcal{S}^-(t) = \{j : W_j \leq -t\}\end{array}$$



$$\mathsf{FDP}(\tau) \leq q \cdot \frac{\overbrace{\#\{j \text{ null} : j \in \mathcal{S}^+(\tau)\}}^{V^+(\tau)}}{1 + \underbrace{\#\{j \text{ null} : j \in \mathcal{S}^-(\tau)\}}_{V^-(\tau)}}$$

Why does all this work?

$$\tau = \text{min}\left\{t : \frac{1+|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \le q\right\} \qquad \begin{array}{l} \mathcal{S}^+(t) = \{j : W_j \ge t\} \\ \mathcal{S}^-(t) = \{j : W_j \le -t\} \end{array}$$



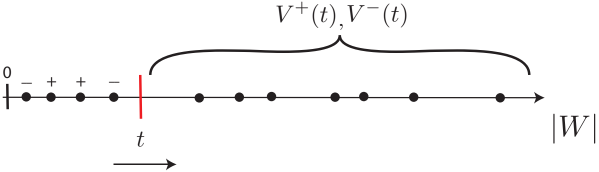
$$\text{FDP}(\tau) \le q \cdot \frac{\overbrace{\#\{j \text{ null} : j \in \mathcal{S}^+(\tau)\}}^{V^+(\tau)}}{1 + \underbrace{\#\{j \text{ null} : j \in \mathcal{S}^-(\tau)\}}_{V^-(\tau)}}$$

To show

$$\mathbb{E}\left[\frac{V^+(\tau)}{1 + V^-(\tau)}\right] \le 1$$

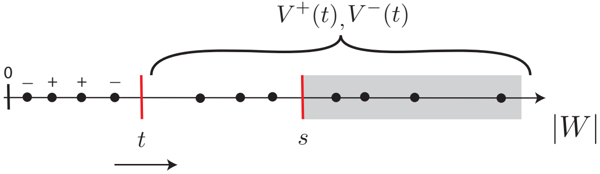
Martingales

$\dfrac{V^+(t)}{1+V^-(t)}$ is a (super)martingale wrt $\mathcal{F}_t = \{\sigma(V^\pm(u))\}_{u\le t}$



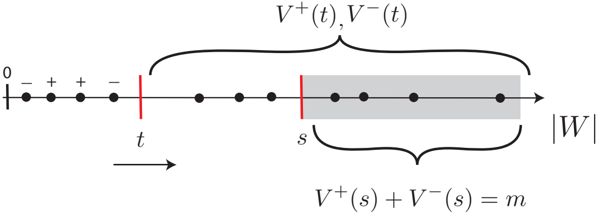
Martingales

$\dfrac{V^+(t)}{1+V^-(t)}$ is a (super)martingale wrt $\mathcal{F}_t = \{\sigma(V^\pm(u))\}_{u\le t}$



Martingales

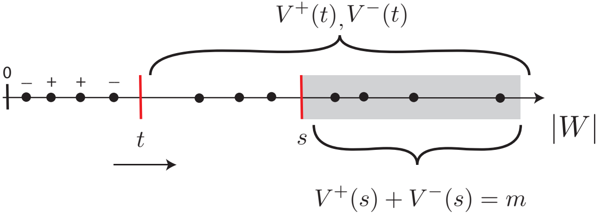
$\frac{V^+(t)}{1+V^-(t)}$ is a (super)martingale wrt $\mathcal{F}_t = \{\sigma(V^\pm(u))\}_{u \le t}$



Conditioned on $V^+(s) + V^-(s)$, $V^+(s)$ is hypergeometric

Martingales

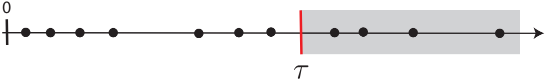
$$\frac{V^+(t)}{1+V^-(t)} \text{ is a (super)martingale wrt } \mathcal{F}_t = \{\sigma(V^\pm(u))\}_{u\le t}$$



Conditioned on $V^+(s)+V^-(s)$, $V^+(s)$ is hypergeometric

$$\mathbb{E}\left[\frac{V^+(s)}{1+V^-(s)}\,|\,V^\pm(t),\,V^+(s)+V^-(s)\right] \le \frac{V^+(t)}{1+V^-(t)}$$

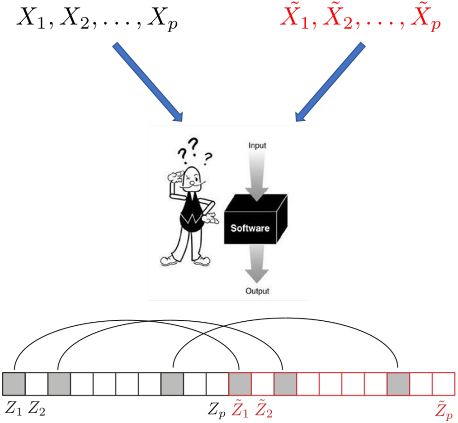
Optional stopping theorem



$$\mathsf{FDR} \le q\ \mathbb{E}\left[\frac{V^+(\tau)}{1+V^-(\tau)}\right] \le q\ \mathbb{E}\left[\frac{\overbrace{V^+(0)}^{\mathsf{Bin}(\#\mathsf{nulls},1/2)}}{1+V^-(0)}\right] \le q$$

Knockoffs for Random Features

Joint with Fan, Janson & Lv


$X_1, X_2, \ldots, X_p$
$\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_p$
Input
Software
Output
Z_1 Z_2
Z_p $\tilde{Z}_1$ $\tilde{Z}_2$
$\tilde{Z}_p$

Variable selection in arbitrary models

Random pair (X, Y) (perhaps thousands/millions of covariates)

$p(Y \mid X)$ depends on X through which variables?

Variable selection in arbitrary models

Random pair (X, Y) (perhaps thousands/millions of covariates)

$p(Y \mid X)$ depends on X through which variables?

Working definition of null variables

Say $j \in \mathcal{H}_0$ is null iff $Y \perp\!\!\!\perp X_j \mid X_{-j}$

Variable selection in arbitrary models

Random pair (X, Y) (perhaps thousands/millions of covariates)

$p(Y \mid X)$ depends on X through which variables?

Working definition of null variables

Say $j \in \mathcal{H}_0$ is null iff $Y \perp\!\!\!\perp X_j \mid X_{-j}$

Local Markov property $\Longrightarrow$ non nulls are smallest subset $\mathcal{S}$ (Markov blanket) s.t.

$$Y \perp\!\!\!\perp \{X_j\}_{j \in \mathcal{S}^c} \mid \{X_j\}_{j \in \mathcal{S}}$$

Variable selection in arbitrary models

Random pair (X, Y) (perhaps thousands/millions of covariates)

$p(Y \mid X)$ depends on X through which variables?

Working definition of null variables

Say $j \in \mathcal{H}_0$ is null iff $Y \perp\!\!\!\perp X_j \mid X_{-j}$

Local Markov property $\Longrightarrow$ non nulls are smallest subset $\mathcal{S}$ (Markov blanket) s.t.

$$Y \perp\!\!\!\perp \{X_j\}_{j \in \mathcal{S}^c} \mid \{X_j\}_{j \in \mathcal{S}}$$

Logistic model: $$\mathbb{P}(Y = 0|X) = \frac{1}{1 + e^{X^\top \beta}}$$

If variables $X_{1:p}$ are not perfectly dependent, then $j \in \mathcal{H}_0 \iff \beta_j = 0$

Knockoff features (random X)

i.i.d. samples from $p(X, Y)$

- Distribution of X known
- Distribution of $Y \mid X$ (likelihood) completely unknown

Knockoff features (random X)

i.i.d. samples from $p(X, Y)$

- Distribution of X known
- Distribution of $Y \mid X$ (likelihood) completely unknown

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

Knockoff features (random X)

i.i.d. samples from $p(X, Y)$

- Distribution of X known
- Distribution of $Y \mid X$ (likelihood) completely unknown

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

(1) Pairwise exchangeability

$$(X, \tilde{X})_{\mathsf{swap}(S)} \overset{d}{=} (X, \tilde{X})$$

e.g.

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\mathsf{swap}(\{2,3\})} \overset{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$

Knockoff features (random X)

i.i.d. samples from $p(X, Y)$

- Distribution of X known
- Distribution of $Y \mid X$ (likelihood) completely unknown

- Originals $X = (X_1, \dots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_p)$

(1) Pairwise exchangeability

$$(X,\, \tilde{X})_{\mathsf{swap}(S)} \overset{d}{=} (X,\, \tilde{X})$$

e.g.

