

# The knockoff filter

Statistical Learning

CLAMSES - University of Milano-Bicocca

Aldo Solari

# References

- Barber, Candès (2015) Controlling the False Discovery Rate via Knockoffs. *Ann. Statist.* 43:2504–2537
- Candès, Fan, Janson, Lv (2018). Panning for gold: model-X knockoffs for high dimensional controlled variable selection. *JRSS-B* 80:551–577.

There are two main approaches:

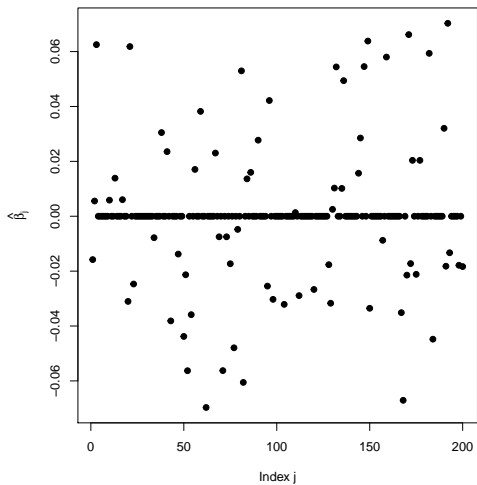
- *Fixed-X knockoffs*

Requires that  $X$  is full rank with  $n \geq 2p$

- *Model-X knockoffs*

Requires assumptions on  $X$  but works with  $p > n$

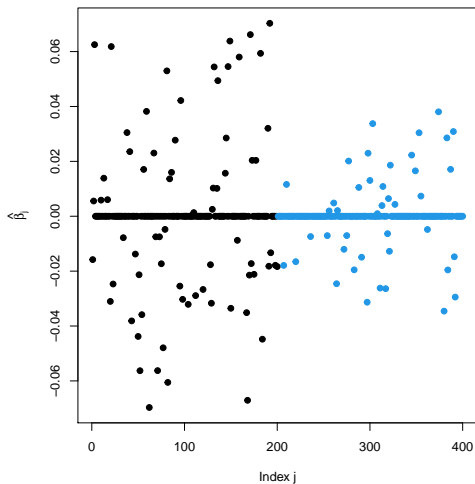
Fixed-X knockoffs



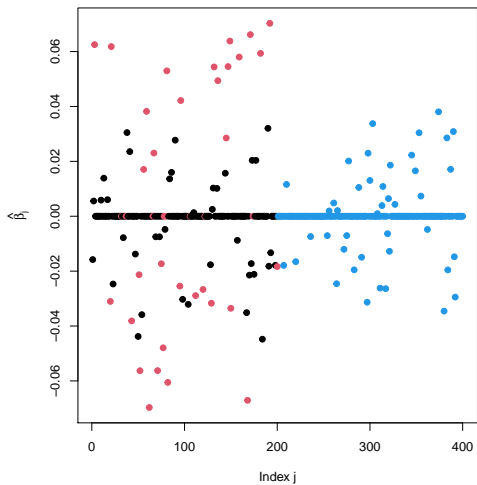
Lasso selects 67 features:  $\text{FDP}(\hat{S}) = 67/200$

# Main idea

- For each feature  $X_j$ , construct a *knockoff* copy  $\tilde{X}_j$
- Knockoffs  $\tilde{X}_1, \dots, \tilde{X}_p$  are independent of  $y$  and mimic the original variables  $X_1, \dots, X_p$  if they were null



Lasso selects 70 original and 43 knockoff:  $\widehat{\text{FDP}}(\hat{S}) = 43/70 \approx 61\%$



$$\text{True FDP}(\hat{S}) = 34/70 \approx 54\%$$



# Knockoff construction

- Suppose without loss of generality that the features are centered and scaled such that  $\|X_j\|_2^2 = 1$  for all  $j$
- Let  $\Sigma = X^t X$  be the correlation matrix of the features
- The method begins by augmenting the design matrix  $X$  with a second matrix  $\tilde{X} \in \mathbb{R}^{n \times p}$  of knockoff variables, constructed to satisfy

$$\begin{aligned} G = [X \ \tilde{X}]^t [X \ \tilde{X}] &= \begin{bmatrix} X^t X & X^t \tilde{X} \\ \tilde{X}^t X & \tilde{X}^t \tilde{X} \end{bmatrix} \\ &= \begin{bmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{bmatrix} \end{aligned}$$

for some diagonal matrix  $D = \text{diag}(d_1, \dots, d_p)$  such that  $G$  is positive definite

