# 1 Variable Selection

Let $\mathsf{y}$ be response variable and $\mathsf{x}$ 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 general, we want to find a subset $\mathcal{S} \subset [p]$ such that $\mathcal{S}$ retains the relevant information in $\mathsf{x}$ for making inference about $\mathsf{y}$,

$$\begin{aligned} \text{maximize}_{S\subset[p]} \quad & \mathcal{Q}(S) \\ \text{subject to} \quad & \|\mathcal{S}\|_0 \leq t \end{aligned}$$

where $\mathcal{Q}(\cdot)$ quantifies the relevance of a feature subset to response. The choice of $\mathcal{Q}(\cdot)$ should be 1) capable of detecting the desired functional dependence between the covariates and the response and 2) concentrated with respect to the underlying measure (generalize well to test data) [1]. Example of criteria $\mathcal{Q}(\cdot)$ include leave-one-out error bound of SVM, mutual information $I(\mathsf{x};\mathsf{y})$, and Hilbert Space-based estimator like Hilbert-Schmidt Independence Criterion (HSIC) [2].

$$\begin{aligned} \text{maximize}_{S\subset[p]} \quad & I(\mathsf{x}_S;\mathsf{y}) \\ \text{subject to} \quad & \|\mathcal{S}\|_0 \leq t \end{aligned}$$

# 2 Instance-wise Variable Selection

The goal of instance-wise variable selection is to find a subset $\mathcal{S}(x) \subset [p]$ most informative in making inference about $\mathsf{y}$. Here $\mathcal{S} : \mathcal{X} \to \{0,1\}^d$ is a function dependent on a particular covariates $x$, and there fore $\mathcal{S}(\mathsf{x})$ is a random variable. For example, L2X maximizes the lower bound of mutual information between response and selected features [3]

$$\begin{aligned} \text{maximize}_{S} \quad & I(\mathsf{x}_{S(x)};\mathsf{y}) \\ \text{subject to} \quad & \|\mathcal{S}(x)\|_0 \leq t \end{aligned}$$

Similarly, INVASE finds $\mathcal{S}$ such that $\mathsf{y} \perp\!\!\!\perp \mathsf{x}_{S(x)} \mid \mathsf{x}_{S(x)}$ or that $p_{\mathsf{y}|\mathsf{x}}(\cdot|x) \overset{d}{=} p_{\mathsf{y}|\mathsf{x}_{S(x)}}(\cdot|x_S)$ [4],

$$\begin{aligned} \text{minimize}_{S} \quad & KL\left(p_{\mathsf{y}|\mathsf{x}}(\cdot|x)\|p_{\mathsf{y}|\mathsf{x}_{S(x)}}(\cdot|x_S)\right) \\ \text{subject to} \quad & \|\mathcal{S}(x)\|_0 \leq t \end{aligned}$$

# 3 Variable Selection as Finding Markov Blancket

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 [5]. Formally, we want to find smallest $\mathcal{S} \subset [p]$ s.t.

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

A natural interpretation is that the other variables $\mathsf{x}_{\setminus\mathcal{S}}$ do not provide additional information about $\mathsf{y}$. 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 $\mathsf{y}$. Alternatively, we can interpretate $\mathsf{x}_{\mathcal{S}}$ as minimal sufficient statistics for predicting $\mathsf{y}$. This connection exists in literature related to information bottleneck method. We can pose the problem of finding the Markov blanket of $\mathsf{y}$ as a multiple

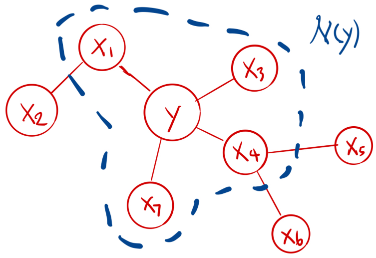


Figure 1: $\mathcal{S} = \{\mathsf{x}_1, \mathsf{x}_3, \mathsf{x}_4, \mathsf{x}_7\}$

