

Topic 17: False Discovery Rate (FDR) and Knockoffs

by Sai Zhang

Key points: Constructing knockoff variables to control FDR when estimating regression coefficients.

Disclaimer: The note is built on Prof. [Jinchi Lv](#)'s lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.

17.1 Motivation

Consider the classical linear regression setting

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown vector of coefficients and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$. In a high-dimensional problem, we would like to just select a subset of all variables $\hat{S} \subset \{1, \dots, p\}$ s.t. conditional on $\{\mathbf{X}_j\}_{j \in \hat{S}}$, \mathbf{y} is **independent** of all other variables, we can define the **False Discovery Rate (FDR)** in can be defined as

Definition 17.1.1: False Discovery Rate (FDR)

$$\text{FDR} = \mathbb{E}(\text{FDP}) = \mathbb{E} \left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}|} = \frac{\#\{j : j \in \hat{S} \setminus \mathcal{S}\}}{\#\{j : j \in \hat{S}\}} \right]$$

where $\mathcal{H}_0 \subset \{1, \dots, p\}$ is the set of **null** variables: \mathbf{X}_j is **null** iff \mathbf{Y} is independent of \mathbf{X}_j conditional on the other variables $\mathbf{X}_{-j} = \{\mathbf{X}_1, \dots, \mathbf{X}_p\} \setminus \{\mathbf{X}_j\}$.

In this note, we consider a series of knockoff-based methods to control FDR. They all follow a common procedure:

- **Step 1:** Construct Knockoffs
- **Step 2:** Calculate test statistics for both original and knockoff variables
- **Step 3:** Calculate a threshold for the test statistics, controlling for a desired FDR level
- **Step 4:** Select variables that pass the threshold

17.2 Barber and Candès (2015)

Constructing the knockoffs [Barber and Candès \(2015\)](#) construct the knockoffs by the following procedure

- Calculate the Gram matrix $\boldsymbol{\Sigma} = \mathbf{X}'\mathbf{X}$ for the normalized original variables, where $\Sigma_{jj} = \|\mathbf{X}_j\|_2^2 = 1$

- Construct the knockoffs $\tilde{\mathbf{X}}$ s.t.

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \Sigma \quad \mathbf{X}'\tilde{\mathbf{X}} = \Sigma - \text{diag}\{\mathbf{s}\}$$

where $\mathbf{s} \in \mathbb{R}_+^p$ is a p -dimensional non-negative vector (larger s_j indicates higher power) and

- $\tilde{\mathbf{X}}$ exhibits the **same** covariance structure as the original design \mathbf{X}
- The correlation between distinct original variables and knockoffs are the same as between the originals:

$$\mathbf{X}_j'\tilde{\mathbf{X}}_k = \mathbf{X}_j'\mathbf{X}_k, \quad \forall j \neq k$$

- The correlation between the original variables and their own knockoffs is **less than 1**

$$\mathbf{X}_j'\tilde{\mathbf{X}}_j = \Sigma_{jj} - s_j = 1 - s_j$$

To construct such knockoffs,

- Given a proper \mathbf{s} , if $n \geq 2p$, then

$$\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \Sigma^{-1}\text{diag}\{\mathbf{s}\}) + \tilde{\mathbf{U}}\mathbf{C}$$

where $\tilde{\mathbf{U}} \in \mathbb{R}^{n \times p}$ is an **orthonormal** matrix s.t. $\tilde{\mathbf{U}}'\mathbf{X} = \mathbf{0}$ and $\mathbf{C}'\mathbf{C} = 2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\Sigma^{-1}\text{diag}\{\mathbf{s}\} \geq \mathbf{0}$

- A sufficient and necessary condition for $\tilde{\mathbf{X}}$ to exist: $\text{diag}\{\mathbf{s}\} \leq 2\Sigma$

2 types of knockoffs can be constructed, following these procedures

T1 **Equi-correlated** knockoffs: set $s_j = 2\lambda_{\min}(\Sigma) \wedge 1$ for all j , then $\langle \mathbf{X}_j, \tilde{\mathbf{X}}_j \rangle = 1 - 2\lambda_{\min}(\Sigma) \wedge 1$ for all j . This is essentially minimizing $|\langle \mathbf{X}_j, \tilde{\mathbf{X}}_j \rangle|$

