Automated feature selection of predictors in electronic medical records data

Gronsbell, Jessica, Jessica Minnier, Sheng Yu, Katherine Liao, and Tianxi Cai. “Automated Feature Selection of Predictors in Electronic Medical Records Data.” *Biometrics* 75, no. 1 (2019): 268–77. https://doi.org/10.1111/biom.12987.

Jianhui Gao

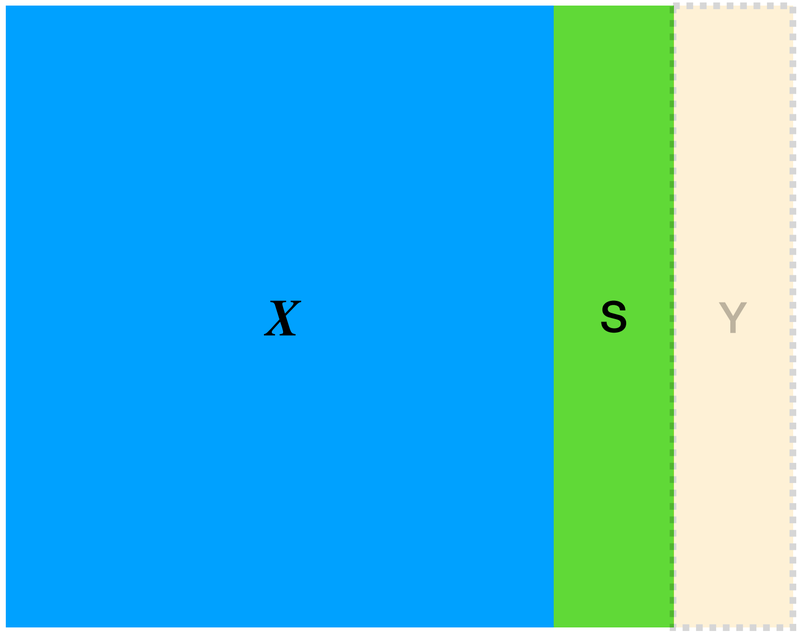
EHR Journal Club Nov 23, 2021

Agenda

- (Quick) Introduction
- Proposed Methods
- Simulation Study
- Data Analysis
- Discussion

Introduction

EHR Data



- X: wide range of candidate features
- Interested in true disease status Y, but…
  - Chart review is labor-intensive
- S: easily extracted from EHR data and closely related to y

Can we accurately extract features that are predictive of Y using only S?

Methods

Notation Set-up

- N *i.i.d.* random vectors $\mathcal{F} = \{(Y_i, \boldsymbol{X}_i^T, \boldsymbol{S}_i^T)^T, i = 1,..,N\}$
- only $\mathcal{D} = \{\boldsymbol{W}_i = (\boldsymbol{X}_i^T, \boldsymbol{S}_i^T)^T, i = 1,...,N\}$ is observed
- Assumptions
  - $P(Y = 1 \mid \boldsymbol{X}) = g(\alpha_0 + \boldsymbol{X}^T\boldsymbol{\beta}_0) = g(\overrightarrow{\boldsymbol{X}}^T\boldsymbol{\theta}_0)$; Y follows a GLM
    $\overrightarrow{\boldsymbol{X}} = (1, \boldsymbol{X}^T)^T, \boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0^T)^T$
  - $\boldsymbol{S} \perp \boldsymbol{X} \mid Y$ $\boldsymbol{S}$ depends on $\boldsymbol{X}$ only through Y
  - $\boldsymbol{X}$ is elliptical symmetric

Methods

Unsupervised feature selection procedure

- Step I :
  - Estimate $\pi_S = P(Y = 1 \mid \boldsymbol{S})$
- Step II:
  - Penalized regression $\hat{\pi}_S$ against $\boldsymbol{X}_i$
  - Candidate features are in $\widehat{\mathscr{A}} = \{j : \hat{\beta}_j \neq 0\}$

Methods

Step I: Clustering

$S \sim \tau f_{\boldsymbol{\Theta}_1} + (1-\tau) f_{\boldsymbol{\Theta}_0}$ *working* parametric mixture model