(independent) binary hypothesis test

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

Let $\mathcal{H}_0 = \left\{ j \mid H_0^{(j)} \text{ holds} \right\}$ 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) [6].

$$\text{maximize}_{\mathcal{S}\subset[p]} \quad \mathbb{E}\left[\frac{|\mathcal{S} \setminus \mathcal{H}_0|}{|\mathcal{S}|}\right] \tag{2}$$

$$\text{subject to} \quad \mathsf{FDR} := \mathbb{E}\left[\frac{|\mathcal{S} \cap \mathcal{H}_0|}{\max\{|\mathcal{S}|, 1\}}\right] \leq q \tag{3}$$

where expectation is take w.r.t. randomness in $\mathsf{x}$ and $\mathsf{y}$. If $p_{\mathsf{y}|\mathsf{x}}(\cdot|x)$ assumes a parametric generalized linear model form,

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

Then by [7], testing for conditional independence (1) is equivalent to the following test,

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

## 4 Model-X Knockoff

Traditionally, $p_{\mathsf{y}|\mathsf{x}}$ 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 [6]. Recently, [8, 7] designed a *knockoff* framework for performing variable selection on high-dimensional nonparametric models with finite sample guarantees over the constraints in (3). The framework requires significant knowledge of $p_{\mathsf{x}}$ and assumes nothing about the $p_{\mathsf{y}|\mathsf{x}}$. This might give way to performing reproducible and robust variable selection where the $p_{\mathsf{y}|\mathsf{x}}$ is parameterized by highly complex mappings, e.g. neural networks. In addition, modeling $p_{\mathsf{x}}$ might be a suitable task for problems where we have large amount of unsupervised data, or we know a priori some structure about $p_{\mathsf{x}}$, which are often the case for large scale machine learning applications.

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

$$\tilde{\mathsf{x}} \perp\!\!\!\perp \mathsf{y} \mid \mathsf{x} \tag{4}$$

$$(\mathsf{x}, \tilde{\mathsf{x}})_{\mathsf{swap}(\mathcal{S})} \overset{d}{=} (\mathsf{x}, \tilde{\mathsf{x}}) \qquad \textit{for any} \qquad \mathcal{S} \subset [p] \tag{5}$$

*where $(\cdot)_{\mathsf{swap}(\mathcal{S})}$ swaps coordinates for all $j \in \mathcal{S}$ with coordinate $j+p$ and leaves other coordinate unchanged. Note (5) is equivalent to below*

$$(\mathsf{x}_1, \cdots, \mathsf{x}_j, \cdots, \mathsf{x}_p, \tilde{\mathsf{x}}_1, \cdots, \tilde{\mathsf{x}}_j, \cdots, \tilde{\mathsf{x}}) \overset{d}{=} (\mathsf{x}_1, \cdots, \tilde{\mathsf{x}}_j, \cdots, \mathsf{x}_p, \tilde{\mathsf{x}}_1, \cdots, \mathsf{x}_j, \cdots, \tilde{\mathsf{x}}) \tag{6}$$

*for any $j = 1, \cdots, p$. (4) is guaranteed if $\tilde{\mathsf{x}}$ is constructed without knowledge of $\mathsf{y}$.*

