A Feature Selection Study for Medical Image deep learning

Sanghyuck Lee

2021. 06 08.

Outline

- Deep Feature Selection study
- Thyroid Associated Ophthalmopathy Research
- Industry-University Collaboration Project

Deep Feature Selection study

- Knockoffs
- **Methodology**
    - **Problem**
    - **Outline**
    - **Details**
- Software
- Tutorials

Referenced by “https://web.stanford.edu/group/candes/knockoffs”

The Methodology of Knockoffs: Problem

The variable selection problem

The very general problem addressed by the knockoff methodology is the following. Suppose that we can observe a response Y and p potential explanatory variables $X = (X_1, \dots, X_p)$. Given n samples $(X_{i,1}, \dots, X_{i,p}, Y_i)_{i=1}^{n}$, we would like to know which predictors are important for the response.

We assume that, conditionally on the predictors, the responses are independent and the conditional distribution of Y_i only depends on its corresponding vector of predictors $(X_{i,1}, \dots, X_{i,p})$. Formally, we write this as:

$$Y_i | (X_{i,1}, \dots, X_{i,p}) \overset{\text{ind.}}{\sim} F_{Y|X}, i = 1, \dots, n,$$

for some conditional distribution $F_{Y|X}$

The Methodology of Knockoffs: Problem

The variable selection problem

Formally, we write this as: $Y_i | \left(X_{i,1}, \ldots, X_{i,p}\right) \overset{\text{ind.}}{\sim} F_{Y|X}, i = 1, \ldots, n,$

for some conditional distribution $F_{Y|X}$

> The variable selection problem is motivated by the belief that, in many practical applications, $F_{Y|X}$ actually only depends on a (small) subset $S \subset \{1, \ldots, p\}$ of the predictors, such that conditionally on $\{X_j\}_{j \in S}$, Y is independent of all other variables.

This is a very intuitive definition, that can be informally restated by saying that the other variables are not important because they do not provide any additional information about Y. A minimal set S with this property is usually known as a *Markov blanket*. Under very mild conditions on $F_{Y|X}$, this can be shown to be unique and the variable selection problem is cleanly defined.

The Methodology of Knockoffs: Problem

The variable selection problem

In order to avoid any ambiguity in those pathological cases in which the Markov blanket is not unique, we will say that the j-th predictor is *null* if and only if Y is independent of X_j, conditionally on all other predictors $X_{-j} = \{X_1, \dots, X_p\} \backslash \{X_j\}$. We denote the subset of null variables by $H_0 \subset \{1, \dots, p\}$ and call the j-th variable *relevant* (or non-null) if $j \notin H_0$.

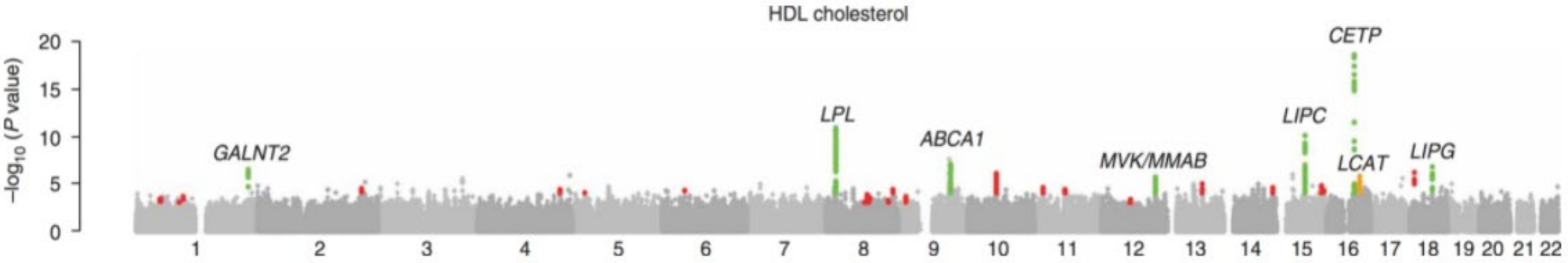
Our goal is to discover as many relevant variables as possible while keeping the false discovery rate (FDR) under control. For a selection rule that selects a subset $\hat{S}$ of the predictors, the FDR is defined as

$$FDR = \mathbb{E}\left[\frac{|\hat{S} \cap H_0|}{\max(1, |\hat{S}|)}\right].$$

The Methodology of Knockoffs: Problem

An important application

Controlled variable selection is particularly relevant in the context of statistical genetics. For instance, a genome-wide association study aims at finding genetic variants that are associated with or influence a trait, choosing from a pool of hundreds of thousands to millions of single-nucleotide polymorphisms. This trait could be the level of cholesterol or a major disease.



The Methodology of Knockoffs: Outline

Why controlled variable selection is hard

A multitude of machine learning tools have been developed for the purpose of predicting a response variable from a vector of covariates, inspiring sophisticated variable selection techniques. To name a few examples, think of penalized regression or non-parametric methods based on trees and ensemble learning. Many of these techniques are extremely popular and enjoy wide applications.