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{\mathsf{swap}(\{2,3\})} \overset{d}{=} (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3)$$

(2) $\tilde{X} \perp\!\!\!\perp Y \mid X$ (ignore Y when constructing knockoffs)

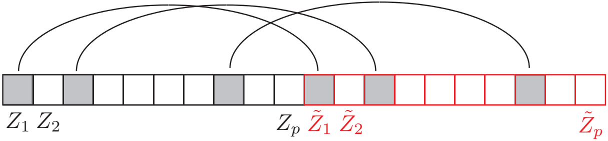
Exchangeability of feature importance statistics

Theorem (C., Fan, Janson Lv ('16))

For knockoff-agnostic scores and any subset $\mathcal{T}$ of nulls

$$(Z, Z)_{\textsf{swap}(\mathcal{T})} \overset{d}{=} (Z, \tilde{Z})$$

- *This holds no matter the relationship between Y and X*
- *This holds conditionally on Y*



Exchangeability of feature importance statistics

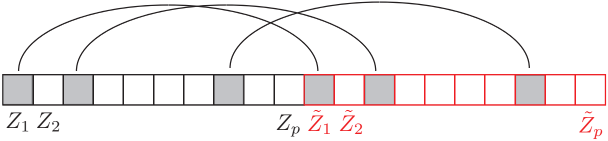
Theorem (C., Fan, Janson Lv ('16))

For knockoff-agnostic scores and any subset $\mathcal{T}$ of nulls

$$(Z, Z)_{\mathsf{swap}(\mathcal{T})} \overset{d}{=} (Z, \tilde{Z})$$

- *This holds no matter the relationship between Y and X*
- *This holds conditionally on Y*

$\Longrightarrow$ *FDR control (conditional on Y) no matter the relationship between X and Y*



Knockoffs for Gaussian features

Swapping any subset of original and knockoff features leaves (joint) dist. invariant

e.g. $\mathcal{T} = \{2,3\}$ $\quad (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$

Note $\tilde{X} \overset{d}{=} X$

Knockoffs for Gaussian features

Swapping any subset of original and knockoff features leaves (joint) dist. invariant

e.g. $\mathcal{T} = \{2, 3\}$ $\quad (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$

Note $\tilde{X} \overset{d}{=} X$

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Knockoffs for Gaussian features

Swapping any subset of original and knockoff features leaves (joint) dist. invariant

e.g. $\mathcal{T} = \{2, 3\}$ $\quad (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$

Note $\tilde{X} \overset{d}{=} X$

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **)$$

Knockoffs for Gaussian features

Swapping any subset of original and knockoff features leaves (joint) dist. invariant

e.g. $\mathcal{T} = \{2, 3\}$ $\qquad (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$

Note $\tilde{X} \overset{d}{=} X$

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

Knockoffs for Gaussian features

Swapping any subset of original and knockoff features leaves (joint) dist. invariant

e.g. $\mathcal{T} = \{2, 3\}$ $\quad (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$

Note $\tilde{X} \overset{d}{=} X$

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X,\, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

$\boldsymbol{s}$ such that $** \succeq 0$

Knockoffs for Gaussian features

Swapping any subset of original and knockoff features leaves (joint) dist. invariant

e.g. $\mathcal{T} = \{2, 3\}$ $\qquad (X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$

Note $\tilde{X} \overset{d}{=} X$

- $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Possible solution

$$(X, \tilde{X}) \sim \mathcal{N}(*, **) \qquad * = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad ** = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} \\ \boldsymbol{\Sigma} - \operatorname{diag}\{\boldsymbol{s}\} & \boldsymbol{\Sigma} \end{bmatrix}$$

  $\boldsymbol{s}$ such that $** \succeq 0$

- Given X, sample $\tilde{X}$ from $\tilde{X} \mid X$ (regression formula)

Different from knockoff features for fixed X!

Knockoffs inference with random features

Pros:

- No parameters
- No p-values

- Holds for finite samples
- No matter the dependence between Y and X
- No matter the dimensionality

Cons: Need to know distribution of covariates

Relationship with classical setup

Classical	MF Knockoffs

Relationship with classical setup

Classical	MF Knockoffs
Observations of X are fixed	Observations of X are random[1]
Inference is conditional on obs. values	

[1] Often appropriate in ‘big’ data apps: e.g. SNPs of subjects randomly sampled

Relationship with classical setup

Classical	MF Knockoffs
Observations of X are fixed Inference is conditional on obs. values	Observations of X are random^1
Strong model linking Y and X	Model free^2

1 Often appropriate in 'big' data apps: e.g. SNPs of subjects randomly sampled

2 Shifts the 'burden' of knowledge

Relationship with classical setup

Classical	MF Knockoffs
Observations of X are fixed Inference is conditional on obs. values	Observations of X are random^1
Strong model linking Y and X	Model free^2
Useful inference even if model inexact	Useful inference even if model inexact^3

1. Often appropriate in ‘big’ data apps: e.g. SNPs of subjects randomly sampled
2. Shifts the ‘burden’ of knowledge
3. More later

Shift in the burden of knowledge

When are our assumptions useful?

- When we have large amounts of unsupervised data (e.g. economic studies with same covariate info but different responses)
- When we have more prior information about the covariates than about their relationship with a response (e.g. GWAS)
- When we control the distribution of X (experimental crosses in genetics, gene knockout experiments,...)

Obstacles to obtaining p-values

$$Y \mid X \sim \text{Bernoulli}(\text{logit}(X^\top \beta))$$

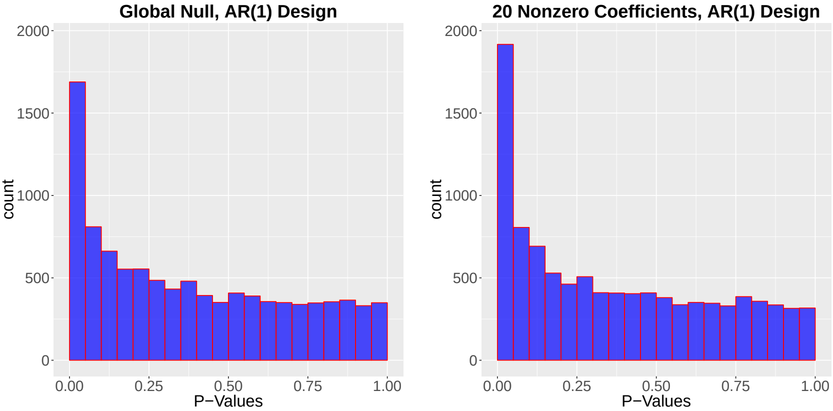


Figure: Distribution of null logistic regression p-values with $n = 500$ and $p = 200$

Obstacles to obtaining p-values

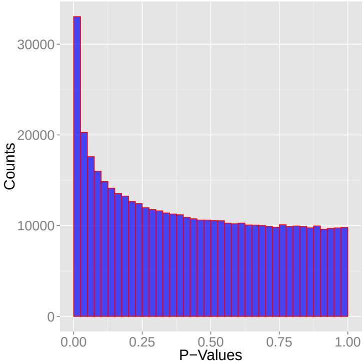
$\mathbb{P}\{p\text{-val} \leq \ldots\%\}$	Sett. (1)	Sett. (2)	Sett. (3)	Sett. (4)
5%	16.89% (0.37)	19.17% (0.39)	16.88% (0.37)	16.78% (0.37)
1%	6.78% (0.25)	8.49% (0.28)	7.02% (0.26)	7.03% (0.26)
0.1%	1.53% (0.12)	2.27% (0.15)	1.87% (0.14)	2.04% (0.14)

Table: Inflated p-value probabilities with estimated Monte Carlo SEs

Shameless plug: distribution of high-dimensional LRTs

Wilks' phenomenon (1938)

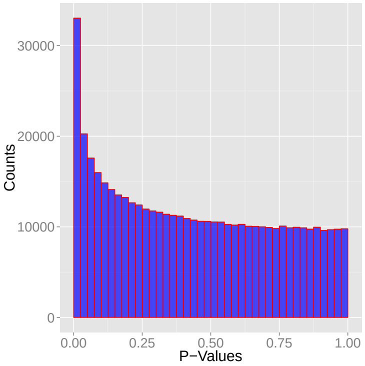
$$2 \log L \ \xrightarrow{\mathrm{d}} \ \chi^2_{\mathrm{df}}$$



Shameless plug: distribution of high-dimensional LRTs

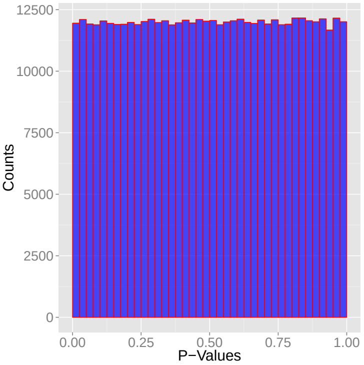
Wilks' phenomenon (1938)

$$2 \log L \overset{\mathrm{d}}{\to} \chi^2_{\mathrm{df}}$$



Sur, Chen, Candès (2017)

$$2 \log L \overset{\mathrm{d}}{\to} \kappa\left(\frac{p}{n}\right)\chi^2_{\mathrm{df}}$$



