

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 , \mathbf{X}_j is null iff \mathbf{Y} is independent of \mathbf{X}_j conditional on the other variables $\mathbf{X}_{-j} = \{\mathbf{X}_1, \dots, \mathbf{X}_p\} \setminus \{\mathbf{X}_j\}$. The subset of of all null variables is denoted by $\mathcal{H}_0 \subset \{1, \dots, p\}$. Variable \mathbf{X}_j is non-null or relevant, if $j \in \mathcal{H}_0$

Candès et al. (2018)

References

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.