Econometrics **June 18, 2023**

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 $\hat{S} \subset \{1, \cdots, p\}$ 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)

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

where $\mathcal{H}_0 \subset \{1, \cdots, p\}$ 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 $\mathbf{X}_{-j} = \{\mathbf{X}_1, \cdots, \mathbf{X}_p\} \setminus \{\mathbf{X}_j\}$.

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, controling for a desired FDR level
- **Step 4**: Select variables that pass the threshold

17.2 Barber and Candes (2015)

Constructing the knockoffs Barber and Candes (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.

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \boldsymbol{\Sigma} \qquad\qquad \mathbf{X}'\tilde{\mathbf{X}} = \boldsymbol{\Sigma} - \operatorname{diag}\{\mathbf{s}\}$$

  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 structrue as the original design $\mathbf{X}$
  - The correlation between distinct original variables and knockoffs are the same as between the originals:

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

  - 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

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

    where $\tilde{\mathbf{U}} \in \mathbb{R}^{n\times p}$ is an **orthonormal** matrix s.t. $\tilde{\mathbf{U}}'\mathbf{X} = \mathbf{0}$ and $\mathbf{C}'\mathbf{C} = 2\operatorname{diag}\{\mathbf{s}\} - \operatorname{diag}\{\mathbf{s}\}\boldsymbol{\Sigma}^{-1}\operatorname{diag}\{\mathbf{s}\} \geq \mathbf{0}$
  - A sufficient and necessary condition for $\tilde{\mathbf{X}}$ to exist: $\operatorname{diag}\{\mathbf{s}\} \leq 2\boldsymbol{\Sigma}$

  2 types of knockoffs can be constructed, following these procedures

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

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

$$\arg\min_{\mathbf{x}} \sum_j (1 - s_j) \qquad\qquad s.t. 0 \leq s_j \leq 1, \operatorname{diag}\{\mathbf{s}\} \leq 2\boldsymbol{\Sigma}$$

  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 $\beta_j \in \{1, \cdots, p\}$ using $\begin{bmatrix}\mathbf{X} & \tilde{\mathbf{X}}\end{bmatrix}$:

- 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 $\begin{bmatrix}\mathbf{X} & \tilde{\mathbf{X}}\end{bmatrix}$

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

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

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

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

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

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

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 <u>*only*</u> on the Gram matrix and on feature-response inner products, that is

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

- 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

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

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

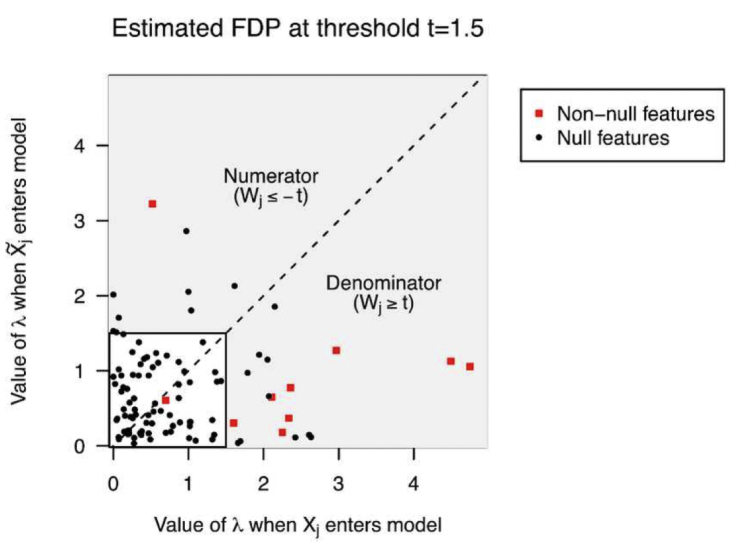

Figure 17.1: Visualizing Test Statistic Thresholding

Variable selection after building the threshold,

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

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

17.2.1 Intuition and Theory

Why knockoffs work?

- **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

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

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

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

References

Rina Foygel Barber and Emmanuel J. Candes. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.