This begs the following question: how do we make sure that we do not select too many null variables? In statistical terms, how do we control for a Type-I error?

For simplicity, consider for a moment the typical example of the lasso:

$$\hat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}$$

The Methodology of Knockoffs: Outline

Why controlled variable selection is hard

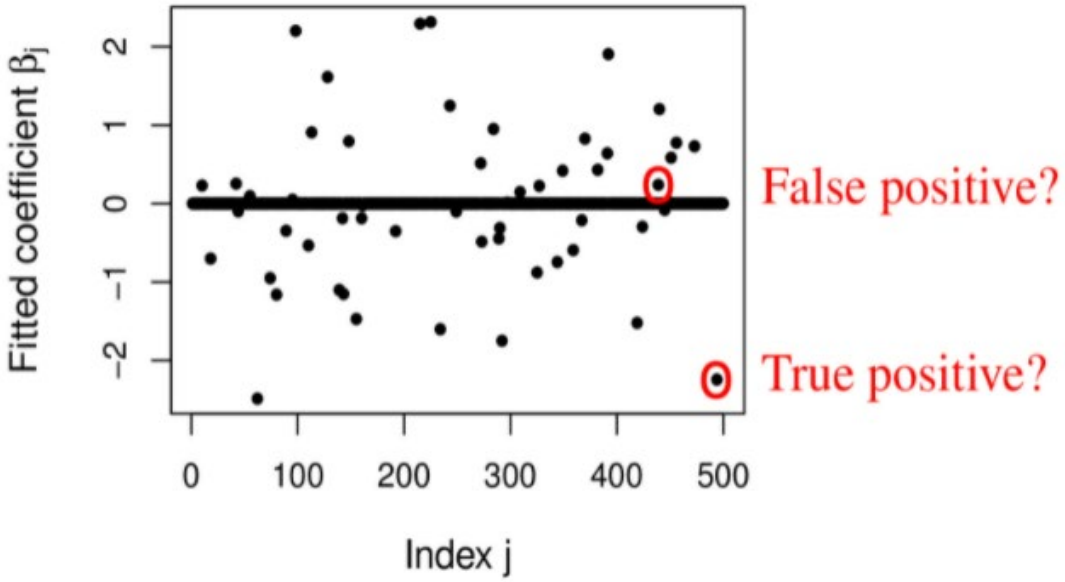
$$\hat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}$$

By varying the regularization parameter λ (perhaps tuning it with cross-validation), we obtain different models in which more or less variables have a non-zero coefficient. Intuitively, it seems reasonable to select variables whose fitted coefficient is (in absolute value) above some significance threshold. However, it is not an easy task to choose the threshold (or the value of λ) in such a way as to control the Type-I error.

The Methodology of Knockoffs: Outline

Why controlled variable selection is hard

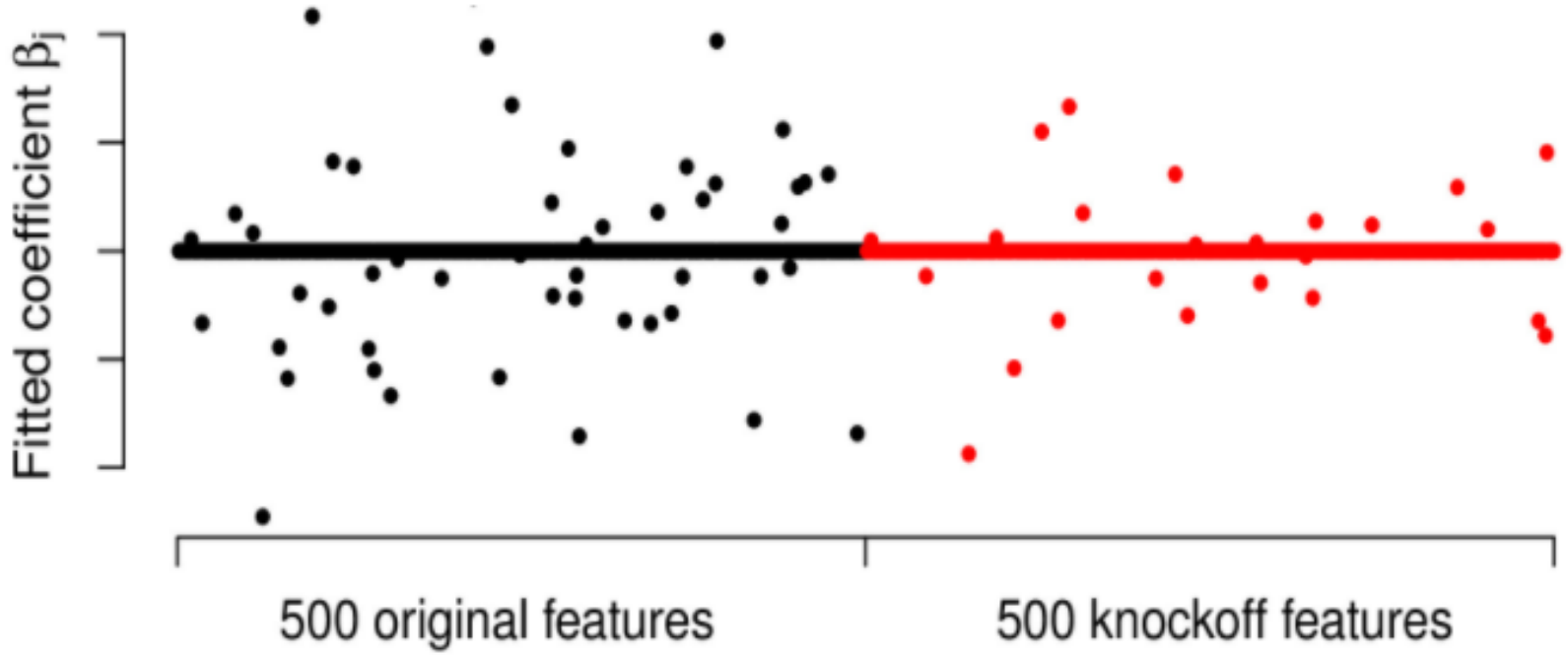
The difficulty arises from the distribution of the estimated coefficients for the null variables being unknown, but at least some of them most likely being non-zero. Moreover, the fitted coefficients are correlated among each other and an incorrect threshold can yield either a very high proportion of false discoveries (if too low) or very low power (if too high).



