# A Feature Selection Study for Medical Image deep learning

Sanghyuck Lee

2021. 05 04.

# Outline

- Knockoff study
- Thyroid Research Project Report

# Knockoff study

- Input: $X, y, q, O, P$ where X is design matrix, y is response vector, q is target FDR level, $O$ is a optimization strategy for calculating $s$ and P is a calculating method for feature statistics.
- Output: Not-null feature set $\hat{S}$

Step 1: Calculate knockoffs

$$\tilde{X}(X, O) = X(1 - (X^T X)^{-1}\text{diag}\{s\}) + \tilde{U}C.$$

Step 2: Calculate feature statistics $W\left(X, \tilde{X}, y, P\right) = P(X, \tilde{X}, y) = W_1, \dots, W_p)$.

Step 3: Calculate threshold $\widehat{\mathrm{T}}$

$$\widehat{\mathrm{T}}(W, q) = \min\{\mathrm{T} \in \{|W_i| : i \in [p]\}\{0\} : \frac{|\{\, i \in [p] : W_i \leq -\mathrm{T}\}|}{|\{\, i \in [p] : W_i \geq \mathrm{T}\}| \vee 1} \leq \mathrm{q}\}$$

if null, $\widehat{\mathrm{T}} = \infty$,

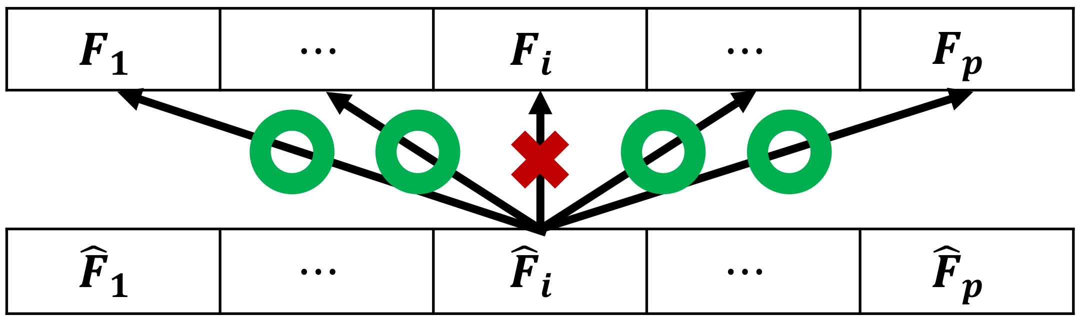
Finally, $\hat{S} = \{i \in [p] : W_i \geq \widehat{\mathrm{T}}\}$

# Knockoff study

- $[X\tilde{X}]^T[X\tilde{X}] = \begin{vmatrix} X^TX & X^T\tilde{X} \\ \tilde{X}^TX & \tilde{X}^T\tilde{X} \end{vmatrix} = \begin{vmatrix} X^TX & X^TX - \mathrm{diag}\{s\} \\ X^TX - \mathrm{diag}\{s\} & X^TX \end{vmatrix}$

- $\tilde{X}_j^T\tilde{X}_k = X_j^TX_k$ for $\forall\ j,\ k$

- $\tilde{X}_j^TX_k = X_j^TX_k$ for $\forall\ j \neq k$



# Knockoff study

Equi-correlated knockoffs:

$s_j = 2\lambda_{min}(\Sigma) \wedge 1$, where $\tilde{X}_j^T X_j = 1 - s_j, \mathrm{diag}\{s\} \leq 2X^T X$

=> minimize $|(\tilde{X}_j^T \quad X_j)|$

$$[X\tilde{X}]^T[X\tilde{X}] = \begin{vmatrix} X^T X & X^T \tilde{X} \\ \tilde{X}^T X & \tilde{X}^T \tilde{X} \end{vmatrix} = \begin{vmatrix} X^T X & X^T X - \mathrm{diag}\{s\} \\ X^T X - \mathrm{diag}\{s\} & X^T X \end{vmatrix}$$