T2 **SDP** knockoffs: solve the convex problem

$$\arg \min_{\mathbf{s}} \sum_j (1 - s_j) \quad \text{s.t. } 0 \leq s_j \leq 1, \text{diag}\{\mathbf{s}\} \leq 2\Sigma$$

which is essentially minimizing the average of $\langle \mathbf{X}_j, \tilde{\mathbf{X}}_j \rangle$

Calculate test statistics Define and calculate test statistics W_j for each $\beta_j \in \{1, \dots, p\}$ using $[\mathbf{X} \quad \tilde{\mathbf{X}}]$:

- the test statistic W_j should be constructed s.t. large positive values are evidence against the null hypothesis $\beta_j = 0$, for example, consider a Lasso on $[\mathbf{X} \quad \tilde{\mathbf{X}}]$

$$\hat{\beta}(\lambda) = \arg \min_{\mathbf{b}} \left\{ \frac{1}{2} \|\mathbf{y} - [\mathbf{X} \quad \tilde{\mathbf{X}}] \mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}$$

where λ is the point on the Lasso path at which the feature enters the model as

$$Z_j = \sup \{ \lambda : \hat{\beta}_j(\lambda) \neq 0 \}$$

$$\text{and set } W_j = (Z_j \vee \tilde{Z}_j) \cdot \begin{cases} +1, & Z_j > \tilde{Z}_j \\ -1, & Z_j < \tilde{Z}_j \end{cases}$$

- In general, the statistics W should satisfy the **sufficient** property and **anti-symmetry** property:

¹Other choices of W_j are $W_j = |\mathbf{X}_j'\mathbf{y}| - |\tilde{\mathbf{X}}_j'\mathbf{y}|$, or $|\hat{\beta}_j^{\text{LS}}| - |\hat{\beta}_{j+p}^{\text{LS}}|$

Definition 17.2.1: Property of Test Statistics W_j

The test statistic W_j is said to obey

- the **sufficient** property if \mathbf{W} depends only on the Gram matrix and on feature-response inner products, that is

$$\mathbf{W} = f\left([\mathbf{X} \ \tilde{\mathbf{X}}]' [\mathbf{X} \ \tilde{\mathbf{X}}], [\mathbf{X} \ \tilde{\mathbf{X}}]' \mathbf{y}\right)$$

- the **antisymmetry** property if swapping the original \mathbf{X}_j and its knockoff $\tilde{\mathbf{X}}_j$ has the effect of **switching the sign** of W_j , that is

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

Calculate a threshold for the test statistics After defining the test statistic, we then

- Let q be the target FDR, define the data-dependent threshold T as

$$T = \min \left\{ t \in \mathcal{W} : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\}$$

where $\mathcal{W} = \{|W_j| : j = 1, \dots, p\} \setminus \{0\}$ is the set of unique non-zero values attained by $|W_j|$'s.

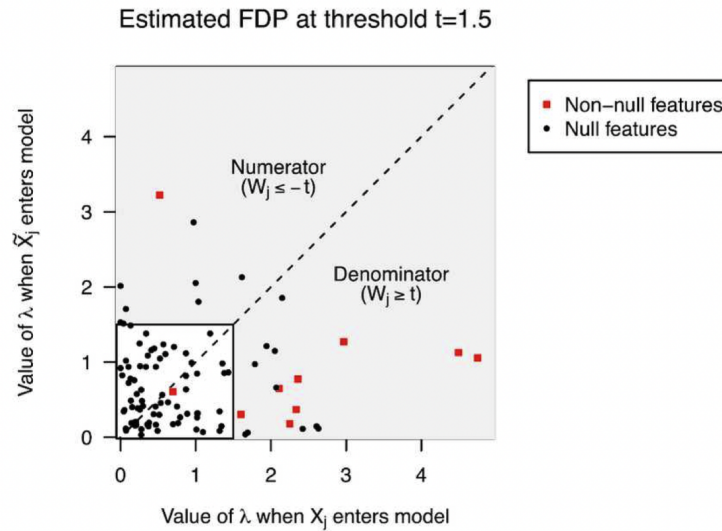


Figure 17.1: Visualizing Test Statistic Thresholding

Variable selection after building the threshold,

- for each $j = 1, \dots, p$, reject $H_{0,j} : \beta_j = 0$ if $W_j \geq T$, the knockoff filter selects the model

$$\hat{S} = \{j : W_j \geq T\}$$

17.2.1 Intuition and Theory

Why knockoffs work?