The Methodology of Knockoffs: Outline

How knockoffs work

Knockoffs solve the controlled variable selection problem by providing a *negative control group* for the predictors that behaves in the same way as the original null variables but, unlike them, is known to be null.



The Methodology of Knockoffs: Outline

How knockoffs work: The construction of the knockoffs

A rigorous definition of knockoffs depends on some additional modeling choices, but intuitively their nature can be understood as follows.

What knockoffs are

For each observation of a predictor X_j, we construct a knockoff copy $\tilde{X}_j$, without using any additional data, such that:
1. the correlation between distinct knockoffs $\tilde{X}_j$ and $\tilde{X}_k$ (for $j \neq k$) is the same as the correlation between $\tilde{X}_j$ and $\tilde{X}_k$,
2. the correlation between X_j and $\tilde{X}_k$ (for $j \neq k$) is the same as the correlation between the original variables X_j and X_k,
3. the knockoffs are created without looking at Y.

The Methodology of Knockoffs: Outline

How knockoffs work: The construction of the knockoffs

By construction, the knockoffs are not important for Y, since they are created without even looking at it (as long as the original predictors X are not removed from the model).

What knockoffs do

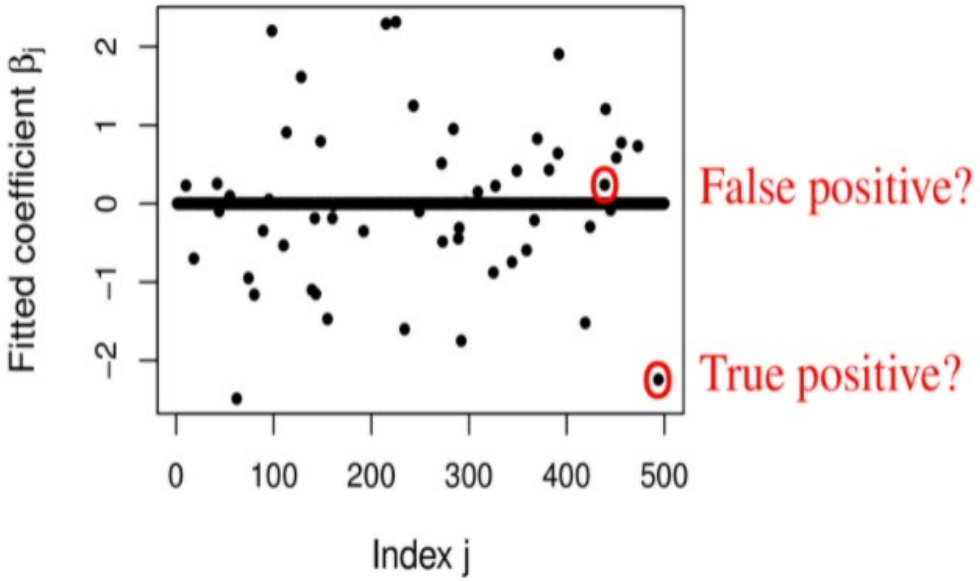
> Knockoffs can be used as *negative controls* for the original variables. The true importance of an explanatory variable X_j can be deduced by comparing its predictive power for Y to that of its knockoff copy $\tilde{X}_j$.

