

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 $p_{x,y}$, we are interested in asking the question

which of the many covariates x_1, \dots, 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.

$$y \perp\!\!\!\perp x_{\mathcal{S}} \mid x_{\setminus \mathcal{S}}$$

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

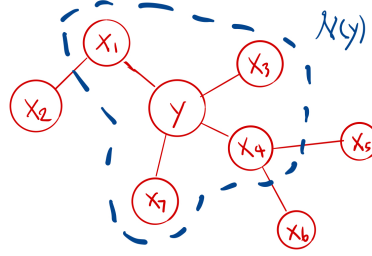


Figure 1: $\mathcal{S} = \{x_1, x_3, x_4, x_7\}$

a multiple (independent) binary hypothesis test

$$H_0^{(j)} : y \perp\!\!\!\perp x_j \mid x_{\setminus \{j\}} \quad \text{for } j = 1, \dots, p \quad (1)$$

Let $\mathcal{H}_0 = \{j \mid H_0^{(j)} \text{ holds}\}$ 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].

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

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

$$\mathbb{E}[y|x] = g^{-1}(\eta) \quad \eta = \beta_1 x_1 + \dots + \beta_p x_p$$

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

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

2 Model-X Knockoff

Traditionally, $p_{y|x}$ 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 $p_{y|x}$. This might give way to performing reproducible and robust variable selection where the $p_{y|x}$ 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.

Definition. \tilde{x} is a model-X knockoff for x if

$$\tilde{x} \perp\!\!\!\perp y \mid x \quad (3)$$

$$(x, \tilde{x})_{\text{swap}(\mathcal{S})} \stackrel{d}{=} (x, \tilde{x}) \quad \text{for any} \quad \mathcal{S} \subset [p] \quad (4)$$

where $(\cdot)_{\text{swap}(\mathcal{S})}$ swaps coordinates for all $j \in \mathcal{S}$ with coordinate $j + p$ and leaves other coordinate unchanged. (3) is guaranteed if \tilde{x} is constructed without knowledge of y .

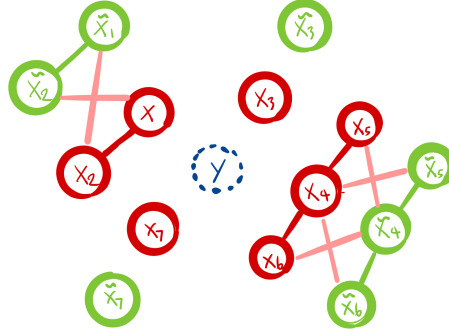


Figure 2: \mathcal{G} represents $p_{x, \tilde{x}}$. (3) implies \tilde{x} is not in the Markov blanket of node y for a graph representing joint distribution $p_{x, \tilde{x}}$. (4) implies that the x, \tilde{x} are pairwise exchangeable, i.e. taking any subset of (green) variables and swap them with their (red) knockoff creates an isomorphism of \mathcal{G} , i.e. edges preserved from the perspective of swapped variables. In practice, we want x_j, \tilde{x}_j be as independent as possible, i.e. no edge connecting x_j, \tilde{x}_j for all $j \in [p]$

Intuitively, *knockoffs* mimics the dependence structure as the original covariates x , while being invariant to $\text{swap}(\cdot)$ operation, and is independent of the response y . It serves as a control for evaluating how much of dependence on the response is due to dependence structure of other variables and how much of it is due to dependence with response y .

2.1 Knockoff Procedure for LASSO

Consider a linear Gaussian model $y = \beta^T x$ where $x \sim \mathcal{N}(\mu, \Sigma)$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be any design matrix with $n > p$. The knockoff filtering procedure for computing variable selection with controlled FDR is given by

1. **(Generate Knockoffs)** To ensure exchangeability property (4), it must be

$$\begin{pmatrix} \mathbf{x} \\ \tilde{\mathbf{x}} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} \\ \boldsymbol{\Sigma} - \text{diag}\{\mathbf{s}\} & \boldsymbol{\Sigma} \end{bmatrix} \right)$$