- \mathbf{W} is constructed (**antisymmetry** and **sufficiency**) such that the signs of the W_j 's are i.i.d. random for the null
- for any threshold t , we have

$$\#\{j : \beta_j = 0, W_j \geq t\} \stackrel{d}{=} \#\{j : \beta_j = 0, W_j \leq -t\}$$

, and the false discovery proportion (FDP) can be estimated as

$$\begin{aligned} \frac{\#\{j : \beta_j = 0, W_j \geq t\}}{\max(\#\{j : W_j \geq t\}, 1)} &\simeq \frac{\#\{j : \beta_j = 0, W_j \leq -t\}}{\max(\#\{j : W_j \geq t\}, 1)} \\ &\leq \frac{\#\{j : W_j \leq -t\}}{\max(\#\{j : W_j \geq t\}, 1)} := \widehat{\text{FDP}}(t) \end{aligned}$$

then the knockoff procedure can be interpreted as finding a threshold via $T = \min \left\{ t \in \mathcal{W} : \widehat{\text{FDR}}(t) \leq q \right\}$

The knockoff procedure essentially controls a quantity **nearly equal** to the FDR. To control the FDR **exactly**, we have, **textbfknockoff+**, a more conservative modification of the knockoff procedure, where the threshold is

$$T = \min \left\{ t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leq -t\}}{\max(\#\{j : W_j \geq t\}, 1)} \leq q \right\}$$

the **+1** part makes it harder to reject the null:

$$\begin{aligned} \text{FDP} &= \frac{\#\{j : \beta_j = 0, W_j \geq -T\}}{\#\{j : W_j \geq T\} \vee 1} \cdot \frac{1 + \#\{j : \beta_j = 0, W_j \leq -T\}}{1 + \#\{j : \beta_j = 0, W_j \leq -T\}} \\ &\leq \frac{1 + \#\{j : W_j \leq -T\}}{\#\{j : W_j \geq T\} \vee 1} \cdot \frac{\#\{j : \beta_j = 0, W_j \geq T\}}{1 + \#\{j : \beta_j = 0, W_j \leq -T\}} \\ &\leq q \cdot 1 \end{aligned}$$

Then, we have the following theorem

Theorem 17.2.2: Property of the Knockoff Method

For any $q \in [0, 1]$, the **knockoff** method satisfies

$$\mathbb{E} \left[\frac{\#\{j : \beta_j = 0, j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} + q^{-1}} \right] \leq q$$

and the **knockoff+** method satisfies

$$\mathbb{E} \left[\frac{\#\{j : \beta_j = 0, j \in \hat{S}\}}{\#\{j : j \in \hat{S}\}} \right] \leq q$$

in both cases, the expectation is taken over the Gaussian noise in the model, while treating original variables \mathbf{X} and knockoffs $\tilde{\mathbf{X}}$ as fixed

17.3 Candès et al. (2018)

Another way of constructing knockoffs, introduced by Candès et al. (2018), is by a swapping method:

Constructing the knockoffs for the family of random variables $\mathbf{X} = (X_1, \dots, X_p)$ are a new family of random variables $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_p)$ constructed with the following 2 properties

- for any subset $S \subset \{1, \dots, p\}$,

$$(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)} \stackrel{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})$$

- $\tilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}$ if there is a response \mathbf{Y}

Suppose $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, then $(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)}$ satisfies $(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)} \stackrel{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})$ if

$$(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)} \stackrel{d}{=} (\mathbf{X}, \tilde{\mathbf{X}}) \sim \mathcal{N}(0, \mathbf{G}), \quad \text{where } \mathbf{G} = \begin{pmatrix} \Sigma & \Sigma - \text{diag}(s) \\ \Sigma - \text{diag}(s) & \Sigma \end{pmatrix}$$

where $\text{diag}(s)$ is any **diagonal matrix** s.t. \mathbf{G} is **positive semidefinite**. The knockoffs constructed this way are named **MX knockoffs**. For \mathbf{P} , the permutation matrix encoding the swap,

$$\mathbf{P}\mathbf{G}\mathbf{P} = \mathbf{G}$$

then we can sample the knockoff vector $\tilde{\mathbf{X}}$ from the conditional distribution

$$\tilde{\mathbf{X}} \mid \mathbf{X} \stackrel{d}{=} \mathcal{N}(\mu, \mathbf{V})$$

where

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

An important lemma is

Lemma 17.3.1: MX Knockoff Construction

For **MX knockoffs**, swapping **null** covariates with their knockoffs would **not** change the joint distribution of the original covariate \mathbf{X} and their knockoffs $\tilde{\mathbf{X}}$, conditional on the response \mathbf{Y} : Take any subset $S \subset \mathcal{H}_0$ of nulls, then

$$(\mathbf{X}, \tilde{\mathbf{X}}) \mid \mathbf{y} \stackrel{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)} \mid \mathbf{y}$$