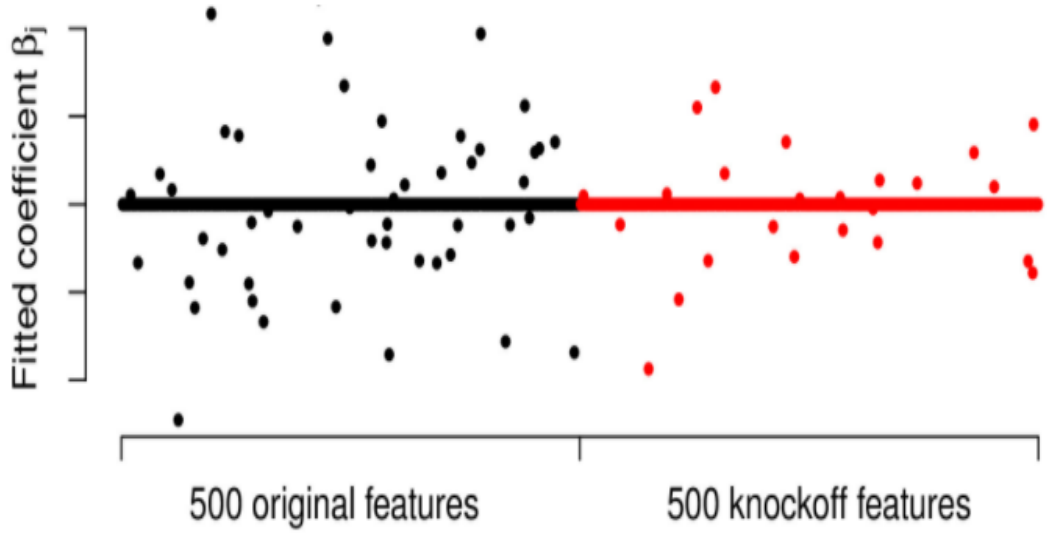
The Methodology of Knockoffs: Outline

How knockoffs work: The construction of the knockoffs

Why controlled variable selection is hard

The difficulty arises from the distribution of the estimated coefficients for the null variables being unknown, but at least some of them most likely being non-zero. Moreover, the fitted coefficients are correlated among each other.





The Methodology of Knockoffs: Outline

How knockoffs work: Measuring variable importance

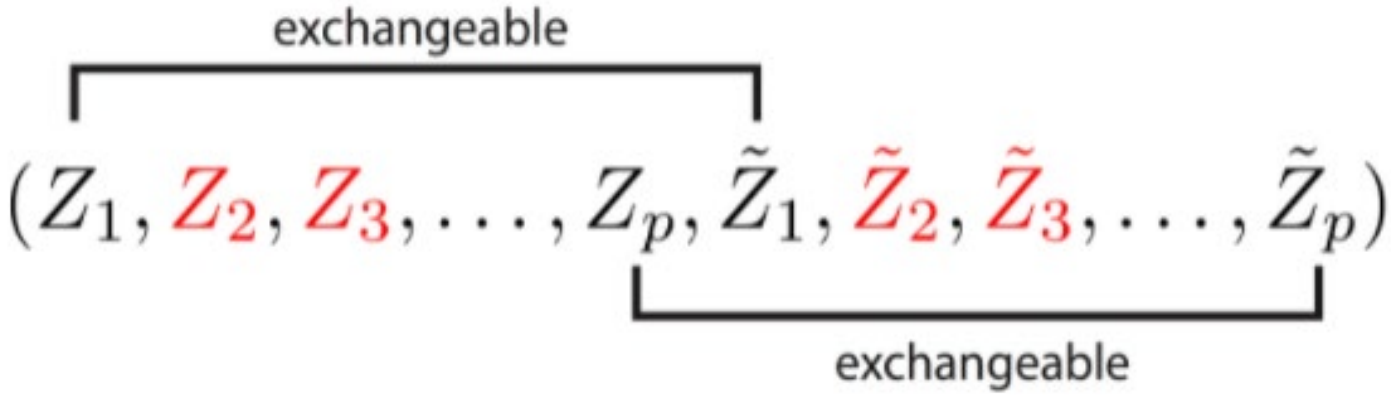
Having constructed the knockoffs, we can proceed by applying a traditional variable selection procedure on the augmented set of predictors $[X\tilde{X}]$. For each $j = 1, \ldots, p$, we compute statistics Z_j and $\tilde{Z}_j$ measuring the importance of X_j and $\tilde{X}_j$, respectively, in predicting Y. For instance, we can apply the lasso on the augmented data and compute:

$$Z_j = \left|\hat{\beta}_j(\lambda)\right|, \qquad \tilde{Z}_j = \left|\hat{\beta}_{j+p}(\lambda)\right|, \qquad j = 1, \ldots, p.$$

The Methodology of Knockoffs: Outline

How knockoffs work: Measuring variable importance

The knockoff filter works by comparing the Z_j's to the $\tilde{Z}_j$'s and selecting only variables that are clearly better than their knockoff copy. The reason why this can be done is that, by construction of the knockoffs, the null statistics are pairwise exchangeable. This means that swapping the Z_j and $\tilde{Z}_j$'s corresponding to null variables leaves the joint distribution of $(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p)$ unchanged.