$$\tilde{X}(X, O) = X(1 - (X^T X)^{-1}\mathrm{diag}\{s\}) + \tilde{U}C.$$

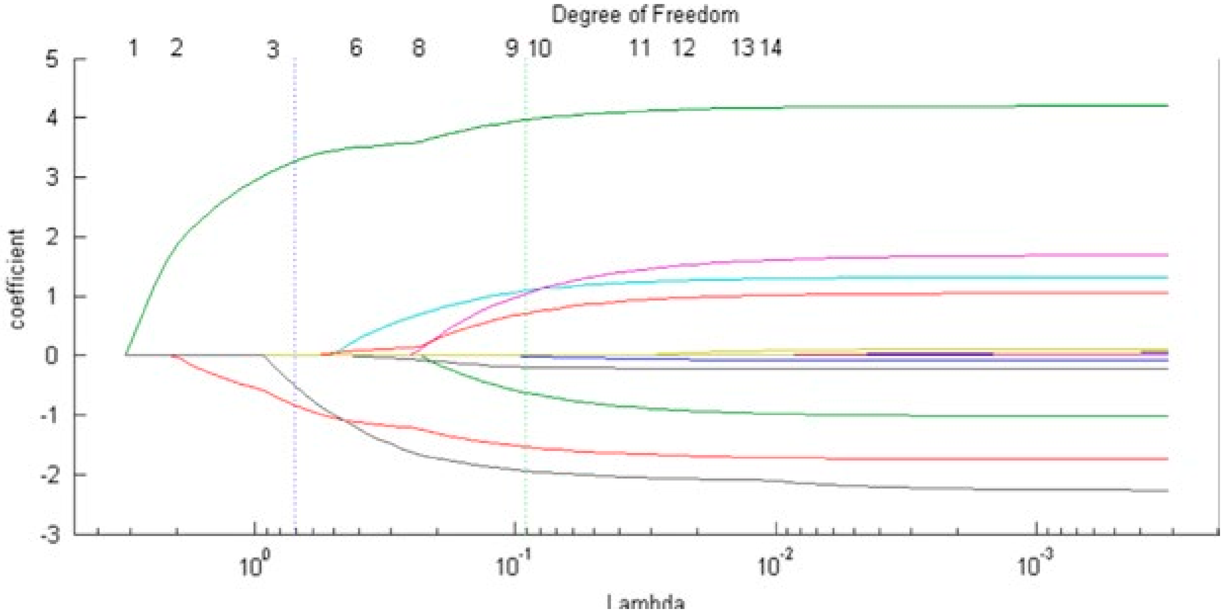
where $\tilde{U}$ is orthonormal to the column space of $X$

and $C^T C = 2diag\{s\} - \mathrm{diag}\{\mathrm{s}\}\,(X^T X)^{-1}$

# Knockoff study

$W_j$ = (largest $\lambda$ such that $X_j$ or $\tilde{X}_j$ enters Lasso path)

$$\times \begin{cases} +1 & \text{if } X_j \text{ enters before } \tilde{X}_j \\ -1 & \text{if } X_j \text{ enters after } \tilde{X}_j \end{cases}$$

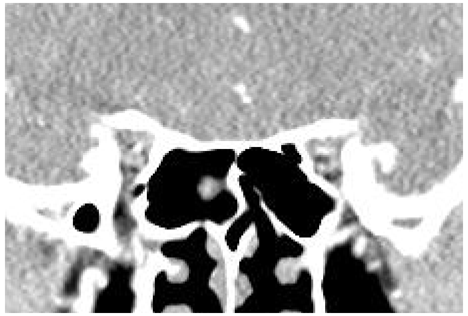
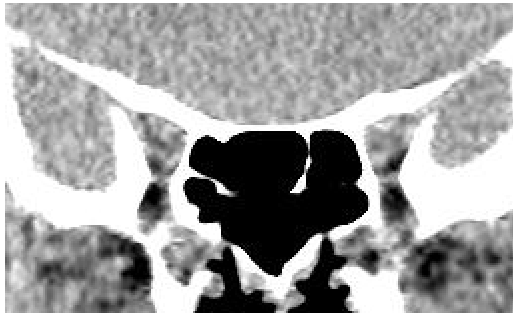
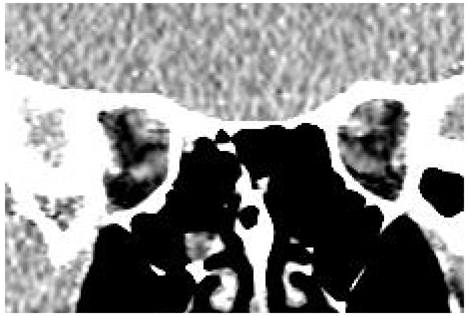



$$minimize\ \frac{1}{n}||y - X\beta||_2^2 + \lambda||\beta||_1$$

$$\widehat{\mathrm{T}}(W, q) = \min\{\mathrm{T} \in \{|W_i| : i \in [p]\}\{0\} : \frac{|\{i \in [p] : w_i \leq -\mathrm{T}\}|}{|\{i \in [p] : w_i \geq \mathrm{T}\}| \vee 1} \leq \mathrm{q}\}$$