- $\tau := P(Y = 1)$
- $S \mid Y \sim f_{\boldsymbol{\Theta}_y}$; $f_{\boldsymbol{\Theta}_y}$ is specified up to unknown parameters $\boldsymbol{\Theta}_y$; e.g. $S \mid Y = 1 \sim N(\boldsymbol{\mu}_1, \Sigma_1)$
- unknown $\boldsymbol{\Theta}_\cdot = (\boldsymbol{\Theta}_1^T, \boldsymbol{\Theta}_0^T, \tau)^T$

$\widehat{\boldsymbol{\Theta}_\cdot}^{MLE} = argmin_{\boldsymbol{\Theta}_\cdot}\{-N^{-1}\sum_{i=1}^{N} l(\boldsymbol{\Theta}_\cdot \mid S_i)\}$ can be found using EM alogirthm

$$\hat{\pi}_S = \frac{P(Y=1,S)}{P(S)} = \frac{\hat{\tau} f_{\hat{\boldsymbol{\Theta}}_1}}{\hat{\tau} f_{\hat{\boldsymbol{\Theta}}_1} + (1-\hat{\tau}) f_{\hat{\boldsymbol{\Theta}}_0}}$$

Methods

Step II: Regularized Estimation

$$\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\boldsymbol{\beta}}^T)^T = argmin_{\boldsymbol{\theta}} \Bigg\{ N^{-1} \sum_{i=i}^{N} \underbrace{l(\boldsymbol{\theta}^T \overrightarrow{\boldsymbol{X}}_i \,|\, \hat{\pi}_{S_i})}_{\text{negative likelihood}} + \underbrace{\lambda_N \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|}}_{\text{adaptive lasso penalty}} \Bigg\}$$

let $\tilde{\boldsymbol{\theta}}$ be the MLE of the average negative likelihood.

$$\approx argmin_{\boldsymbol{\theta}} \Bigg\{ (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \Sigma_{\boldsymbol{X}}^{-1} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \lambda_N \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|} \Bigg\}$$

Methods

Summary

- $E(\bar{\pi}_S \mid \boldsymbol{X}) \approx g(\boldsymbol{\theta}_0^T \overrightarrow{\boldsymbol{X}}) = E(Y \mid \boldsymbol{X})$ provided that $E(\bar{\pi}_S \mid Y = y) \approx y$
- Fitting $\hat{\pi}_S$ (the "error-corrupted" version of Y) can be viewed as mis-specification of link function (Neuhaus, 1999)
- Even mixture model fails to hold, if $\boldsymbol{X}$ follows elliptical symmetric distribution, then
  - $E(\hat{\boldsymbol{\beta}}) = c\boldsymbol{\beta}_0$ provided that $cov(\bar{\pi}_S, \boldsymbol{\theta}_0^T \overrightarrow{\boldsymbol{X}}) \neq 0$ (Li and Duan, 1989)
  - $P(\widehat{\mathscr{A}} = \mathscr{A}) \rightarrow 1$ as $N \rightarrow \infty$ (Appendix)

Neuhaus, John M. "Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression." *Biometrika* 86, no. 4 (1999): 843–55.
Li, Ker-Chau, and Naihua Duan. "Regression Analysis Under Link Violation." *The Annals of Statistics* 17, no. 3 (September 1989): 1009–52. https://doi.org/10.1214/aos/1176347254.

Methods

Variable selection via resampling

- $\boldsymbol{X}_{\widehat{\mathscr{A}}}$ does not perform as well as $\boldsymbol{X}_{\mathscr{A}}$ due to uncertainty in $\widehat{\mathscr{A}}$
- Let $\hat{\boldsymbol{\theta}}^{(m)}$ be the minimizer solved in m^{th} **subsample** of size N_b
  - $\hat{\rho}_{0j} := P(\hat{\beta}_j = 0) = \mathbb{P}_m(I\{\hat{\beta}_j^{(m)} = 0\})$
  - Variable selected if $\hat{\rho}_{0j} < 0.5$

