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.

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

A natural interpretation is that the other variables $\mathsf{x}_{\backslash \mathcal{S}}$ 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

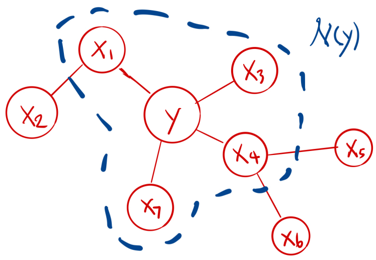

Figure 1: $\mathcal{S} = \{\mathsf{x}_1, \mathsf{x}_3, \mathsf{x}_4, \mathsf{x}_7\}$

a multiple binary hypothesis test

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

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

$$\begin{aligned} \text{maximize}_{\hat{\mathcal{S}} \subset [p]} \quad & \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \setminus \mathcal{H}_0|}{|\hat{\mathcal{S}}|} \right] \\ \text{subject to} \quad & \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{\max\left\{ |\hat{\mathcal{S}}|, 1 \right\}} \right] \leq q \end{aligned} \tag{2}$$

If $p_{\mathsf{y}|\mathsf{x}}(\cdot|x)$ assumes a parametric generalized linear model form,

$$\mathbb{E}\left[\mathsf{y}|\mathsf{x}\right] = g^{-1}(\eta) \qquad \eta = \beta_1 x_1 + \cdots + \beta_p x_p$$

Then by [3], testing for conditional independence (1) is equivalent to the following test,

$$H_0^{(j)} : \beta_j = 0 \qquad \text{for} \quad j = 1, \cdots, p$$

2 Model-X Knockoff

Traditionally, $p_{\mathsf{y|x}}$ is chosen to be in some parametric family, e.g. GLM, and variable selection with FDR control is performed by computing & plugging p-values into the BHq procedure [2]. Recently, [4, 3] designed a *knockoff* framework for performing variable selection on high-dimensional nonparametric models with finite sample guarantees over the constraints in (2). The framework requires significant knowledge of p_{x} and assumes nothing about the $p_{\mathsf{y|x}}$. This might give way to performing reproducible and robust variable selection where the $p_{\mathsf{y|x}}$ is parameterized by highly complex mappings, e.g. neural networks. In addition, modeling p_{x} might be a suitable task for problems where we have large amount of unsupervised data, or we know a priori some structure about p_{x}, which are often the case for large scale machine learning applications.

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.

[3] Emmanuel Candes et al. "Panning for Gold: Model-X Knockoffs for High-dimensional Controlled Variable Selection". In: *arXiv:1610.02351 [math, stat]* (Dec. 12, 2017). arXiv: 1610.02351. URL: http://arxiv.org/abs/1610.02351 (visited on 02/03/2020).

[4] Rina Foygel Barber and Emmanuel J. Candès. "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5 (Oct. 2015), pp. 2055–2085. ISSN: 0090-5364. DOI: 10.1214/15-AOS1337. arXiv: 1404.5609. URL: http://arxiv.org/abs/1404.5609 (visited on 04/14/2020).