# 갑상선 연구 과제 보고

- 1273 환자에 대해 CO2, CO3 선정 완료
- CO2: 근육의 ½ 지점, CO3: 아랫쪽 V자 뼈의 모서리 지점
- CO3: 1) 근육 2/3 지점으로 변경하는 것은 어떨지

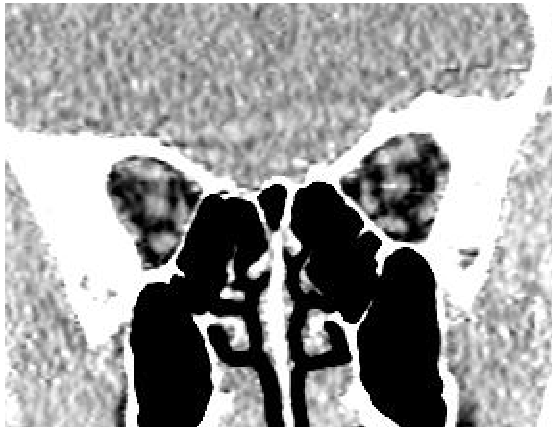





- 근육 2/3

: 근육의 길이에 비례해 선택하여, 안와의 크기가 꽤 의미있게 선택 지점 비율에 의거하여, 줄어드는 것처럼 보이도록 슬라이스를 선택함 -> 비슷한 안와 크기에서 근육이 신경을 누르는지 확인할 수 있음

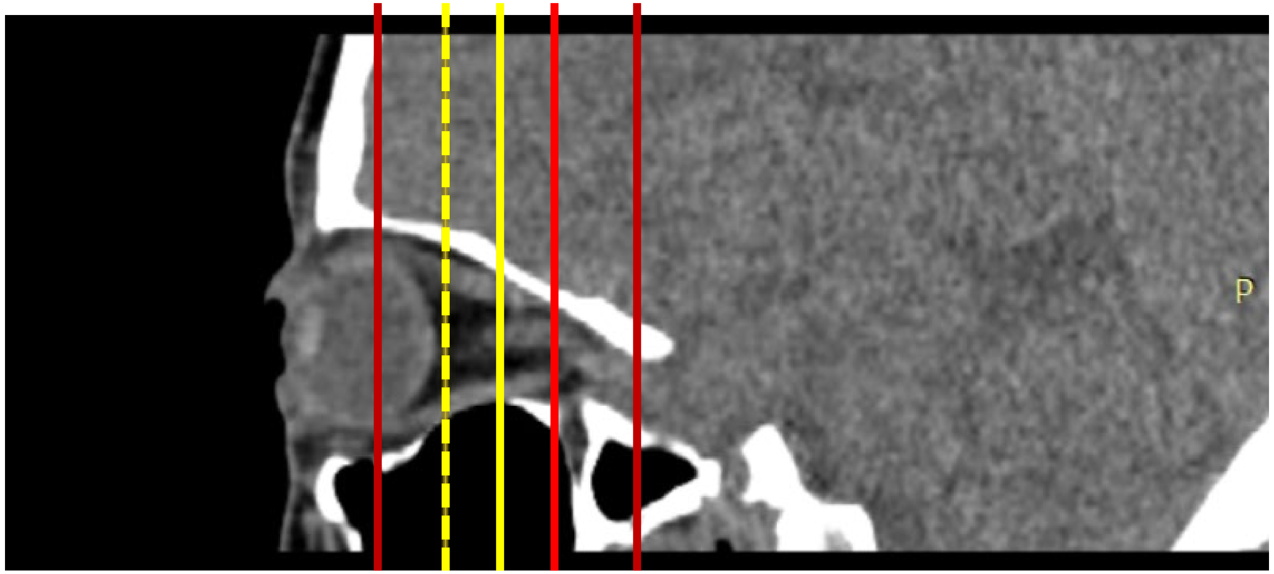
# 갑상선 연구 과제 보고

- CO3: 2) 안와 전체만 세그멘테이션 진행하는 것은 어떤지
  - 4근육+신경이 SFS에 입력되어야할 것으로 보임
  - 맞다면, 기존 방식은 세그멘테이션 오류가 발생할 수 있고, 이는 성능에 꽤 큰 영향을 줄 수 있을 것 같음
  - 세그멘테이션 작업에 효율성



# 갑상선 연구 과제 보고

- CO2) ½지점 -> 눈 바로 뒤로 변경
  - 식별이 더 명확해짐
  - CO3과 함께 예측을 진행할 경우,
    근육 전체를 더 잘 반영할 수 있음




# 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).

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).