Simulations

Methods compared

- Suppose training data $n = 100$, validation set $N = 5000$, two surrogate marker $\boldsymbol{S}$
- AUC is used to compare prediction performance
- Supervised approach:
    - L_{100}: $Y \sim \boldsymbol{X} + \boldsymbol{S}$ using ALASSO
    - L_{100}^{2step}: $Y \sim \boldsymbol{X}_{\mathscr{A}} + \boldsymbol{S}$; $\boldsymbol{X}_{\mathscr{A}}$ selected using $Y \sim \boldsymbol{X}$
- Proposed unsupervised approach:
    - AutoClust: $Y \sim \boldsymbol{X}_{\mathscr{A}} + \boldsymbol{S}$; $\boldsymbol{X}_{\mathscr{A}}$ selected using $\hat{\pi}_{\boldsymbol{S}} \sim X$
    - AutoClust$_R$: $Y \sim \boldsymbol{X}_{\mathscr{A}} + \boldsymbol{S}$; $\boldsymbol{X}_{\mathscr{A}}$ selected using $\hat{\pi}_{\boldsymbol{S}} \sim X$ with resampling

Simulations

Methods compared

- Other unsupervised approach:
  - PenReg$_{S_1+S_2}$: $Y \sim \boldsymbol{X}_{\mathscr{A}} + \boldsymbol{S}$; $\boldsymbol{X}_{\mathscr{A}}$ selected using $S_1 + S_2 \sim \boldsymbol{X}$
  - PenReg$_{\boldsymbol{S}}$: $Y \sim \boldsymbol{X}_{\mathscr{A}} + \boldsymbol{S}$; $\boldsymbol{X}_{\mathscr{A}}$ selected using $\boldsymbol{S} \sim \boldsymbol{X}$
  - RankCor$_{S_1+S_2}$: $Y \sim \boldsymbol{X}_{\mathscr{A}} + \boldsymbol{S}$; $\boldsymbol{X}_{\mathscr{A}}$ selected based on rank correlation [1]
  - Extreme: $Y \sim \boldsymbol{X}_{\mathscr{A}} + \boldsymbol{S}$; $\boldsymbol{X}_{\mathscr{A}}$ selected based on extreme sampling [2]

1. Yu, Sheng, Katherine P Liao, Stanley Y Shaw, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S. Kohane, and Tianxi Cai. “Toward High-Throughput Phenotyping: Unbiased Automated Feature Extraction and Selection from Knowledge Sources.” *Journal of the American Medical Informatics Association* 22, no. 5 (September 1, 2015): 993–1000. https://doi.org/10.1093/jamia/ocv034.
2. Yu, Sheng, Abhishek Chakrabortty, Katherine P Liao, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, et al. “Surrogate-Assisted Feature Extraction for High-Throughput Phenotyping.” *Journal of the American Medical Informatics Association: JAMIA* 24, no. e1 (April 2017): e143–49. https://doi.org/10.1093/jamia/ocw135.

Simulations

Setting 1: Correct model specifications

Generate data:

- $Y \sim Bin(N = 5000, p = 0.3)$
- $\boldsymbol{X}_i \sim MVN(y_i \Sigma^{\boldsymbol{X}} \boldsymbol{\beta}_0, \Sigma^{\boldsymbol{X}})$ $logit(P(Y = 1 \mid \boldsymbol{X})) = (\alpha_0 + \boldsymbol{\beta}_0^T)\boldsymbol{X}$
- $\boldsymbol{S}_i^0 \sim MVN(0, \Sigma^{\boldsymbol{S}}) + \boldsymbol{\mu}^{\boldsymbol{S}} + y_i \Delta_0^{\boldsymbol{S}}$ $\boldsymbol{S} \perp \boldsymbol{X} \mid Y$
- $\boldsymbol{S} = log\{\lfloor \exp(\boldsymbol{S}^0) \rceil + 1\}$ Count Data; log-transformation to stabilize model fitting and standardize unit variance

Parameter values:

- Strong signals: $\boldsymbol{\beta}_0 = [1.2, -1.2, 0.5, -0.3, 0.3, 0.1, 0.1, \mathbf{0}^T_{(p-7)\times 1}]^T$, $\Delta_0^{\boldsymbol{S}} = [0.75, 0.3]^T$

