1 Variable Selection

Let y be response variable and x be explanatory variables or covariates. Given i.i.d. samples $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ from the joint distribution $p_{\mathsf{x},\mathsf{y}}$, we are interested in asking the question

which of the many covariates $x_1, \cdots, x_p$ does the response y depend on?

assuming that the response does depend on a sparse set of variables. In reality, we are interested in the causal relationship. However, quantifying causal effects requires interventions and not possible from purely observational data. A natural relaxation is to find covariates dependent (in a statistical sense) on the response, conditioned on all other observed features [1]. Formally, we want to find smallest $\mathcal{S} \subset [p]$ s.t.

$$\mathsf{y} \perp\!\!\!\perp \mathsf{x}_{\mathcal{S}} \mid \mathsf{x}_{\setminus \mathcal{S}}$$

A natural interpretation is that the other variables $\mathsf{x}_{\setminus \mathcal{S}}$ do not provide additional information about y. If we think of $\mathcal{G}$ as graph representing the joint distribution $p_{\mathsf{x},\mathsf{y}}$, then $\mathcal{S}$ is the markov blanket for node y. We can pose the problem of finding the Markov blanket of y as

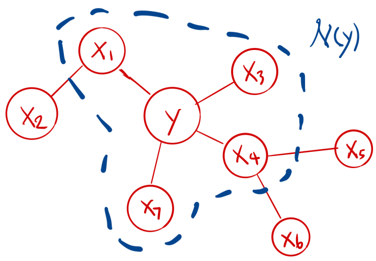

Figure 1: $\mathcal{S} = \{\mathsf{x}_1, \mathsf{x}_3, \mathsf{x}_4, \mathsf{x}_7\}$

a multiple binary hypothesis test

$$H_0^{(j)} : \mathsf{y} \perp\!\!\!\perp \mathsf{x}_j \mid \mathsf{x}_{\setminus \{j\}} \qquad \text{for} \quad j = 1, \cdots, p$$

Let $\mathcal{H}_0 = \left\{ \mathsf{x}_j \mid H_0^{(j)} \text{ holds} \right\}$ be the set of truly irrelevant covariates. In essence we are interested in maximizing *power* while controlling the number of false positives. A global threshold for p-values of each tests is overly conservative for large p, an alternative approach is to control *false discovery rate* (FDR) [2].

$$\begin{aligned} \text{maximize}_{\hat{\mathcal{S}} \subset [p]} \quad & \mathbb{E}\left[\frac{\hat{\mathcal{S}} \setminus \mathcal{H}_0}{|\hat{\mathcal{S}}|} \right] && \text{(maximize power)} \\ \text{subject to} \quad & \mathbb{E}\left[\frac{\hat{\mathcal{S}} \cap \mathcal{H}_0}{|\hat{\mathcal{S}}|} \right] \leq q && \text{(control FDR)} \end{aligned}$$

2 Model-X Knockoff

References

[1] Jaime Roquero Gimenez, Amirata Ghorbani, and James Zou. "Knockoffs for the mass: new feature importance statistics with false discovery guarantees". In: *arXiv:1807.06214 [cs, stat]* (May 28, 2019). arXiv: 1807.06214. URL: http://arxiv.org/abs/1807.06214 (visited on 04/17/2020).

[2] Yoav Benjamini and Yosef Hochberg. "Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing". In: *J. Royal Statist. Soc., Series B* 57 (Nov. 30, 1995), pp. 289–300. DOI: 10.2307/2346101.