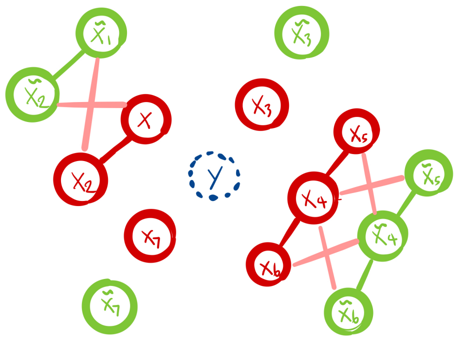

Figure 2: $\mathcal{G}$ represents $p_{\mathsf{x},\tilde{\mathsf{x}}}$. (4) implies $\tilde{\mathsf{x}}$ is not in the Markov blanket of node $\mathsf{y}$ for a graph representing joint distribution $p_{\mathsf{x},\tilde{\mathsf{x}}}$. (5) implies that the $\mathsf{x}, \tilde{\mathsf{x}}$ are pairwise 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 $\mathsf{x}_j, \tilde{\mathsf{x}}_j$ be as independent as possible, i.e. no edge connecting $\mathsf{x}_j, \tilde{\mathsf{x}}_j$ for all $j \in [p]$

Intuitively, *knockoffs* mimics the dependence structure as the original covariates $\mathsf{x}$, while being invariant to $\mathsf{swap}(\cdot)$ operation, and is independent of the response $\mathsf{y}$. 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 $\mathsf{y}$.

## 4.1 Knockoff Procedure for LASSO

Consider a linear Gaussian model $\mathsf{y} = \beta^T \mathsf{x} + \epsilon$ where $\mathsf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\epsilon \sim \mathcal{N}(0,1)$. 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 (5), it must be

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

One way to construct knockoff is to sample $\tilde{x}$ from the conditional distribution [7, 5],

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

Alternatively, we can match empirical first and second moment [8] and construct design for knockoff as

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

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

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

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

Now we compute statistics $\mathsf{w}$ for each pair of original and knockoff variables

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

which satisfy the coin-flip property. One important consequence is that for any $j \in \mathcal{H}_0$, $\mathsf{w}_j$ is a symmtric distribution about the origin [8, 7].

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

$$\tau_+ = \min\left\{t > 0 \mid \widehat{\mathsf{FDP}}(t) \le q\right\} \quad \text{where} \quad \widehat{\mathsf{FDP}} = \frac{1 + \#\{j \mid w_j \le -t\}}{\#\{j \mid w_j \ge t\}}$$

4. (**Perform Test**) with threshold $\tau_+$,

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

ensures that $\mathsf{FDR} \le q$

### 4.2 Current Problems

1. Generating knockoffs is a hard problem, there has been work to sample knockoffs using MCMC [9], with generative model [10], in particular with GAN [11] and with latent variable models [5].

2. There has been a few papers that tries to analyze power of the knockoff framework assuming Gaussian data [12]

3. There has been some preliminary work on adopting multiple hypothesis testing and FDR control when $p_{\mathsf{y}|\mathsf{x}}$ is parameterized by neural networks, e.g. by MLP [13]. 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] Le Song et al. "Feature Selection via Dependence Maximization". In: *Journal of Machine Learning Research* 13.47 (2012), pp. 1393–1434. URL: http://jmlr.org/papers/v13/song12a.html (visited on 04/24/2020).

[2] Arthur Gretton et al. "Measuring Statistical Dependence with Hilbert-Schmidt Norms". In: *Algorithmic Learning Theory*. Ed. by Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita. Red. by David Hutchison et al. Vol. 3734. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 63–77. ISBN: 978-3-540-29242-5 978-3-540-31696-1. DOI: 10.1007/11564089_7. URL: http://link.springer.com/10.1007/11564089_7 (visited on 04/24/2020).

[3] Pieter-Jan Kindermans et al. "Learning how to explain neural networks: PatternNet and PatternAttribution". In: *arXiv:1705.05598 [cs, stat]* (Oct. 24, 2017). arXiv: 1705.05598. URL: http://arxiv.org/abs/1705.05598 (visited on 01/16/2020).

[4] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. "INVASE: Instance-wise Variable Selection using Neural Networks". In: International Conference on Learning Representations. Sept. 27, 2018. URL: https://openreview.net/forum?id=BJg_roAcK7 (visited on 04/26/2020).

[5] 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).

[6] 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.

[7] 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).

[8] 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).

[9] 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).

[10] 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).

[11] 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).

[12] 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).

[13] 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).