TABLE 1 Percent of times each feature was selected over 500 replications as well as the average AUC attained for final algorithm training with 100 or 200 labeled examples (AUC_{100}, AUC_{200}) when the covariate distribution is elliptically symmetric for the strong signal based on (i) supervised training with all features and 100 or 200 labeled samples (L_{100}, L_{200}) or with features selected via the two-step approach (L_{100}^{2step}, L_{200}^{2step}) (ii) the automated feature selection method with and without resampling (AutoClust, $AutoClust_R$), (iii) a penalized regression with $S_1 + S_2$ or **S** as the outcome ($PenReg_{S_1+S_2}$, $PenReg_{\mathbf{S}}$), (iii) the rank correlation method based on $S_1 + S_2$ ($RCor_{S_1+S_2}$) (iv) the extreme sampling method (Extreme) and (iv) the method of Agarwal et al. (2016) based on $S_1 + S_2 \geq 1$ as the silver standard label ($Agarwal_{S_1+S_2\geq 1}$).

(a) $p = 50$

Method	1.2	−1.2	0.5	−0.3	0.3	0.1	0.1	0	AUC_{100}	AUC_{200}
L_{100}	85	46	3	6	14	11	9	3	79.9	
L_{100}^{2step}	88	59	4	5	19	13	9	4	81.1	
L_{200}	100	99	26	12	46	32	28	10		84.5
L_{200}^{2step}	100	99	32	12	47	33	27	11		84.3
AutoClust	100	100	91	43	66	28	29	12	84.2	86.3
$AutoClust_R$	100	100	74	15	42	12	10	2	85.3	86.6
$PenReg_{S_1+S_2}$	100	100	100	96	97	65	66	45	81.9	85.4
$PenReg_{\mathbf{S}}$	100	100	99	65	96	79	72	34	82.4	85.6
$RankCor_{S_1+S_2}$	36	0	0	0	0	0	0	0	79.7	80.2
Extreme	100	95	1	0	6	4	1	0	85.3	86.2
$Agarwal_{S_1+S_2\geq 1}$	100	100	99	66	85	38	38	14	83.9	86.2

(b) $p = 100$

Method	1.2	−1.2	0.5	−0.3	0.3	0.1	0.1	0	AUC_{100}	AUC_{200}
L_{100}	58	13	2	0	4	4	3	0	76.5	
L_{100}^{2step}	47	8	1	0	2	4	2	0	79.6	
L_{200}	98	84	6	3	25	20	13	3		82.5
L_{200}^{2step}	100	92	8	3	31	23	15	3		83.9
AutoClust	100	100	90	39	65	31	24	11	83.5	85.8
$AutoClust_R$	100	100	68	12	34	11	6	2	85.2	86.5
$PenReg_{S_1+S_2}$	100	100	100	96	99	67	62	46	79.6	84.4
$PenReg_{\mathbf{S}}$	100	100	97	42	95	78	61	20	82.1	85.4
$RankCor_{S_1+S_2}$	36	0	0	0	0	0	0	0	79.6	80.2
Extreme	98	58	0	0	3	0	0	0	84.1	84.8
$Agarwal_{S_1+S_2\geq 1}$	100	100	98	59	81	35	32	11	83.4	85.9

Simulations

Setting 2: Distribution of X not elliptical symmetric

Generate data:

- $Y \sim Bin(N = 5000, p = 0.3)$
- $\boldsymbol{X}_i \sim MVN(y_i \Sigma^{\boldsymbol{X}} \boldsymbol{\beta}_0, \Sigma^{\boldsymbol{X}}) \Rightarrow \log(\lfloor \exp(\boldsymbol{X}) \rceil + 1)$
- $\boldsymbol{S}_i^0 \sim MVN(0, \Sigma^{\boldsymbol{S}}) + \boldsymbol{\mu}^{\boldsymbol{S}} + y_i \Delta_0^{\boldsymbol{S}}$
- $\boldsymbol{S} = log\{\lfloor \exp(\boldsymbol{S}^0) \rceil + 1\}$

Simulations

Setting 3: $S \not\perp\!\!\!\perp X \mid Y$

Generate data:

- $Y \sim Bin(N = 5000, p = 0.3)$
- $\boldsymbol{X}_i \sim MVN(y_i \Sigma^{\boldsymbol{X}} \boldsymbol{\beta}_0, \Sigma^{\boldsymbol{X}})$
- $\boldsymbol{S}_i^0 \sim MVN(0, \Sigma^{\boldsymbol{S}}) + \boldsymbol{\mu}^{\boldsymbol{S}} + y_i \Delta_0^{\boldsymbol{S}} + [X_1, X_3]$
- $\boldsymbol{S} = log\{\lfloor \exp(\boldsymbol{S}^0) \rceil + 1\}$