The Methodology of Knockoffs: Outline

How knockoffs work: Measuring variable importance

Despite the simple example that we presented, the knockoffs procedure is by no means restricted to statistics based on the lasso, as many other options are available for assessing the importance of X_j and $\tilde{X}_j$. In general, it is required that the method used to compute the Z_j and $\tilde{Z}_j$'s satisfy a *fairness* requirement, so that swapping $\tilde{X}_j$ with X_j would have the only effect of swapping Z_j with $\tilde{Z}_j$. However, under certain modeling scenarios, an additional *sufficiency* requirement is also required.

The Methodology of Knockoffs: Outline

How knockoffs work: Measuring variable importance

Once the Z_j and $\tilde{Z}_j$'s have been computed, different contrast functions can be used to compare them. In general, we must choose an anti-symmetric function h and we compute the *symmetrized knockoff statistics*

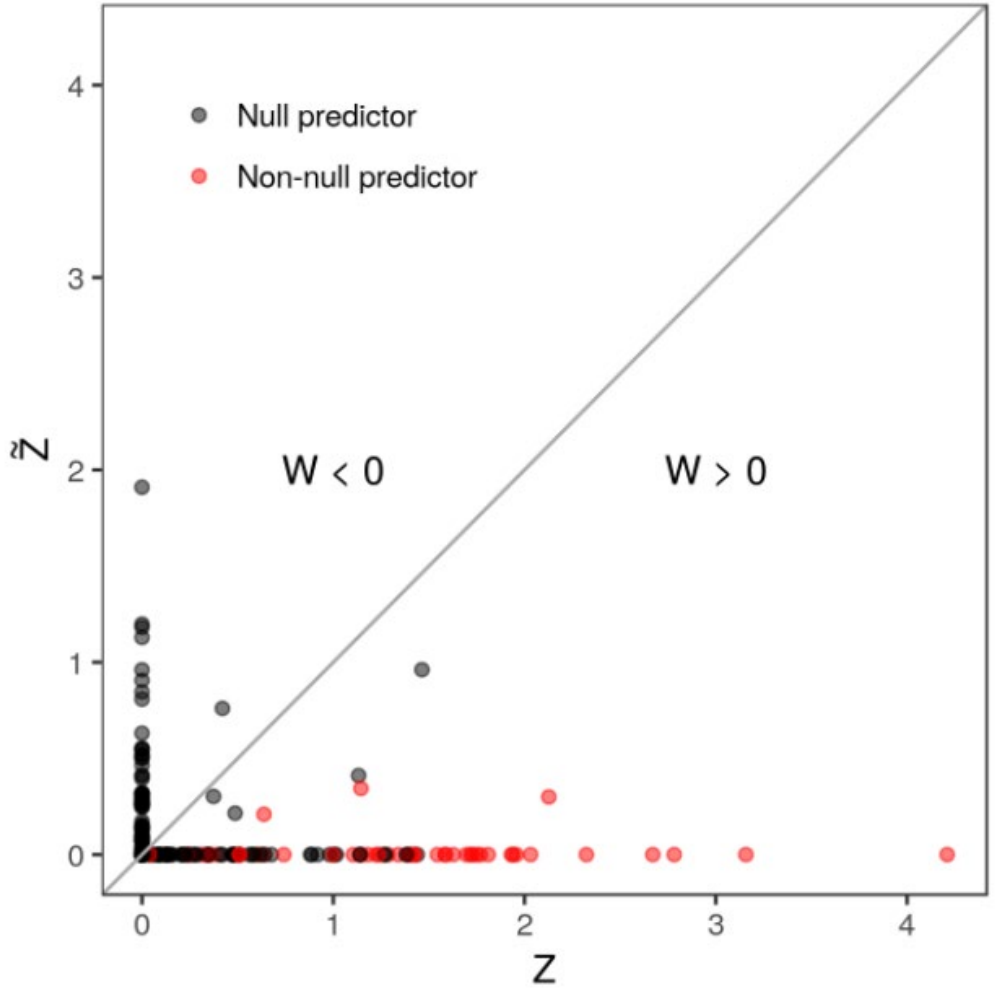
$$W_j = h(Z_j, \tilde{Z}_j) = -h(\tilde{Z}_j, Z_j), \qquad j = 1, \ldots, p$$

such that $W_j > 0$ indicates that X_j appears to be more important than its own knockoff copy. A simple example may be $W_j = Z_j - \tilde{Z}_j$, but many other alternatives are possible.