Here, the main assumption of model-X knockoffs is assuming **known** joint distribution of covariates, and this leads to

Proposition 17.3.2: Conditional Exchangeability of MX Knockoffs

The random variables $(\tilde{X}_1, \dots, \tilde{X}_p)$ are **MX knockoffs** for (X_1, \dots, X_p) if and only if for any $j \in \{1, \dots, p\}$, the pair (X_j, \tilde{X}_j) is exchangeable conditional on all the other variables and their knockoffs.

under Prop.17.3.2, we can use the following algorithm to construct the MX Knockoffs

Algorithm 17.3.3: Sequential Conditional Independent Pairs

```

j = 1
while j ≤ p do
  sample  $\tilde{X}_j$  from  $\mathcal{L}(X_j \mid X_{-j}, \tilde{X}_{1:j-1})$ 
  j = j + 1
enda

```

^aExample with $p = 3$

- $j = 1$: sample \tilde{X}_1 from $\mathcal{L}(X_1 \mid X_{2:3})$
- $j = 2$: sample \tilde{X}_2 from $\mathcal{L}(X_2 \mid X_1, X_3, \tilde{X}_1)$
- $j = 3$: sample \tilde{X}_3 from $\mathcal{L}(X_3 \mid X_{1:2}, \tilde{X}_{1:2})$

And an **approximate** construction can be achieved via matching the first 2 moments of $(X, \tilde{X})_{\text{swap}(S)}$ and (X, \tilde{X}) ,

$$\text{cov}(X, \tilde{X}) = G \quad G = \begin{pmatrix} \Sigma & \Sigma - \text{diag}(s) \\ \Sigma - \text{diag}(s) & \Sigma \end{pmatrix}$$

which can be achieved through 2 ways:

- **equicorrelated** construction

$$s_j^{\text{EQ}} = 2\lambda_{\min}(\Sigma) \wedge 1, \forall j$$

minimizing the **correlation between variable knockoff pairs** subject to the constraint that all such pairs *must have the same correlation*.

ISSUE with large p : $\lambda_{\min}(\Sigma)$ tends to be extremely small: computationally easy, but **low power** of s_j^{EQ}

- **semidefinite programme** construction

$$\begin{aligned} &\text{minimize} && \sum_j |1 - s_j^{\text{SDP}}| \\ &\text{subject to} && s_j^{\text{SDP}} \geq 0, \text{diag}(s^{\text{SDP}}) \leq 2\Sigma \end{aligned}$$

minimizing the **sum of the absolute values** of variable knockoff correlations between all suitable s

ISSUE with large p : SDP (a convex problem) is computationally expensive

Hence, in high-dimensional situation, follow a 2-step procedure to combine both

- Step 1: choose an **approximation** Σ_{approx} of Σ and solve

$$\begin{aligned} &\text{minimize} && \sum_j |1 - \hat{s}_j| \\ &\text{subject to} && \hat{s}_j \geq 0, \text{diag}(\hat{s}) \leq 2\Sigma_{\text{approx}} \end{aligned}$$

- Step 2: solve

$$\begin{aligned} &\text{maximize} && \gamma \\ &\text{subject to} && \text{diag}(\gamma \hat{s}) \leq 2\Sigma \end{aligned}$$

and set $s^{\text{ASDP}} = \gamma \hat{s}$

It's easy to see that this 2-step procedure can be reduced to equicorrelated or semidefinite programme

- **equicorrelated**: $\Sigma = \mathbf{I} \Rightarrow \hat{s}_j = 1, \gamma = 2 \times \lambda_{\min} \Sigma \wedge 1$
- **semidefinite programme**: $\Sigma = \Sigma, \hat{s}_j = s^{\text{SDP}}, \gamma = 1$