TABLE 2 Model Sizes and AUCs for final algorithm training with (a) $p = 100$ or (b) $p = 200$ labeled samples (AUC_{100}, AUC_{200}) for the strong signal when $\mathbf{X}$ is not elliptically symmetric based on (i) supervised training with all features and 100 or 200 labeled samples (L_{100}, L_{200}) or with features selected via the two-step approach (L_{100}^{2step}, L_{200}^{2step}) (ii) the automated feature selection method with and without resampling (AutoClust, $AutoClust_R$), (iii) a penalized regression with $S_1 + S_2$ or $\mathbf{S}$ as the outcome ($PenReg_{S_1+S_2}$, $PenReg_{\mathbf{S}}$), (iii) the rank correlation method based on $S_1 + S_2$ ($RCor_{S_1+S_2}$) (iv) the extreme sampling method (Extreme) and (iv) the method of Agarwal et al. (2016) based on $S_1 + S_2 \geq 1$ as the silver standard label ($Agarwal_{S_1+S_2\geq 1}$).

(a) $p = 50$

Method	Model Size	AUC_{100}	AUC_{200}
L_{100}	1	77.8	
L_{100}^{2step}	1.1	79.8	
L_{200}	4.7		82.8
L_{200}^{2step}	5.4		83.4
AutoClust	10.1	82.8	85
$AutoClust_R$	4.3	84	85.4
$PenReg_{S_1+S_2}$	26	80.3	84.3
$PenReg_{\mathbf{S}}$	19.6	81.3	84.6
$RankCor_{S_1+S_2}$	0.3	79.3	79.9
Extreme	1.3	82.5	83.3
$Agarwal_{S_1+S_2\geq 1}$	11	82.6	85

(b) $p = 100$

Method	Model Size	AUC_{100}	AUC_{200}
L_{100}	0.5	75.6	
L_{100}^{2step}	0.2	78.7	
L_{200}	1.9		80
L_{200}^{2step}	2.4		82.6
AutoClust	14.9	82	84.7
$AutoClust_R$	4.6	83.9	85.3
$PenReg_{S_1+S_2}$	50.3	77.8	82.6
$PenReg_{\mathbf{S}}$	22	80.9	84.3
$RankCor_{S_1+S_2}$	0.3	79.3	79.8
Extreme	0.8	81	81.6
$Agarwal_{S_1+S_2\geq 1}$	15.1	82	84.8

TABLE 3 Model Sizes and AUCs for final algorithm training with 100 or 200 labeled samples (AUC_{100}, AUC_{200}) for the strong signal when $\mathbf{S} \not\perp\!\!\!\perp \mathbf{X} \mid Y$ based on (i) supervised training with all features and 100 or 200 labeled samples (L_{100}, L_{200}) or with features selected via the two-step approach (L_{100}^{2step}, L_{200}^{2step}) (ii) the automated feature selection method with and without resampling (AutoClust, $AutoClust_R$), (iii) a penalized regression with $S_1 + S_2$ or $\mathbf{S}$ as the outcome ($PenReg_{S_1+S_2}$, $PenReg_{\mathbf{S}}$), (iii) the rank correlation method based on $S_1 + S_2$ ($RCor_{S_1+S_2}$) (iv) the extreme sampling method (Extreme) and (iv) the method of Agarwal et al. (2016) based on $S_1 + S_2 \geq 1$ as the silver standard label ($Agarwal_{S_1+S_2\geq 1}$).