‘Low’ dim. linear model with dependent covariates

$$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\mathsf{CV}})|$$

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

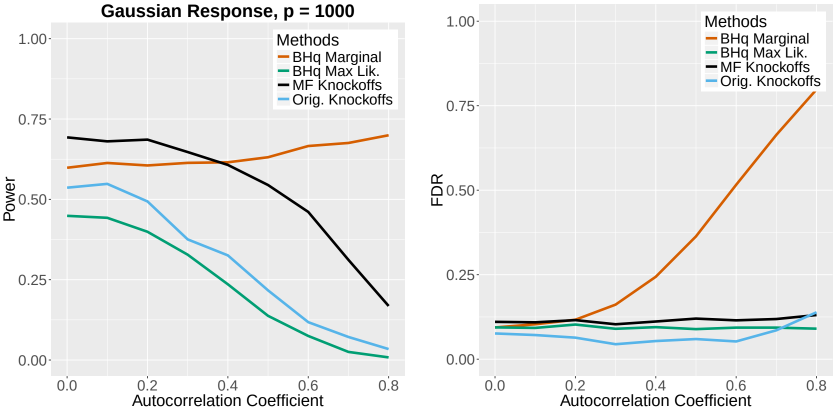


Figure: Low-dimensional setting: $n = 3000$, $p = 1000$

‘Low’ dim. logistic model with indep. covariates

$$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\mathsf{CV}})|$$
$$W_j = Z_j - \tilde{Z}_j$$

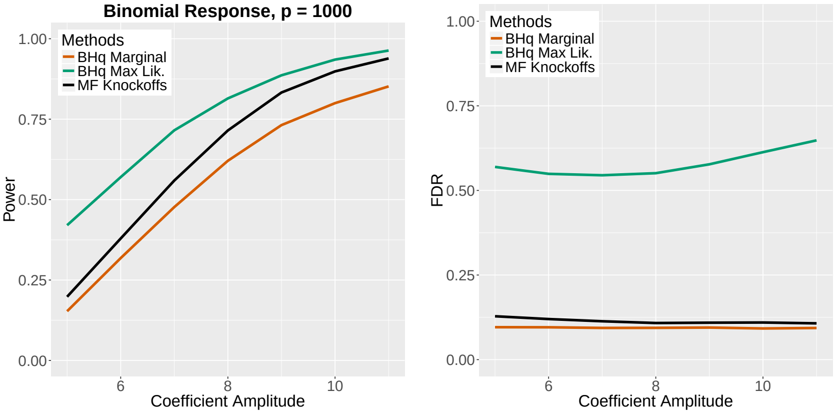


Figure: Low-dimensional setting: $n = 3000$, $p = 1000$

‘High’ dim. logistic model with dependent covariates

$$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\mathsf{CV}})|$$
$$W_j = Z_j - \tilde{Z}_j$$

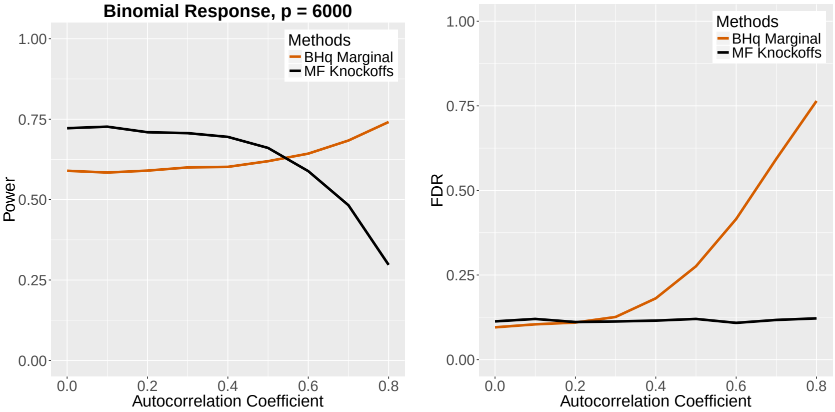


Figure: High-dimensional setting: $n = 3000$, $p = 6000$

Bayesian knockoff statistics

LCD (Lasso coeff. difference)

BVS (Bayesian variable selection)

$$Z_j = \mathbb{P}(\beta_j \neq 0 \,|\, \boldsymbol{y}, \boldsymbol{X})$$

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

Bayesian knockoff statistics

LCD (Lasso coeff. difference)

BVS (Bayesian variable selection)

$$Z_j = \mathbb{P}(\beta_j \neq 0 \,|\, \boldsymbol{y}, \boldsymbol{X})$$

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

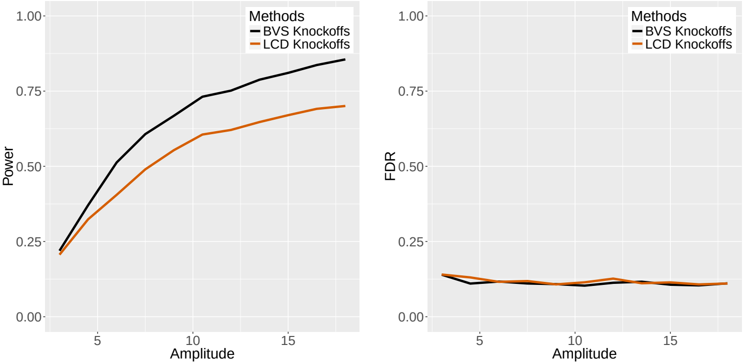


Figure: $n = 300$, $p = 1000$ and Bayesian linear model with 60 expected variables

Inference is correct even if prior is wrong or MCMC has not converged

Partial summary

- No valid p-values even for logistic regression
- Shifts the burden of knowledge to X (covariates); makes sense in many contexts
- Robustness: simulations show properties of inference hold even when the model for X is only approximately right.

  Always have access to these diagnostic checks (later)
- When assumptions are appropriate $\rightsquigarrow$ gain a lot of power, and can use sophisticated selection techniques

How to Construct Knockoffs for some Graphical Models

Joint with Sabatti & Sesia

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \ldots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \mid X_{\text{-}j}, \tilde{X}_{1:j-1}$
end
```

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \dots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \mid X_{\text{-}j}, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \ldots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \,|\, X_{\text{-}j}, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

- Sample $\tilde{X}_1$ from $X_1 \,|\, X_{-1}$

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

```
Algorithm Sequential Conditional Independent Pairs
for $j = \{1, \dots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \mid X_{\text{-}j}, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

- Sample $\tilde{X}_1$ from $X_1 \mid X_{-1}$
- Joint law of $X, \tilde{X}_1$ is known

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \ldots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \,|\, X_{\text{-}j}, \, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

- Sample $\tilde{X}_1$ from $X_1 \,|\, X_{-1}$
- Joint law of $X, \tilde{X}_1$ is known
- Sample $\tilde{X}_2$ from $X_2 \,|\, X_{-2}, \tilde{X}_1$

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \ldots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \,|\, X_{\text{-}j}, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

- Sample $\tilde{X}_1$ from $X_1 \,|\, X_{-1}$
- Joint law of $X, \tilde{X}_1$ is known
- Sample $\tilde{X}_2$ from $X_2 \,|\, X_{-2}, \tilde{X}_1$
- Joint law of $X, \tilde{X}_{1:2}$ is known

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \ldots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \,|\, X_{\text{-}j}, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

- Sample $\tilde{X}_1$ from $X_1 \,|\, X_{-1}$
- Joint law of $X, \tilde{X}_1$ is known
- Sample $\tilde{X}_2$ from $X_2 \,|\, X_{-2}, \tilde{X}_1$
- Joint law of $X, \tilde{X}_{1:2}$ is known
- Sample $\tilde{X}_3$ from $X_3 \,|\, X_{-3}, \tilde{X}_{1:2}$

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \ldots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \,|\, X_{\text{-}j},\, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

- Sample $\tilde{X}_1$ from $X_1 \,|\, X_{-1}$
- Joint law of $X, \tilde{X}_1$ is known
- Sample $\tilde{X}_2$ from $X_2 \,|\, X_{-2}, \tilde{X}_1$
- Joint law of $X, \tilde{X}_{1:2}$ is known
- Sample $\tilde{X}_3$ from $X_3 \,|\, X_{-3}, \tilde{X}_{1:2}$
- Joint law of $X, \tilde{X}$ is known and is pairwise exchangeable!

A general construction (C., Fan, Janson and Lv, '16)

$$(X_1, \tilde{X}_2, \tilde{X}_3, \tilde{X}_1, X_2, X_3) \overset{d}{=} (X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)$$

Algorithm Sequential Conditional Independent Pairs

```
for $j = \{1, \dots, p\}$ do
    Sample $\tilde{X}_j$ from law of $X_j \,|\, X_{\text{-}j}, \, \tilde{X}_{1:j-1}$
end
```

e.g. $p = 3$

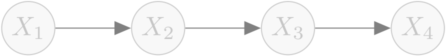
- Sample $\tilde{X}_1$ from $X_1 \,|\, X_{-1}$
- Joint law of $X, \tilde{X}_1$ is known
- Sample $\tilde{X}_2$ from $X_2 \,|\, X_{-2}, \tilde{X}_1$
- Joint law of $X, \tilde{X}_{1:2}$ is known
- Sample $\tilde{X}_3$ from $X_3 \,|\, X_{-3}, \tilde{X}_{1:2}$
- Joint law of $X, \tilde{X}$ is known and is pairwise exchangeable!

