

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)

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

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)

[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.

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

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 $\mathbf{C}'\mathbf{C} = 2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\Sigma^{-1}\text{diag}\{\mathbf{s}\} \geq \mathbf{0}$

References

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