The Methodology of Knockoffs: Outline

How knockoffs work: Measuring variable importance

The figure below shows one realization of the random vector $(Z_j, \tilde{Z}_j)$, and it seems apparent that the nulls (black dots) are distributed symmetrically around the diagonal. For the non-nulls (red dots) on the other hand, we see a propensity for Z_j to be larger than its "control value" $\tilde{Z}_j$ (points below the diagonal).



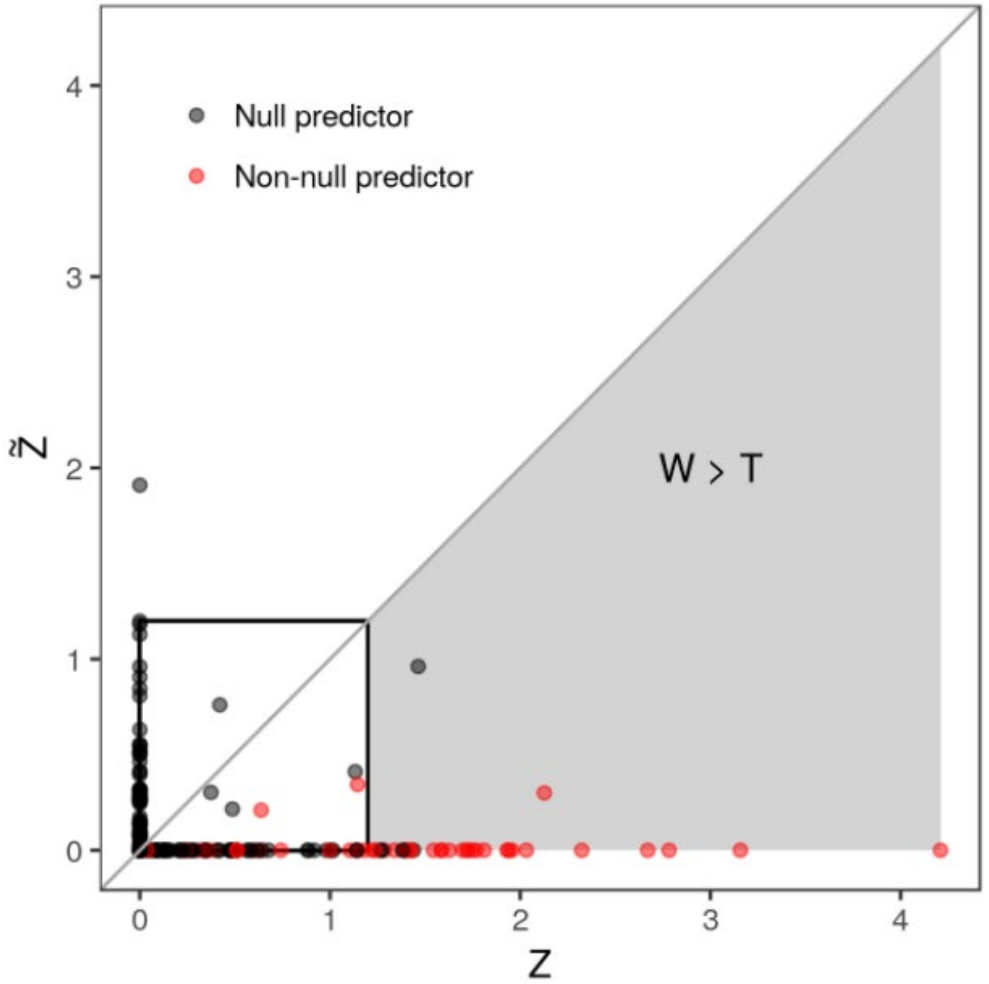
The Methodology of Knockoffs: Outline

How knockoffs work: A data-adaptive significance threshold

Finally, the knockoff procedure selects predictors with large and positive values of W_j, according to the adaptive threshold defined as

$$T = \min\left\{t: \frac{1 + \#\{j: W_j \leq -t\}}{\#\{j: W_j > t\}} \leq \alpha\right\},$$

where α is the (desired) target FDR level.