Usually not practical, easy in some cases (e.g. Markov chains)

Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

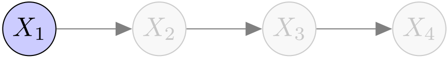
$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$



Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

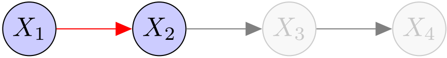
$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$



Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

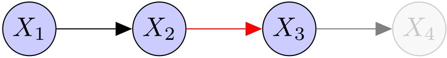
$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$



Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

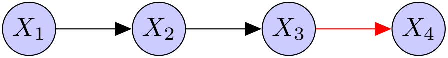
$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$



Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

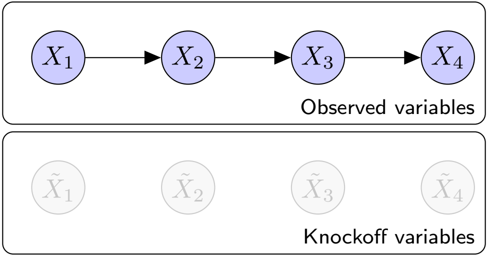
$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$



Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$

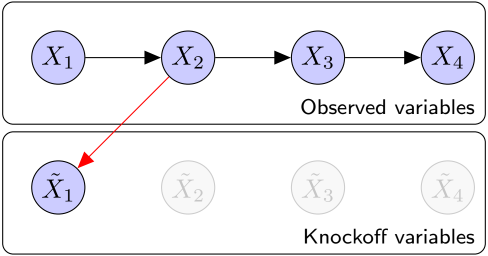


General algorithm can be implemented efficiently in the case of a Markov chain

Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$

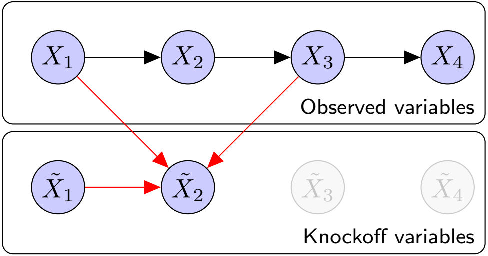


General algorithm can be implemented efficiently in the case of a Markov chain

Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$

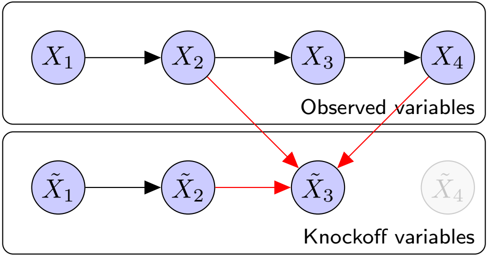


General algorithm can be implemented efficiently in the case of a Markov chain

Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$

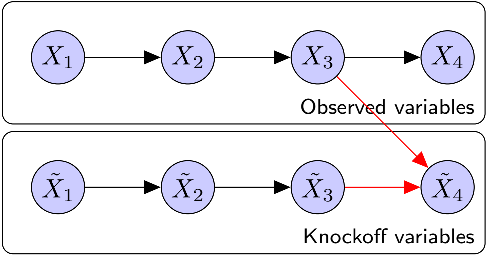


General algorithm can be implemented efficiently in the case of a Markov chain

Knockoff copies of a Markov chain

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$

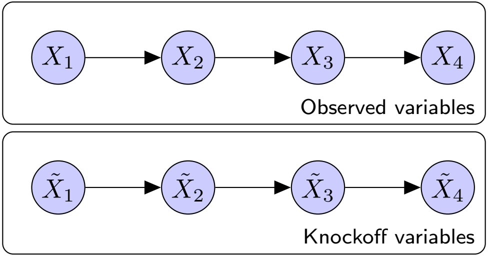


General algorithm can be implemented efficiently in the case of a Markov chain

Knockoff copies of a Markov chain

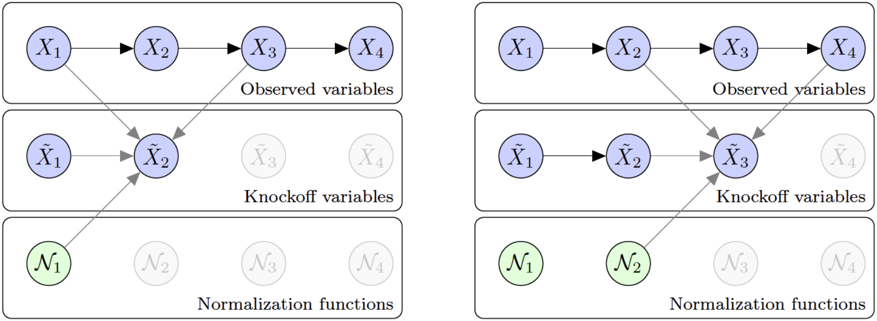
$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a Markov chain

$$p(X_1, \ldots, X_p) = q_1(X_1) \prod_{j=2}^{p} Q_j(X_j | X_{j-1}) \qquad (\boldsymbol{X} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}))$$



General algorithm can be implemented efficiently in the case of a Markov chain

Recursive update of normalizing constants



(a) Sampling $\tilde{X}_2$ at step $j = 2$.

(b) Sampling $\tilde{X}_3$ at step $j = 3$.

- Sampling $\tilde{X}_1$

$$p(X_1|X_{-1}) = p(X_1|X_2)$$

- Sampling $\tilde{X}_1$

$$p(X_1|X_{-1}) = p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

- Sampling $\tilde{X}_1$

$$p(X_1|X_{-1}) = p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{q_1(X_1)\,Q_2(X_2|X_1)}{Z_1(X_2)}$$

$$Z_1(z) = \sum_u q_1(u)\,Q_2(z|u)$$

- Sampling $\tilde{X}_1$

$$p(X_1|X_{-1}) = p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{q_1(X_1)\,Q_2(X_2|X_1)}{Z_1(X_2)}$$

$$Z_1(z) = \sum_u q_1(u)\,Q_2(z|u)$$

- Sampling $\tilde{X}_2$

$$p(X_2|X_{-2}, \tilde{X}_1) = p(X_2|X_1, X_3, \tilde{X}_1)$$

- Sampling $\tilde{X}_1$

$$p(X_1|X_{-1}) = p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{q_1(X_1)\,Q_2(X_2|X_1)}{Z_1(X_2)}$$

$$Z_1(z) = \sum_u q_1(u)\,Q_2(z|u)$$

- Sampling $\tilde{X}_2$

$$p(X_2|X_{-2}, \tilde{X}_1) = p(X_2|X_1, X_3, \tilde{X}_1) \propto Q_2(X_2|X_1)\,Q_3(X_3|X_2)\,\frac{Q_2(X_2|\tilde{X}_1)}{Z_1(X_2)}$$

- Sampling $\tilde{X}_1$

$$p(X_1|X_{-1}) = p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{q_1(X_1)\, Q_2(X_2|X_1)}{Z_1(X_2)}$$

$$Z_1(z) = \sum_u q_1(u)\, Q_2(z|u)$$

- Sampling $\tilde{X}_2$

$$p(X_2|X_{-2}, \tilde{X}_1) = p(X_2|X_1, X_3, \tilde{X}_1) \propto Q_2(X_2|X_1)\, Q_3(X_3|X_2)\, \frac{Q_2(X_2|\tilde{X}_1)}{Z_1(X_2)}$$

normalization constant $Z_2(X_3)$

$$Z_2(z) = \sum_u Q_2(u|X_1)\, Q_3(z|u)\, \frac{Q_2(u|\tilde{X}_1)}{Z_1(u)}$$

- Sampling $\tilde{X}_3$

$$p(X_3|X_{-3},\tilde{X}_1,\tilde{X}_2) = p(X_3|X_2,X_4,\tilde{X}_1,\tilde{X}_2)$$

- Sampling $\tilde{X}_3$

$$\begin{aligned} p(X_3|X_{-3},\tilde{X}_1,\tilde{X}_2) &= p(X_3|X_2,X_4,\tilde{X}_1,\tilde{X}_2) \\ &\propto Q_3(X_3|X_2)\,Q_4(X_4|X_3)\,\frac{Q_3(X_3|\tilde{X}_2)}{Z_2(X_3)} \end{aligned}$$

- Sampling $\tilde{X}_3$

$$\begin{aligned} p(X_3|X_{-3},\tilde{X}_1,\tilde{X}_2) &= p(X_3|X_2,X_4,\tilde{X}_1,\tilde{X}_2) \\ &\propto Q_3(X_3|X_2)\,Q_4(X_4|X_3)\,\frac{Q_3(X_3|\tilde{X}_2)}{Z_2(X_3)} \end{aligned}$$

normalization constant $Z_3(X_4)$

$$Z_3(z) = \sum_u Q_3(u|X_2)\,Q_4(z|u)\,\frac{Q_3(u|\tilde{X}_2)}{Z_2(u)}$$

- Sampling $\tilde{X}_3$

$$p(X_3|X_{-3},\tilde{X}_1,\tilde{X}_2) = p(X_3|X_2,X_4,\tilde{X}_1,\tilde{X}_2)$$

$$\propto Q_3(X_3|X_2)\,Q_4(X_4|X_3)\,\frac{Q_3(X_3|\tilde{X}_2)}{Z_2(X_3)}$$

normalization constant $Z_3(X_4)$

$$Z_3(z) = \sum_u Q_3(u|X_2)\,Q_4(z|u)\,\frac{Q_3(u|\tilde{X}_2)}{Z_2(u)}$$

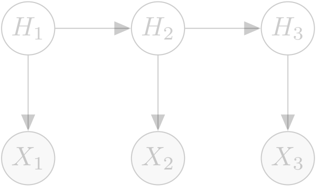
- And so on sampling $\tilde{X}_j$ ...