- The knockoffs have the same correlation structure as the original features

$$\tilde{X}^t \tilde{X} = X^t X = \Sigma$$

- The correlation between  $\tilde{X}_k$  and  $X_j$  is

$$\tilde{X}_j^t X_k = X_j^t X_k \quad \forall k \neq j$$

- The correlation between  $\tilde{X}_j$  and  $X_j$  is

$$\tilde{X}_j^t X_j = 1 - d_j$$

with  $d_j$  as close to 1 as possible

# Equi-correlated knockoffs

Suppose we require  $d_j = d$  for all  $j$ . Define

$$\tilde{X} = X(I_p - d\Sigma^{-1}) + UC$$

where

- $U \in \mathbb{R}^{n \times p}$  is an orthonormal matrix such that  $U^t X = 0$
- $C \in \mathbb{R}^{p \times p}$  from the Cholesky decomposition of

$$C^t C = 4((d/2)I_p - (d/2)^2 \Sigma^{-1})$$

This approach corresponds to `method="equi"` in the `knockoff` package. A semidefinite programming approach is used to determine the values that minimize  $\sum_{j=1}^p (1 - d_j)$  subject to the constraints (`method="sdp"`)

# The knockoff statistics

- Fit the lasso to the augmented design matrix  $[X \tilde{X}]$  for  $\lambda \in \Lambda$
- Let  $[\hat{\beta}(\lambda) \tilde{\beta}(\lambda)]$ ,  $\lambda \in \Lambda$  denote the coefficient estimates
- Compute

$Z_j = \sup\{\lambda \in \Lambda : \hat{\beta}_j(\lambda) \neq 0\} = \text{first time } X_j \text{ enters the lasso path}$

$\tilde{Z}_j = \sup\{\lambda \in \Lambda : \tilde{\beta}_j(\lambda) \neq 0\} = \text{first time } \tilde{X}_j \text{ enters the lasso path}$

- Then define the statistics

$$W_j = \max(Z_j, \tilde{Z}_j) \cdot \text{sign}(Z_j - \tilde{Z}_j) = \begin{cases} Z_j & \text{if } X_j \text{ enters first } (Z_j > \tilde{Z}_j) \\ 0 & \text{if } Z_j = \tilde{Z}_j \\ -\tilde{Z}_j & \text{if } \tilde{X}_j \text{ enters first } (Z_j < \tilde{Z}_j) \end{cases}$$

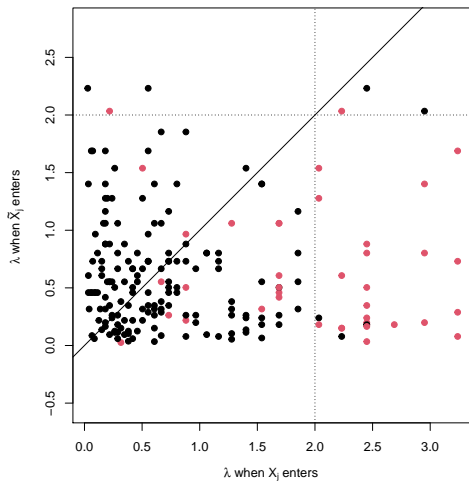
## FDP estimate

- For some threshold  $\tau \geq 0$ , select

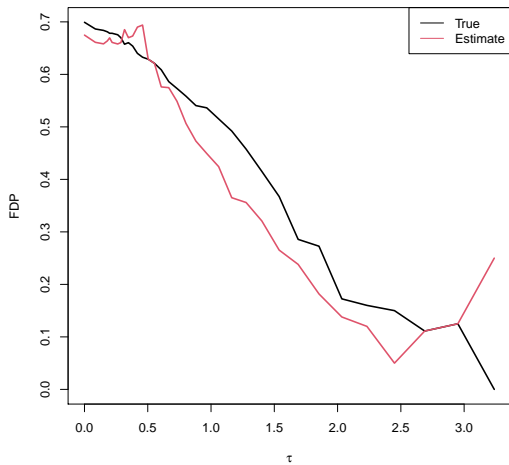
$$\hat{S}_\tau = \{j \in \{1, \dots, p\} : W_j \geq \tau\}$$

- The knockoff estimate of the FDP is

$$\begin{aligned} \text{FDP}(\hat{S}_\tau) &= \frac{\#\{j \in N : W_j \geq t\}}{\#\{j : W_j \geq t\}} \\ &\approx \frac{\#\{j \in N : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \\ &\leq \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} = \widehat{\text{FDP}}(\hat{S}_\tau) \end{aligned}$$



For  $\tau = 2$ ,  $|\hat{S}_\tau| = 29$  with  $\widehat{\text{FDP}}(\hat{S}_\tau) = 4/29$  and  $\text{FDP}(\hat{S}_\tau) = 5/29$



The knockoff procedure chooses a data-dependent threshold

$$\hat{\tau} = \min \left\{ \tau > 0 : \widehat{\text{FDP}}(\hat{S}_{\tau}) \leq \alpha \right\}$$

with  $\hat{\tau} = +\infty$  if no such  $\tau$  exists.