One way to construct knockoff is to sample \tilde{x} from the conditional distribution [3, 1],

$$\begin{aligned} \tilde{x} \mid (x = x) &\sim \mathcal{N} \left(\boldsymbol{\mu}_{\tilde{x}|x}(x), \boldsymbol{\Sigma}_{\tilde{x}|x}(x) \right) \\ \boldsymbol{\mu}_{\tilde{x}|x}(x) &= (I - \text{diag}\{\mathbf{s}\} \boldsymbol{\Sigma}^{-1}) x + \text{diag}\{\mathbf{s}\} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ \boldsymbol{\Sigma}_{\tilde{x}|x}(x) &= 2 \text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\} \boldsymbol{\Sigma}^{-1} \text{diag}\{\mathbf{s}\} \end{aligned}$$

Alternatively, we can match empirical first and second moment [4] and construct design for knockoff as

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

where $\tilde{\mathbf{U}} \in \mathbb{R}^{n \times p}$ is orthonormal matrix whose column is orthogonal to \mathbf{X} and $\mathbf{C}^T \mathbf{C} = 2 \text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\} \boldsymbol{\Sigma}^{-1} \text{diag}\{\mathbf{s}\}$ is a Cholesky decomposition.

2. **(Compute Pairwise Statistics)** Compute Lasso for the coefficients

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{2p}} \frac{1}{2} \left\| \mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}] \boldsymbol{\beta} \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

Now we compute statistics w for each pair of original and knockoff variables

$$w_j := w_j(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}) = |\hat{\boldsymbol{\beta}}_j| - |\hat{\boldsymbol{\beta}}_{j+p}| \quad \text{for } j = 1, \dots, p$$

which satisfy the coin-flip property. One important consequence is that for any $j \in \mathcal{H}_0$, w_j is a symmetric distribution about the origin [4, 3].

3. **(Compute Threshold for Statistics)** Given $q > 0$ be target FDR, then let

$$\tau_+ = \min \left\{ t > 0 \mid \frac{1 + \#\{j \mid w_j \leq -t\}}{\#\{j \mid w_j \geq t\}} \leq q \right\}$$

4. **(Perform Test)** with threshold τ_+ ,

$$\hat{\mathcal{S}} = \{j \mid w_j \geq \tau_+\}$$

ensures that $\text{FDR} \leq q$

2.2 Current Problems

1. Generating knockoffs is a hard problem, there has been work to sample knockoffs using MCMC [5], with generative model [6], in particular with GAN [7] and with latent variable models [1].
2. There has been a few papers that tries to analyze power of the knockoff framework assuming Gaussian data [8]
3. There has been some preliminary work on adopting multiple hypothesis testing and FDR control when $p_{y|x}$ is parameterized by neural networks, e.g. by MLP [9]. The challenge here seems that it is not easy to maintain power of tests.
4. It seems that currently, experimental datasets is either simulated or small in size of n, p . I have not found any attempts that tried this method on image datasets. Might be interesting to see how this fairs.

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).
- [5] Stephen Bates et al. “Metropolized Knockoff Sampling”. In: *arXiv:1903.00434 [stat]* (Mar. 1, 2019). arXiv: [1903.00434](#). URL: <http://arxiv.org/abs/1903.00434> (visited on 04/18/2020).
- [6] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. “Deep Knockoffs”. In: *arXiv:1811.06687 [math, stat]* (Nov. 16, 2018). arXiv: [1811.06687](#). URL: <http://arxiv.org/abs/1811.06687> (visited on 02/26/2020).
- [7] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. “KnockoffGAN: Generating Knockoffs for Feature Selection using Generative Adversarial Networks”. In: International Conference on Learning Representations. Sept. 27, 2018. URL: https://openreview.net/forum?id=ByeZ5jC5YQ&source=post_page----- (visited on 04/17/2020).
- [8] Jingbo Liu and Philippe Rigollet. “Power analysis of knockoff filters for correlated designs”. In: *arXiv:1910.12428 [cs, math, stat]* (Jan. 9, 2020). arXiv: [1910.12428](#). URL: <http://arxiv.org/abs/1910.12428> (visited on 04/17/2020).
- [9] Yang Lu et al. “DeepPINK: reproducible feature selection in deep neural networks”. In: *Advances in Neural Information Processing Systems* 31. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 8676–8686. URL: <http://papers.nips.cc/paper/8085-deeppink-reproducible-feature-selection-in-deep-neural-networks.pdf> (visited on 02/03/2020).