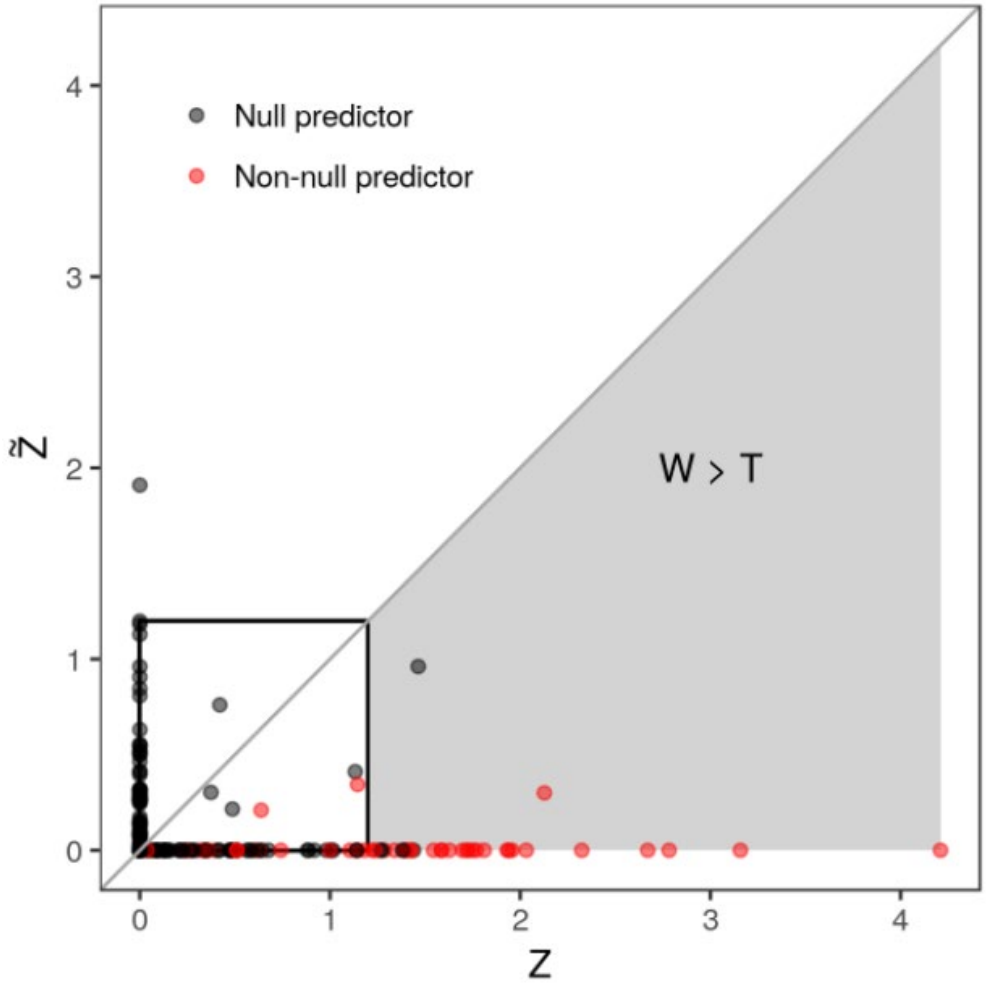


The Methodology of Knockoffs: Outline

How knockoffs work: A data-adaptive significance threshold

$$T = \min\left\{t: \frac{1 + \#\{j: W_j \leq -t\}}{\#\{j: W_j > t\}} \leq \alpha\right\},$$

Intuitively, the reason why this procedure can control the FDR is that, by the the exchangeability property of the null Z_j and $\tilde{Z}_j$'s, it can be shown that the signs of the null W_j's are independent coin flips, conditional on the absolute values $|W|$. Consequently, it can be shown that the fraction inside the definition of the adaptive threshold T is a conservative estimate of proportion of false discoveries.



The Methodology of Knockoffs: Outline

How knockoffs work: A data-adaptive significance threshold

$$T = \min\left\{t: \frac{1 + \#\{j: W_j \le -t\}}{\#\{j: W_j > t\}} \le \alpha\right\},$$

$$W_j = Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1, & Z_j > \tilde{Z}_j \\ -1, & Z_j < \tilde{Z}_j \end{cases}$$

where $\vee$ is max operator.

$$\#\{\text{null } j: W_j \le -t\} \overset{d}{=} \#\{\text{null } j: W_j \ge t\}$$