Computationally efficient $O(p)$

Hidden Markov Models (HMMs)

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a HMM if

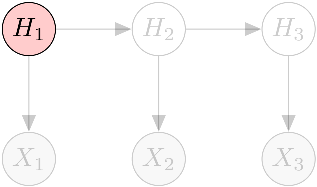
$$\begin{cases} \boldsymbol{H} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}) & \text{(latent Markov chain)} \\ X_j|\boldsymbol{H} \sim X_j|H_j \overset{\text{ind.}}{\sim} f_j(X_j; H_j) & \text{(emission distribution)} \end{cases}$$



Hidden Markov Models (HMMs)

$\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ is a HMM if

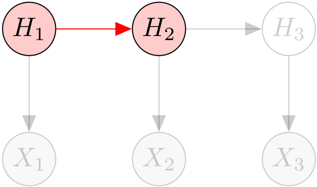
$$\begin{cases} \boldsymbol{H} \sim \text{MC}\,(q_1, \boldsymbol{Q}) & \text{(latent Markov chain)} \\ X_j | \boldsymbol{H} \sim X_j | H_j \overset{\text{ind.}}{\sim} f_j(X_j; H_j) & \text{(emission distribution)} \end{cases}$$



Hidden Markov Models (HMMs)

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a HMM if

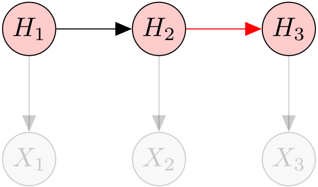
$$\begin{cases} \boldsymbol{H} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}) & \text{(latent Markov chain)} \\ X_j | \boldsymbol{H} \sim X_j | H_j \overset{\text{ind.}}{\sim} f_j(X_j; H_j) & \text{(emission distribution)} \end{cases}$$



Hidden Markov Models (HMMs)

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a HMM if

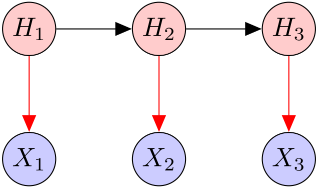
$$\begin{cases} \boldsymbol{H} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}) & \text{(latent Markov chain)} \\ X_j|\boldsymbol{H} \sim X_j|H_j \overset{\text{ind.}}{\sim} f_j(X_j; H_j) & \text{(emission distribution)} \end{cases}$$



Hidden Markov Models (HMMs)

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a HMM if

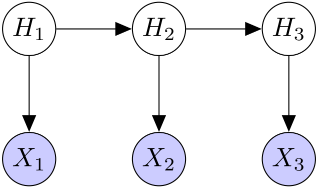
$$\begin{cases} \boldsymbol{H} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}) & \text{(latent Markov chain)} \\ X_j | \boldsymbol{H} \sim X_j | H_j \overset{\text{ind.}}{\sim} f_j(X_j; H_j) & \text{(emission distribution)} \end{cases}$$



Hidden Markov Models (HMMs)

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is a HMM if

$$\begin{cases} \boldsymbol{H} \sim \mathsf{MC}\,(q_1, \boldsymbol{Q}) & \text{(latent Markov chain)} \\ X_j|\boldsymbol{H} \sim X_j|H_j \overset{\text{ind.}}{\sim} f_j(X_j; H_j) & \text{(emission distribution)} \end{cases}$$



The $\boldsymbol{H}$ variables are latent and only the $\boldsymbol{X}$ variables are observed

Haplotypes and genotypes

Haplotype Set of alleles on a single chromosome
0/1 for common/rare allele

Genotype Unordered pair of alleles at a single marker

$$\begin{array}{rcl} & 0\;1\;0\;1\;1\;0 & \text{Haplotype M} \\ + & 1\;1\;0\;0\;1\;1 & \text{Haplotype P} \\ \hline & 1\;2\;0\;1\;2\;1 & \text{Genotypes} \end{array}$$

A phenomenological HMM for haplotype & genotype data

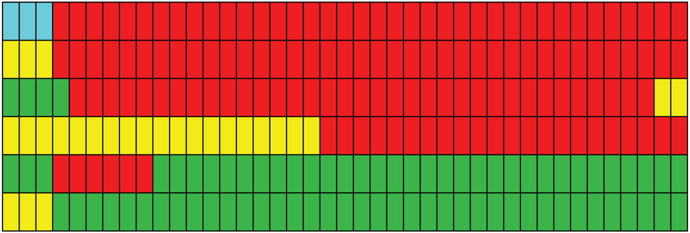


Figure: Six haplotypes: color indicates ‘ancestor’ at each marker (Scheet, ’06)

A phenomenological HMM for haplotype & genotype data

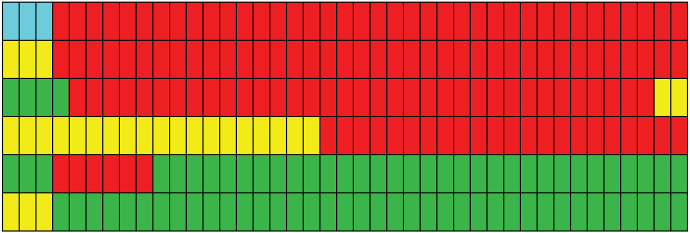


Figure: Six haplotypes: color indicates 'ancestor' at each marker (Scheet, '06)

Haplotype estimation/phasing (Browning, '11)
Imputation of missing SNPs (Marchini, '10)

- `fastPHASE` (Scheet, '06)

- `IMPUTE` (Marchini, '07)
- `MaCH` (Li, '10)

A phenomenological HMM for haplotype & genotype data

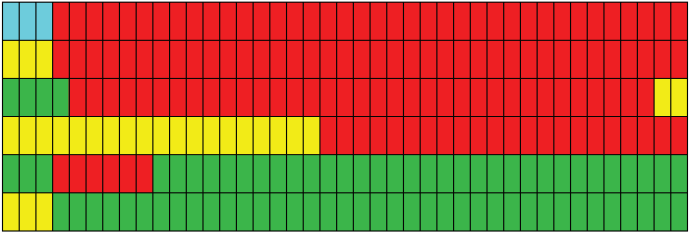


Figure: Six haplotypes: color indicates ‘ancestor’ at each marker (Scheet, ’06)

Haplotype estimation/phasing (Browning, ’11)
Imputation of missing SNPs (Marchini, ’10)

- `fastPHASE` (Scheet, ’06)

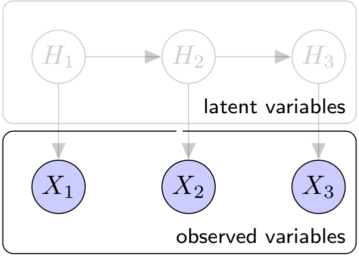
- `IMPUTE` (Marchini, ’07)
- `MaCH` (Li, ’10)

New application of same HMM: generation of knockoff copies of genotypes!
Each genotype: sum of two independent HMM haplotype sequences

Knockoff copies of a hidden Markov model

Theorem (Sesia, Sabatti, C. '17)

A knockoff copy of $\tilde{\boldsymbol{X}}$ of $\boldsymbol{X}$ can be constructed as

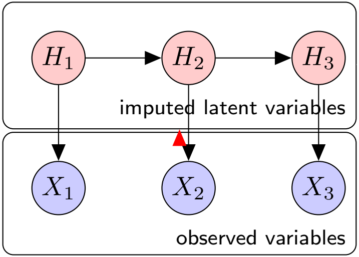


Knockoff copies of a hidden Markov model

Theorem (Sesia, Sabatti, C. '17)

A knockoff copy of $\tilde{\boldsymbol{X}}$ of $\boldsymbol{X}$ can be constructed as

(1) *Sample $\boldsymbol{H}$ from $p(\boldsymbol{H}|\boldsymbol{X})$ using forward-backward algorithm*



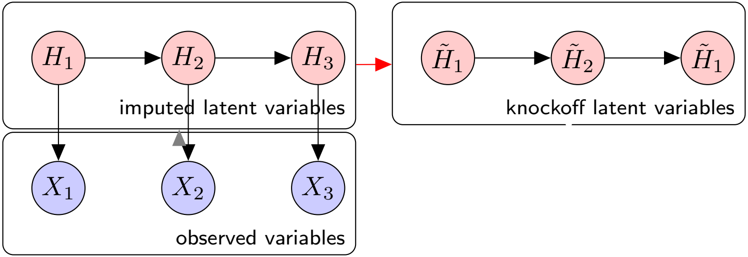
Knockoff copies of a hidden Markov model

Theorem (Sesia, Sabatti, C. '17)

A knockoff copy of $\tilde{\boldsymbol{X}}$ of $\boldsymbol{X}$ can be constructed as

(1) *Sample $\boldsymbol{H}$ from $p(\boldsymbol{H}|\boldsymbol{X})$ using forward-backward algorithm*