Calculate test statistics After constructing the knockoffs, we can construct the feature importance statistics by imposing a **flip sign** property: swapping the j th variable with its knockoff has the effect of changing the sign of W_j

$$w_j \{(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)}, \mathbf{y}\} = \begin{cases} w_j \{(\mathbf{X}, \tilde{\mathbf{X}}), \mathbf{y}\}, & j \notin S \\ -w_j \{(\mathbf{X}, \tilde{\mathbf{X}}), \mathbf{y}\}, & j \in S \end{cases}$$

consider a statistic \mathbf{T} for each original and knockoff variable

$$\mathbf{T} \triangleq (\mathbf{Z}, \tilde{\mathbf{Z}}) = (Z_1, \dots, Z_p, \tilde{Z}_1, \dots, \tilde{Z}_p) = t \{(\mathbf{X}, \tilde{\mathbf{X}}), \mathbf{y}\}$$

if the components of \mathbf{T} are switched in the same way:

$$(\mathbf{Z}, \tilde{\mathbf{Z}})_{\text{swap}(S)} = t \{(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)}, \mathbf{y}\}$$

then the **flip sign** property can be achieved by setting

$$W_j = f_j(Z_j, \tilde{Z}_j)$$

where f_j is any **antisymmetric** function $f(v, u) = -f(u, v)$.

Lemma 17.3.4: Feature Statistics: Lasso Coefficient Difference (LCD)

Consider the Lasso **augmented with knockoffs**

$$\min_{b \in \mathbb{R}^{2p}} \frac{1}{2} \|y - (\mathbf{X}, \tilde{\mathbf{X}})b\|_2^2 + \lambda \|b\|_1$$

which has solution $\hat{b}(\lambda) = (\hat{b}_1(\lambda), \dots, \hat{b}_p(\lambda), \hat{b}_{p+1}(\lambda), \dots, \hat{b}_{2p}(\lambda))$, then the statistic can be constructed as

$$W_j = Z_j - \tilde{Z}_j = |\hat{b}_j(\lambda)| - |\hat{b}_{j+p}(\lambda)|$$

and conditional on $(|W_1|, \dots, |W_p|)$, the sign of the null W_j s ($j \in \mathcal{H}_0$) are i.i.d. coin flips^a.

^aProof: for a sequence independent random variables $\epsilon = (\epsilon_1, \dots, \epsilon_p)$ s.t. $\epsilon_j = \pm 1$ with probability $\frac{1}{2}$ if $j \in \mathcal{H}_0$, and $\epsilon_j = 1$ otherwise, put $S = \{j : \epsilon_j = -1\} \subset \mathcal{H}_0$

- flip sign property: $W_{\text{swap}(S)} \triangleq w \{(\mathbf{X}, \tilde{\mathbf{X}})_{\text{swap}(S)}, \mathbf{y}\} \stackrel{\text{d}}{=} \epsilon \odot W = (\epsilon_1 W_1, \dots, \epsilon_p W_p)$

- Lemma 17.3.1: $W_{\text{swap}(S)} \stackrel{\text{d}}{=} W$

which establishes $W \stackrel{\text{d}}{=} \epsilon \odot W$

- a large positive value of W_j provides some evidence that the distribution of \mathbf{Y} depends on \mathbf{X}_j
- value of λ can be chosen in any data-dependent fashion for a pair of \mathbf{y} and $(\mathbf{X}, \tilde{\mathbf{X}})$

Why **i.i.d. coin flips**? the null W_j s are **symmetric**

$$\# \{j : W_j \leq -t, j \in \mathcal{H}_0\} \stackrel{\text{d}}{=} \# \{j : W_j \geq t, j \in \mathcal{H}_0\}$$

and for any fixed threshold $t > 0$

$$\#\{j : W_j \leq -t\} \geq \#\{j : W_j \leq -t, j \in \mathcal{H}_0\}$$

so for the false discovery proportion (FDP)

$$\text{FDP}(t) = \frac{\#\{j : W_j \geq t, j \in \mathcal{H}_0\}}{\#\{j : W_j \geq t\}}$$

an upward-biased estimate is

$$\widehat{\text{FDP}}(t) = \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}}$$

then Theorem 17.2.2 applies.

17.4 Fan et al. (2020)

The model-X knockoff (Candès et al., 2018) can accommodate an arbitrarily large p , but assumes **known** joint distribution of covariates. Fan et al. (2020) introduced a data-splitting procedure to separately estimate the unknown covariate distribution and then globally construct the knockoff variables

Fan et al. (2020)

References

- Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.
- Emmanuel J Candès, Jianqing Fan, Lucas Janson, and Jinchi Lv. Panning for gold: ‘model-x’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.
- Yingying Fan, Emre Demirkaya, Gaorong Li, and Jinchi Lv. Rank: Large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 115(529):362–379, 2020.