### Theorem

*For any  $\alpha \in (0, 1)$ , the knockoff procedure selects*

$$\hat{S}_{\hat{\tau}} = \{j \in \{1, \dots, p\} : W_j \geq \hat{\tau}\}$$

*with the guarantee that*

$$\text{FDR}(\hat{S}_{\hat{\tau}}) = \mathbb{E} \left( \frac{|N \cap \hat{S}_{\hat{\tau}}|}{|\hat{S}_{\hat{\tau}}|} \right) \leq \alpha$$

*where the expectation is taken over  $\varepsilon$  in the Gaussian linear model  $y = X\beta + \varepsilon$  while treating  $X$  and  $\tilde{X}$  as fixed.*



# Variable importance statistics

- Fit the Random Forest to the augmented design matrix  $[X \tilde{X}]$
- Compute

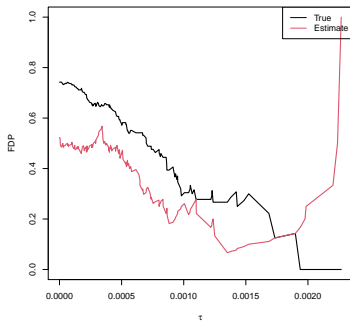
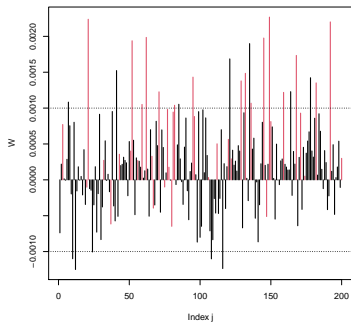
$$Z_j = \text{VariableImportance}(X_j)$$

$$\tilde{Z}_j = \text{VariableImportance}(\tilde{X}_j)$$

The importance of a variable is measured as the total decrease in node impurities from splitting on that variable, averaged over all trees

- Then define the statistics

$$W_j = \text{abs}(Z_j) - \text{abs}(\tilde{Z}_j)$$



For  $\tau = 0.001$ ,  $|\hat{S}_\tau| = 23$  with  $\widehat{\text{FDP}}(\hat{S}_\tau) = 4/23$  and  $\text{FDP}(\hat{S}_\tau) = 7/23$

Model- $X$  knockoff

# Modeling $X$

- $X$  is treated as a random matrix with i.i.d. rows  $x_i$
- $(x_i, y_i)$ ,  $i = 1, \dots, n$  are i.i.d. from some unknown distribution
- Assume we know the *marginal distribution* of  $x_i$ , e.g.

$$x_i = (x_{i1}, \dots, x_{ip}) \sim N_p(\mu, \Sigma)$$

- Null features given by *conditional independence*

$$N = \{j \in \{1, \dots, p\} : y \perp\!\!\!\perp x_j | x_{-j}\}$$

where  $x_{-j} = \{x_1, \dots, x_p\} \setminus \{x_j\}$

# Knockoffs in the Gaussian case

- The joint distribution of original features and knockoff copies satisfies

$$[x \tilde{x}] \sim N(M, V) \quad \text{with } M = \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \quad V = \begin{bmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{bmatrix}$$

where  $D = \text{diag}(d_1, \dots, d_p)$  such that  $V$  is positive definite

- Draw a random  $\tilde{x}_i$  from the conditional distribution  $\tilde{x}_i|x_i$ , which is normal with

$$\begin{aligned} \mathbb{E}(\tilde{x}_i|x_i) &= \mu + (\Sigma - D)\Sigma^{-1}(x_i - \mu) \\ \text{Var}(\tilde{x}_i|x_i) &= \Sigma - (\Sigma - D)\Sigma^{-1}(\Sigma - D) \end{aligned}$$

- If  $\mu$  and  $\Sigma$  are unknown, replace by estimates  $\hat{\mu}$  and  $\hat{\Sigma}$