(2) *Generate a knockoff $\tilde{\boldsymbol{H}}$ of $\boldsymbol{H}$ using the SCIP algorithm for a Markov chain*



Knockoff copies of a hidden Markov model

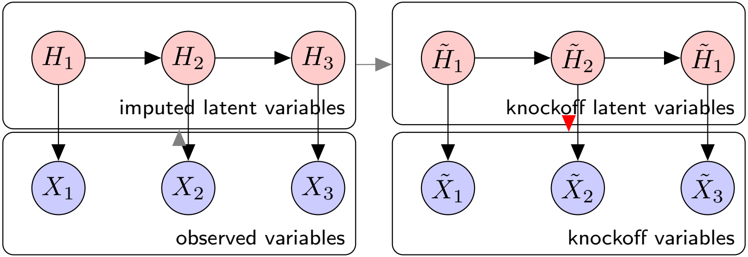
Theorem (Sesia, Sabatti, C. '17)

A knockoff copy of $\tilde{\boldsymbol{X}}$ of $\boldsymbol{X}$ can be constructed as

(1) *Sample $\boldsymbol{H}$ from $p(\boldsymbol{H}|\boldsymbol{X})$ using forward-backward algorithm*

(2) *Generate a knockoff $\tilde{\boldsymbol{H}}$ of $\boldsymbol{H}$ using the SCIP algorithm for a Markov chain*

(3) *Sample $\tilde{\boldsymbol{X}}$ from the emission distribution of $\boldsymbol{X}$ given $\boldsymbol{H} = \tilde{\boldsymbol{H}}$*



Some Examples

Simulations with synthetic Markov chain

Markov chain covariates with 5 hidden states. Binomial response

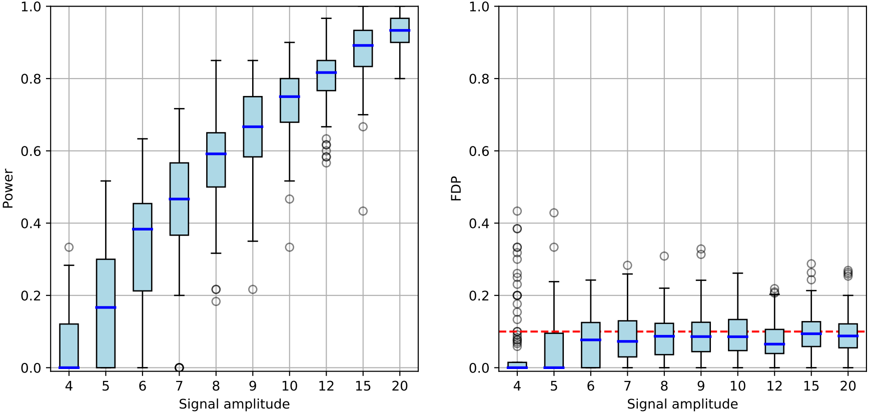


Figure: Power and FDP over 100 repetitions (true F_X)

$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$

$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Robustness

Markov chain covariates with 5 hidden states. Binomial response

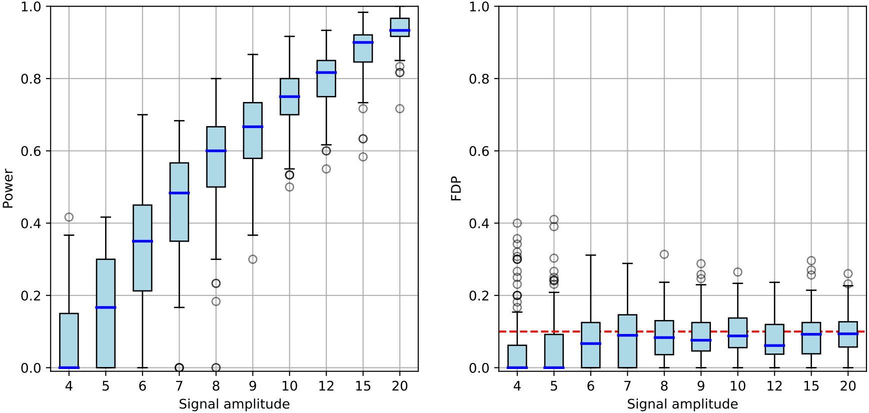


Figure: Power and FDP over 100 repetitions (estimated F_X)

$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$

$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\mathrm{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Simulations with synthetic HMM

HMM covariates with latent “clockwise” Markov chain. Binomial response

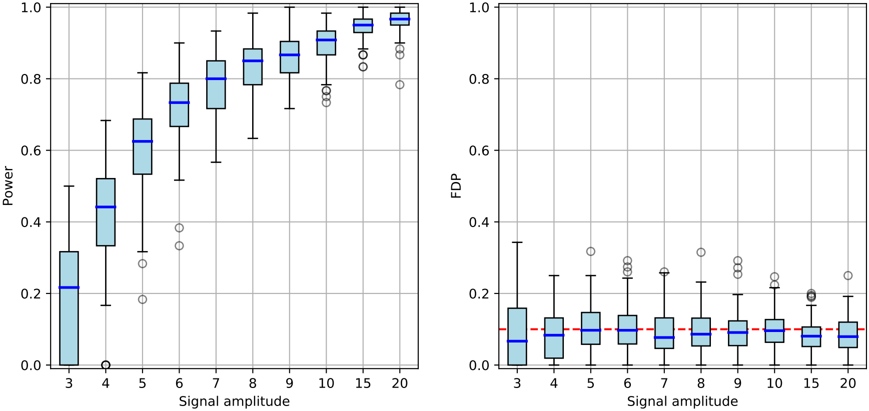


Figure: Power and FDP over 100 repetitions (true F_X)
$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$
$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Robustness

HMM covariates with latent “clockwise” Markov chain. Binomial response

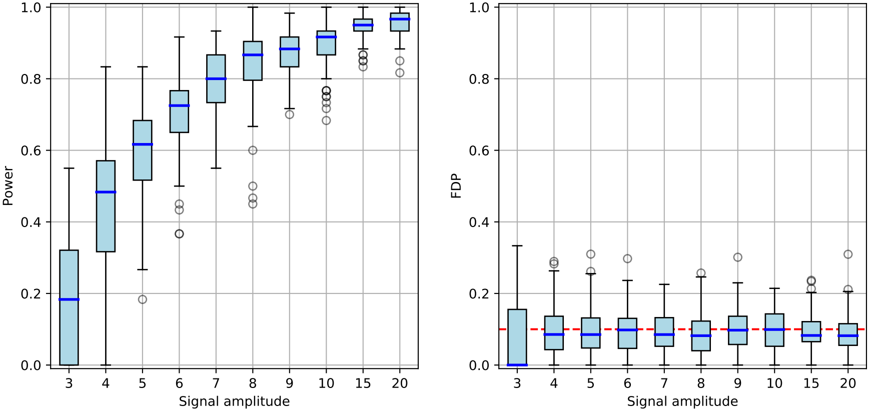


Figure: Power and FDP over 100 repetitions (estimated F_X)
$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$
$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Out-of-sample parameter estimation

Inhomogeneous Markov chain covariates with 5 hidden states. Binomial response

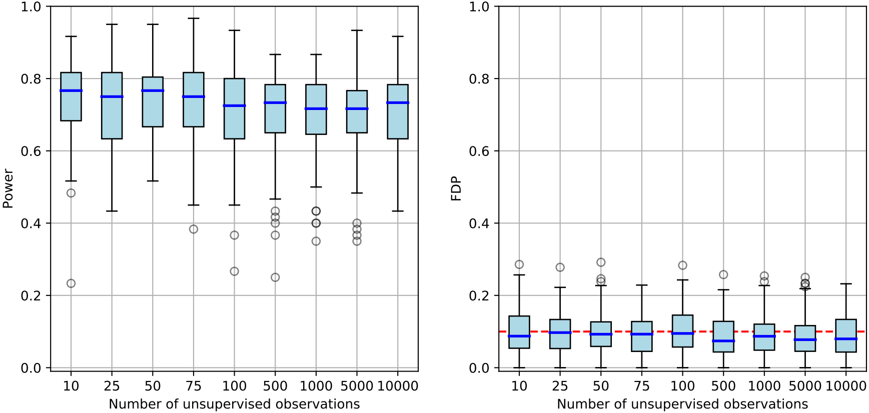


Figure: Power and FDP over 100 repetitions (estimated F_X from independent dataset)
$n = 1000, p = 1000$, target FDR: $\alpha = 0.1$
$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$

Genetic Data Analysis

Genetic analysis

Crohn's disease (CD)

- Wellcome Trust Case Control Consortium (WTCCC)
- $n \approx 5,000$ subjects ($\approx 2,000$ patients, $\approx 3,000$ healthy controls)
- $p \approx 400,000$ SNPs
- Previously analyzed in WTCCC (2007)

Genetic analysis

Crohn's disease (CD)

- Wellcome Trust Case Control Consortium (WTCCC)
- $n \approx 5,000$ subjects ($\approx 2,000$ patients, $\approx 3,000$ healthy controls)
- $p \approx 400,000$ SNPs
- Previously analyzed in WTCCC (2007)