(a) $p = 50$

Method	Model Size	AUC_{100}	AUC_{200}
L_{100}	3.3	81.2	
L_{100}^{2step}	3.6	81.6	
L_{200}	6.8		85.8
L_{200}^{2step}	7.9		84.5
AutoClust	5.7	85.3	86.5
$AutoClust_R$	4	85.6	86.6
$PenReg_{S_1+S_2}$	30.6	83	86.1
$PenReg_{\mathbf{S}}$	19	83.9	86.1
$RankCor_{S_1+S_2}$	2.6	83.6	84.5
Extreme	2.3	82.1	83
$Agarwal_{S_1+S_2\geq 1}$	9.1	84.9	86.4

(b) $p = 100$

Method	Model Size	AUC_{100}	AUC_{200}
L_{100}	1.1	76.9	
L_{100}^{2step}	0.8	81.2	
L_{200}	5.1		84.3
L_{200}^{2step}	5.7		84.1
AutoClust	7.4	85.1	86.4
$AutoClust_R$	4.2	85.6	86.6
$PenReg_{S_1+S_2}$	53.1	80.6	85.2
$PenReg_{\mathbf{S}}$	21.2	83.8	86
$RankCor_{S_1+S_2}$	2.6	83.6	84.4
Extreme	2	81.3	82.1
$Agarwal_{S_1+S_2\geq 1}$	12.6	84.6	86.1

Data Analysis

Rheumatoid arthritis

- $46{,}111$ potential RA subjects
- 435 subjects are labeled via chart review
- S_1 : number of RA icd-9 codes; S_2 : counts of NLP mentions of RA
- $\boldsymbol{X}$: $p = 77$ NLP features
- Randomly sample $N = 5000$ unlabelled data for feature selection
- $n = 100{,}200$ training data; $435 - n$ validation set.

TABLE 4 (a) Number of features included in the algorithm training as well as coefficient estimates and AUC of the resulting algorithms trained with (b) $n = 100$ and (c) $n = 200$ labels when fitting the full model (L) as well as when only including features selected based on the AutoClust, AutoClust$_R$, RankCor$_{S_1+S_2}$, Extreme, and Agarwal$_{S_1+S_2\geq 1}$ methods for the EMR-based study of RA.

	(a) Number of features selected				
	AutoClust	AutoClust$_R$	RankCor$_{S_1+S_2}$	Extreme	Agarwal$_{S_1+S_2\geq 1}$
	33.00	26.80	69.20	23.80	45.10
(b) Coefficient and AUC estimates with $n = 100$ labels					
	L_{100}	AutoClust	AutoClust$_R$	RankCor$_{S_1+S_2}$	Extreme	Agarwal
RA$_{ICD}$	0.35	0.66	0.7	0.4	0.73	0.58
RA$_{NLP}$	0.27	0.68	0.74	0.32	0.85	0.49
Methotraxate$_{NLP}$	0.08	0.14	0.14	0.09	0.13	0.11
AM Stiffness$_{NLP}$	0.04	0.12	0.13	0.05	0.15	
Echography$_{NLP}$					−0.21	
AUC	91.1	92.7	92.8	91.2	93.4	91.7
(c) Coefficient and AUC estimates with $n = 200$ labels						
	L_{200}	AutoClust	AutoClust$_R$	RankCor$_{S_1+S_2}$	Extreme	Agarwal
RA$_{ICD}$	0.7	0.85	0.87	0.72	0.87	0.79
RA$_{NLP}$	0.72	1.26	1.35	0.79	1.52	1.07
Methotraxate$_{NLP}$	0.13	0.15	0.15	0.14	0.14	0.15
MRI$_{NLP}$	−0.05			−0.09		
AM Stiffness$_{NLP}$	0.1	0.28	0.29	0.12	0.29	0.17
Physiotherapy$_{NLP}$	−0.07					−0.15
Echography$_{NLP}$	−0.08				−0.53	−0.26
Note Count$_{NLP}$		−0.37	−0.41		−0.36	
Antinuclear Antibodies$_{NLP}$		−0.11	−0.13			
Redness$_{NLP}$			−0.16	−0.06		
Intravenous Infusion$_{NLP}$					−0.15	
AUC	92.8	94.3	94.5	92.9	94.8	94.2

Discussion

- AutoClust Vs AutoClust$_R$
  - AUC is similar
  - AutoClust$_R$ tends to select smaller number of variables
- Combine label and unlabeled data
  - $$\sum_{i=1}^{N} l(\boldsymbol{\theta}^T \overrightarrow{\boldsymbol{X}}_i \mid \hat{\pi}_S) + \sum_{j=N+1}^{N+n} l(\boldsymbol{\theta}^T \overrightarrow{\boldsymbol{X}}_i \mid Y_i)$$