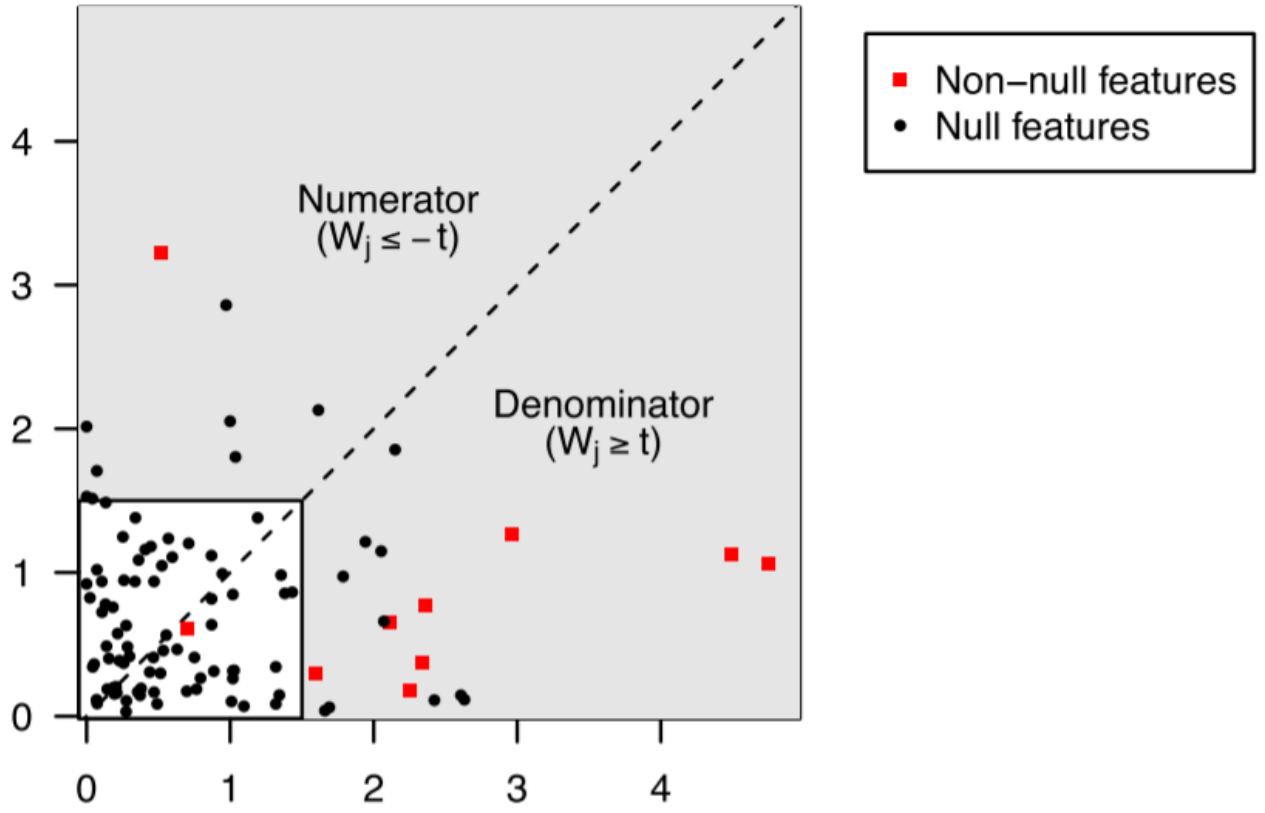


Figure: representation of $(Z_j, \tilde{Z}_j)$ on threshold $t = 1.5$

Thyroid Associated Ophthalmopathy Research

- 완료 현황: 각각 1323명 환자 (기존에는 1267 환자)

유형	눈 (L)	눈 (R)	시신경 (L)	시신경 (R)	내직근 (L)	내직근 (R)	외직근 (L)	외직근 (R)	상직근 (L)	상직근 (R)	하직근 (L)	하직근 (R)	상안검 (L)	상안검 (R)	지방 (L)	지방 (R)	눈물주머니 (L)	눈물주머니 (R)	안와 (L)	안와 (R)	평균
AX 1	0.85	0.81	0.42	0.47	0.64	0.63	0.73	0.66							0	0					0.52
AX 2	0.91	0.89	0.75	0.72	0.93	0.93	0.93	0.9							0	0					0.70
AX 3	0.96	0.96	0.95	0.94	0.95	0.96	0.94	0.95							0.08	0.08					0.78
AX 4	0.94	0.94	0.82	0.84	0.82	0.84	0.79	0.75							0	0					0.67
AX 5	0.81	0.81	0.36	0.37	0.34	0.36	0.32	0.38							0	0					0.38
AX 6																	0	0			0.00
CO 1	0.96	0.96																			0.96
CO 2			0.89	0.89	0.91	0.91	0.91	0.89	0.86	0.9	0.91	0.88									0.90
CO 3																			0.08	0.08	0.08
SA 1		0.96		0.92						0.96		0.93		0.96							0.95
SA 2	0.96		0.92						0.96		0.93		0.96								0.95
계																					0.67

0% 100%

Industry-University Collaboration Project

- 제출 실적
  - SW 제작 (산학 1), 국제학술대회 논문 1편 이상 게재 (산학 2), 국내 특허출원 1건 이상 (산학 2)
- 현재 진행 사항
  - SW 등록: 진행중
  - 특허 출원: 기존 논문 번역본 제작 중
  - 투고 저널 리스트업 완료

References

Barber, Rina Foygel, and Emmanuel J. Candès. "Controlling the false discovery rate via knockoffs." *Annals of Statistics* 43.5 (2015): 2055-2085.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995): 289-300.

Chen, Jiajie, Anthony Hou, and Thomas Y. Hou. "Some analysis of the knockoff filter and its variants." *arXiv preprint arXiv:1706.03400* (2017).

Jordon, James, Jinsung Yoon, and Mihaela van der Schaar. "KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks." *International Conference on Learning Representations*. 2018.

Sarkar, Sanat K., and Cheng Yong Tang. "Adjusting the Benjamini-Hochberg method for controlling the false discovery rate in knockoff assisted variable selection." *arXiv preprint arXiv:2102.09080* (2021).