Lipid traits (HDL, LDL cholesterol)

- Northern Finland 1966 Birth Cohort study of metabolic syndrome (NFBC)
- $n \approx 4,700$ subjects
- $p \approx 330,000$ SNPs
- Previously analyzed in Sabatti et al. (2009)

High-level results

Knockoffs with nominal FDR level of 10%

High-level results

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

High-level results

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)

High-level results

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)
- Knockoffs made a number of new discoveries

High-level results

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)

- Knockoffs made a number of new discoveries
    - Expect some (roughly 10%) of these to be false discoveries

High-level results

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)
- Knockoffs made a number of new discoveries
  - Expect some (roughly 10%) of these to be false discoveries
  - It is likely that many of these correspond to true discoveries

High-level results

Knockoffs with nominal FDR level of 10%

- Power is much higher:

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

- Quite a few of the discoveries made by knockoffs were confirmed by larger GWAS (Franke et al., '10, Willer et al., '13)
- Knockoffs made a number of new discoveries
  - Expect some (roughly 10%) of these to be false discoveries
  - It is likely that many of these correspond to true discoveries
  - Evidence from independent studies about adjacent genes shows some of the top unconfirmed hits to be promising candidates

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Franke et al. '10	WTCCC '07
100%	rs11209026 (2)	1	67.31–67.42	yes	yes
99%	rs6431654 (20)	2	233.94–234.11	yes	yes
98%	rs6688532 (33)	1	169.4–169.65		yes
97%	rs17234657 (1)	5	40.44–40.44	yes	yes
95%	rs11805303 (16)	1	67.31–67.46	yes	yes
91%	rs7095491 (18)	10	101.26–101.32	yes	yes
91%	rs3135503 (16)	16	49.28–49.36	yes	yes
81%	rs7768538 (1145)	6	25.19–32.91	yes	yes
80%	rs6601764 (1)	10	3.85–3.85		yes
75%	rs7655059 (5)	4	89.5–89.53		
73%	rs6500315 (4)	16	49.03–49.07	yes	yes
72%	rs2738758 (5)	20	61.71–61.82	yes	
70%	rs7726744 (46)	5	40.35–40.71	yes	yes
68%	rs11627513 (7)	14	96.61–96.63		
66%	rs4246045 (46)	5	150.07–150.41	yes	yes
62%	rs9783122 (234)	10	106.43–107.61		
61%	rs6825958 (3)	4	55.73–55.77		

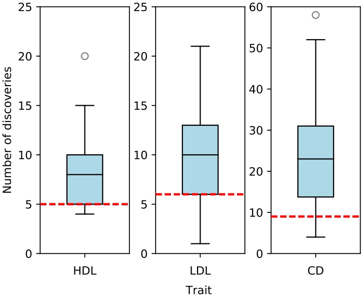
Table: SNP clusters found to be important for CD over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
100%	rs1532085 (4)	15	58.68–58.7	yes	yes
100%	rs7499892 (1)	16	57.01–57.01	yes	yes
100%	rs1800961 (1)	20	43.04–43.04	yes	
99%	rs1532624 (2)	16	56.99–57.01	yes	yes
95%	rs255049 (142)	16	66.41–69.41	yes	yes

Table: SNP clusters found to be important for HDL over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
99%	rs4844614 (34)	1	207.3–207.88		yes
97%	rs646776 (5)	1	109.8–109.82	yes	yes
97%	rs2228671 (2)	19	11.2–11.21	yes	yes
94%	rs157580 (4)	19	45.4–45.41	yes	yes
92%	rs557435 (21)	1	55.52–55.72	yes	
80%	rs10198175 (1)	2	21.13–21.13	yes	yes
76%	rs10953541 (58)	7	106.48–107.3		
62%	rs6575501 (1)	14	95.64–95.64		

Table: SNP clusters found to be important for LDL over 100 repetitions of knockoffs.



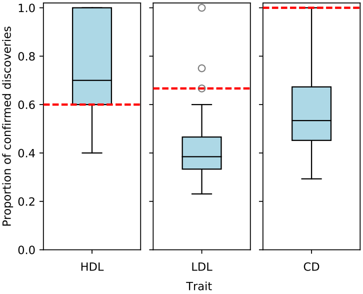


Figure: Number of discoveries made on different GWAS datasets (left) and proportion of discoveries confirmed by a meta-analysis (right). Red lines correspond to results published in papers that first analyzed our datasets

Data analysis issues

(1) Estimate distribution of SNPs (HMM) to build knockoffs

(2) Highly correlated SNPs

Data analysis issues

(1) Estimate distribution of SNPs (HMM) to build knockoffs

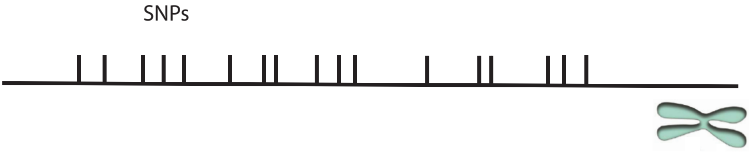
(2) Highly correlated SNPs

(1) Estimating the HMM

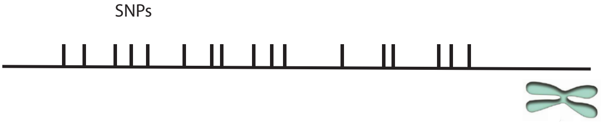
- Methodology of Scheet and Stephens '06
- Fitted with `fastPHASE` (EM), $K \approx 10$ possible hidden states
- For each individual, making a knockoff copy of 70,000 SNPs takes about 1.3 sec on Intel Xeon CPU (2.6GHz) (after parameter estimation)

Highly correlated SNPs

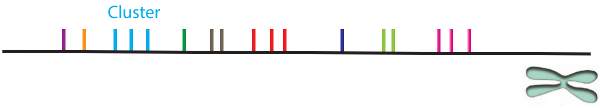
Hard to choose between two or more nearly-identical variables if the data supports at least one of them being selected



Clustering



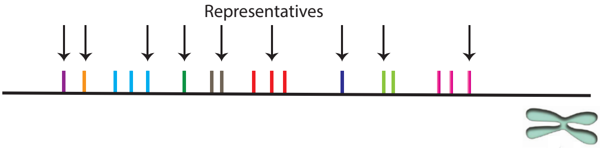
Clustering



- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5

  ⇝ settle for discovering important SNP clusters among 71,145 candidates for CD and 59,005 for cholesterol

Clustering



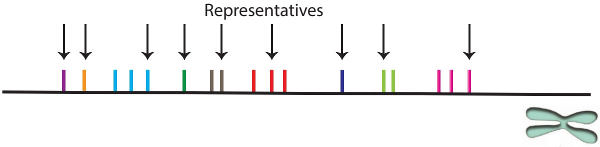
- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5

  $\rightsquigarrow$ settle for discovering important SNP clusters among
  71,145 candidates for CD and 59,005 for cholesterol

- Cluster variables? Choose a representative SNP from each cluster (see also Reid and Tibshirani, '15)

  $\rightsquigarrow$ approximate null: cluster rep $\perp\!\!\!\perp Y \,|\,$ other reps

Clustering



- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5

  $\rightsquigarrow$ settle for discovering important SNP clusters among

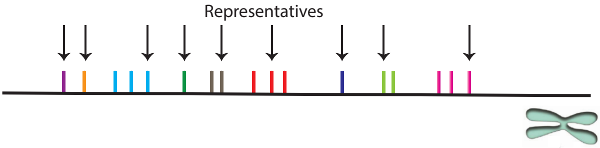
  71,145 candidates for CD and 59,005 for cholesterol

- Cluster variables? Choose a representative SNP from each cluster (see also Reid and Tibshirani, '15)

  $\rightsquigarrow$ approximate null: cluster rep $\perp\!\!\!\perp Y \,|\,$ other reps

- Which rep? Most significant SNP as computed on 20% of the samples

Clustering



- Cluster SNPs using estimated correlations as similarity measure and single-linkage cutoff of 0.5

  $\rightsquigarrow$ settle for discovering important SNP clusters among

  71,145 candidates for CD and 59,005 for cholesterol

- Cluster variables? Choose a representative SNP from each cluster (see also Reid and Tibshirani, '15)

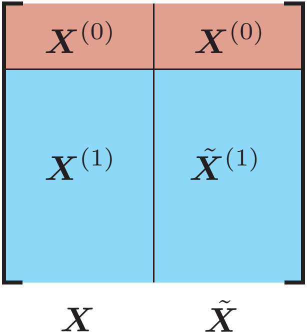
  $\rightsquigarrow$ approximate null: cluster rep $\perp\!\!\!\perp Y \,|\,$ other reps

- Which rep? Most significant SNP as computed on 20% of the samples

- Safe data re-use (optimize power) as in Barber and C. (16)

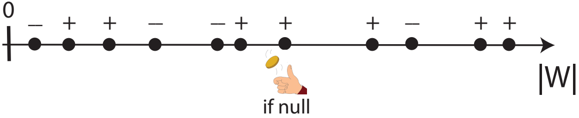
Safe data re-use

We used an independent split of the data
to select representative SNPs



🟥 Used for selecting reps and safely re-used for inference

🟦 Used only for inference



Re-use data to improve ordering but not to compute signs (1-bit p-values)

Simulations with genetic covariates

- Real genetic covariates X
- Logistic conditional model $Y \mid X$ with 60 variables

Simulations with genetic covariates

- Real genetic covariates X
- Logistic conditional model $Y \mid X$ with 60 variables

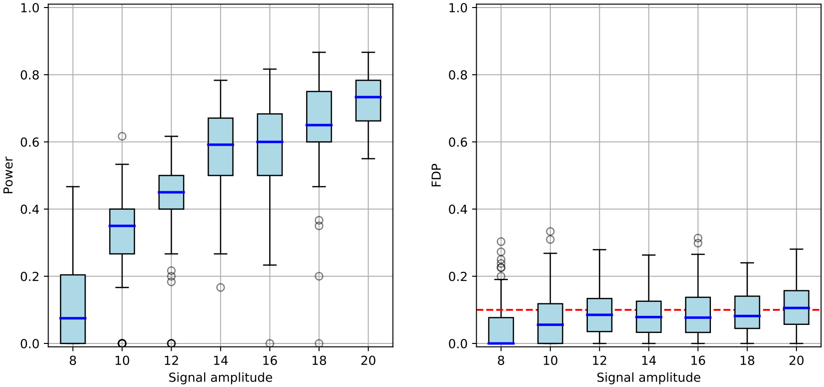
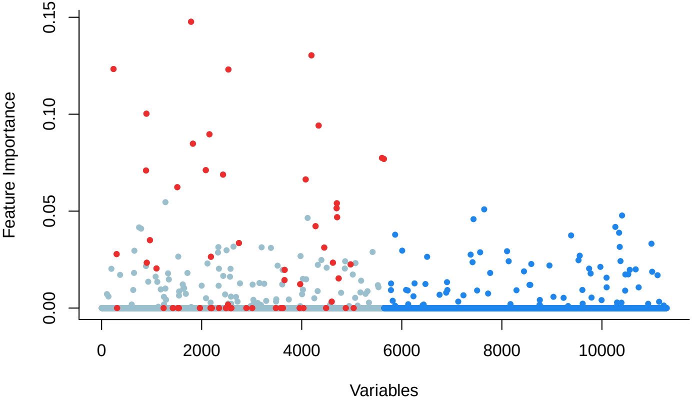


Figure: Power and FDP over 100 repetitions
$Z_j = |\hat{\beta}_j(\hat{\lambda}_{\text{CV}})|$, $W_j = Z_j - \tilde{Z}_j$, target FDR: $\alpha = 0.1$

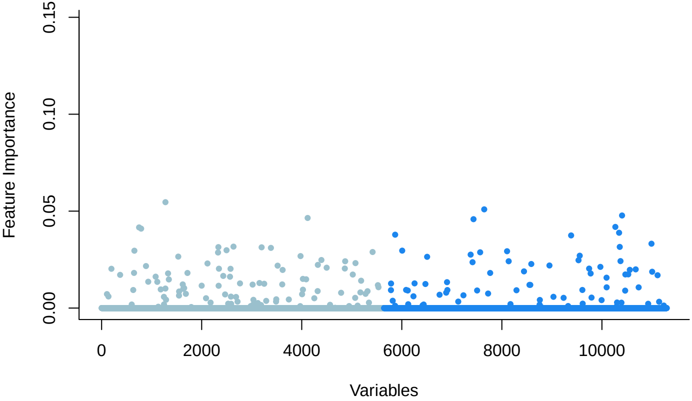
Diagnostic plot: simulation with data from Chromosome 1

Feature importance $Z_j = |\hat{\beta}_j(\lambda_{\mathsf{CV}})|$



Diagnostic plot: simulation with data from Chromosome 1

Feature importance $Z_j = |\hat{\beta}_j(\lambda_{\mathsf{CV}})|$



Results of data analysis

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Franke et al. '10	WTCCC '07
100%	rs11209026 (2)	1	67.31–67.42	yes	yes
99%	rs6431654 (20)	2	233.94–234.11	yes	yes
98%	rs6688532 (33)	1	169.4–169.65		yes
97%	rs17234657 (1)	5	40.44–40.44	yes	yes
95%	rs11805303 (16)	1	67.31–67.46	yes	yes
91%	rs7095491 (18)	10	101.26–101.32	yes	yes
91%	rs3135503 (16)	16	49.28–49.36	yes	yes
81%	rs7768538 (1145)	6	25.19–32.91	yes	yes
80%	rs6601764 (1)	10	3.85–3.85		yes
75%	rs7655059 (5)	4	89.5–89.53		
73%	rs6500315 (4)	16	49.03–49.07	yes	yes
72%	rs2738758 (5)	20	61.71–61.82	yes	
70%	rs7726744 (46)	5	40.35–40.71	yes	yes
68%	rs11627513 (7)	14	96.61–96.63		
66%	rs4246045 (46)	5	150.07–150.41	yes	yes
62%	rs9783122 (234)	10	106.43–107.61		
61%	rs6825958 (3)	4	55.73–55.77		

Table: SNP clusters found to be important for CD over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
100%	rs1532085 (4)	15	58.68–58.7	yes	yes
100%	rs7499892 (1)	16	57.01–57.01	yes	yes
100%	rs1800961 (1)	20	43.04–43.04	yes	
99%	rs1532624 (2)	16	56.99–57.01	yes	yes
95%	rs255049 (142)	16	66.41–69.41	yes	yes

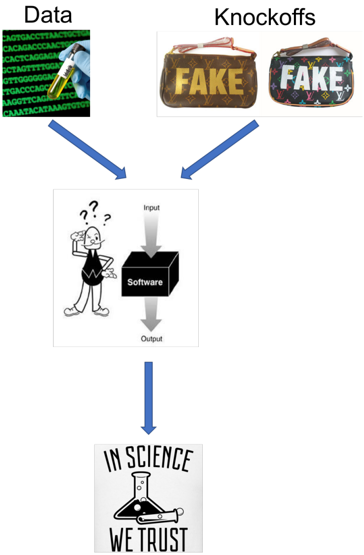
Table: SNP clusters found to be important for HDL over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
99%	rs4844614 (34)	1	207.3–207.88		yes
97%	rs646776 (5)	1	109.8–109.82	yes	yes
97%	rs2228671 (2)	19	11.2–11.21	yes	yes
94%	rs157580 (4)	19	45.4–45.41	yes	yes
92%	rs557435 (21)	1	55.52–55.72	yes	
80%	rs10198175 (1)	2	21.13–21.13	yes	yes
76%	rs10953541 (58)	7	106.48–107.3		
62%	rs6575501 (1)	14	95.64–95.64		

Table: SNP clusters found to be important for LDL over 100 repetitions of knockoffs.

Summary and open questions

- Knockoffs offers finite sample inferential properties in subtle and important problems
- Knockoffs is a powerful, flexible, and robust solution whenever there is considerable outside information on dist. of X such as GWAS
- Knockoffs addresses the replicability issue
- Where is the burden of knowledge?



Data

Knockoffs

Input

Software

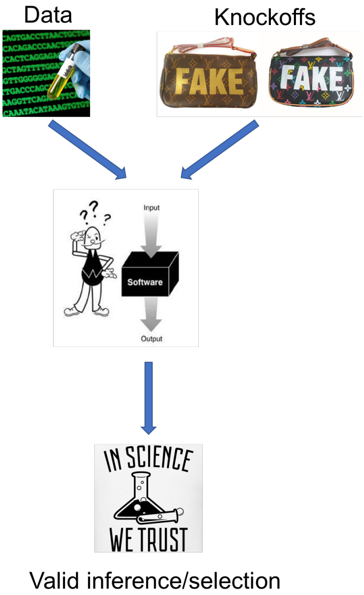
Output

Valid inference/selection

Summary and open questions

- Knockoffs offers finite sample inferential properties in subtle and important problems
- Knockoffs is a powerful, flexible, and robust solution whenever there is considerable outside information on dist. of X such as GWAS
- Knockoffs addresses the replicability issue
- Where is the burden of knowledge?

- Robustness theory (Barber, Samworth and C.)
- Derandomization (multiple knockoffs)
- Knockoff constructions and statistics for other applications



Data

Knockoffs

Valid inference/selection

Thank You!

Derandomization

Combine information from mutiple knockoffs: who's consistently showing up?

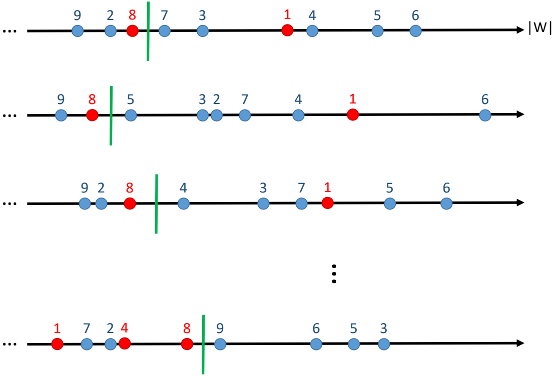


Figure: Cartoon representation of W's from different sample realizations of knockoffs