P8160 Group Project Presentation

Optimization and Bootstrap

Xinlei Chen, Guojing Wu, Yujing Yao

Department of Biostatistics, Columbia University

April 19, 2019

Introduction

Introduction

Introduction

Group project 2: Optimization algorithms on a breast cancer diagnosis dataset

- **Aim:** Build a predictive model based on logistic regression to faciliate cancer diagnosis, and we compared methods including Newton Raphson, Gradient Decent with general logistic regression and Pathwise Coordinate Descent with regularized logistic regression

Introduction

Group project 3: Bootstrapping on developing classification model

- **Aim:** Build a predictive model based on logistic regression to faciliate down syndrome diagnosis, and compared methods including and Pathwise Coordinate Descent with regularized logistic regression and smoothed bootstrap estimation.

Project: Optimization

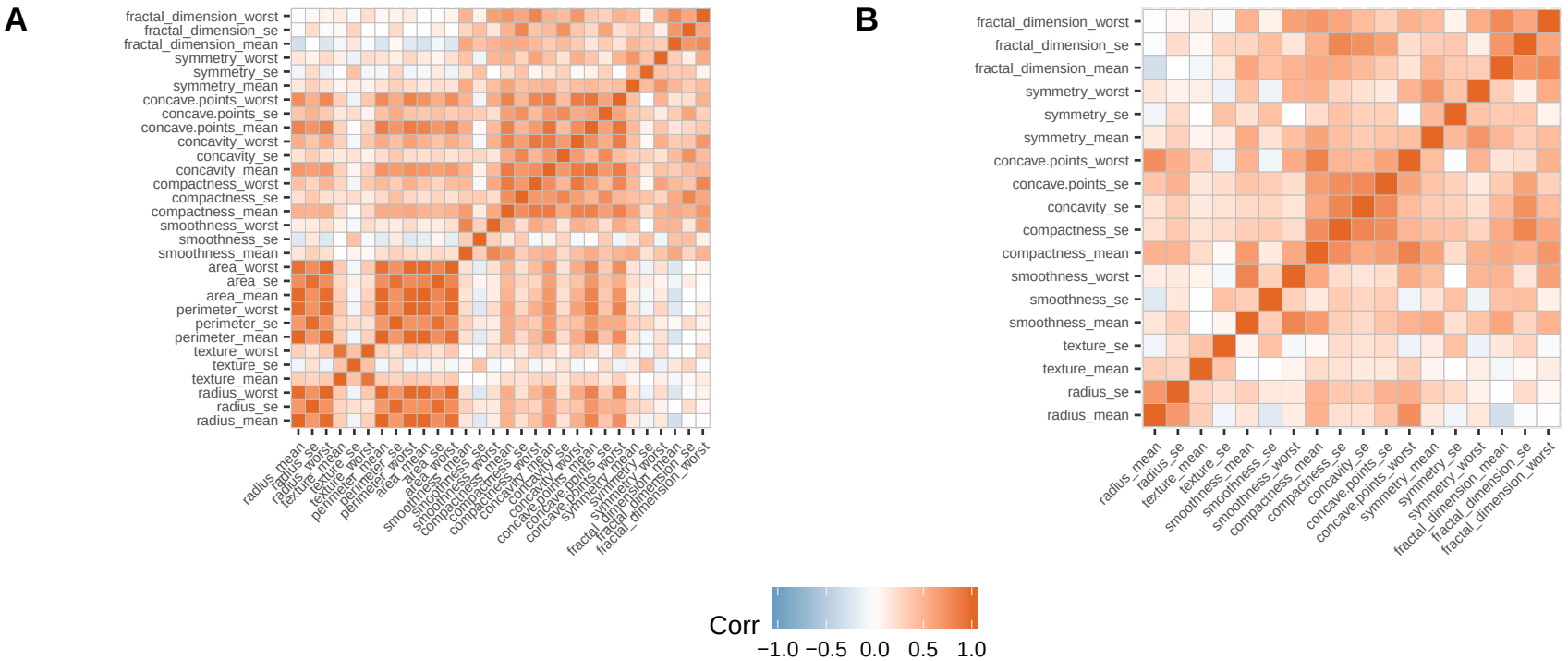
Background

Breast Cancer Data

The data breast-cancer.csv have 569 row and 33 columns. The first two columns inclues covariate "ID" which lables individual breast tissue images and covariate "Diagnonsis" which indentifies if the image is coming from cancer tissue or benign cases. There are 357 benign and 212 malignant cases. The other 30 columns correspond to mean, standard deviation and the largest values (points on the tails) of the distributions of 10 features (radiusm texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension) computed for the cellnuclei.

Multicollinearity Plot of the Dataset

- **Variable Selection:** Reduce multicollinearity based on both correlation coefficient and eigenvalue of correlation matrix



Method

Logistic Model

Logistic Regression:

y: the vector of n response random variable

X: the $n \times p$ design matrix (X_i denote the ith row)

β: the $p \times 1$ coefficient

- Object: maximize log-likelihood function

$$max \sum_{i=1}^{n} \{y_i(X_i\beta) - \log(1 + \exp(X_i\beta))\}$$

Newton Raphson

- The gradient:

$$\nabla l(\beta) = X^T(y - p)$$

where $p = \frac{\exp(X\beta)}{1+\exp(X\beta)}$

- The Hessian:

$$\nabla^2 l(\beta) = -X^T W X$$

where $W = diag(p_i(1 - p_i)), i = 1, \cdots, n$. The Hessian is negative definite.

Newton Raphson

Update coefficients: step-halving

$$\beta_{i+1}(\gamma) = \beta_i - \gamma[\nabla^2 l(\beta_i)]^{-1}\nabla l(\beta_i)$$

- Set $\gamma = 1$
- If $f(\theta_{i+1}(1)) \geq f(\theta_i)$, then set $\theta_{i+1} = \theta_{i+1}(1)$
- If $f(\theta_{i+1}(1)) \leq f(\theta_i)$, search for a value $\gamma \in (0, 1)$ for which $f(\theta_{i+1}(\gamma)) \geq f(\theta_i)$, set $\theta_{i+1} = \theta_{i+1}(\gamma)$

Gradient Descent:

$$\beta_{i+1} = \beta_i + H_i \nabla f(\beta_i)$$

where $H_i = (X^T X)^{-1}$ for every i.

Logistic-LASSO Model with Pathwise Coordinate Descent

- Object: maximize the penalized log likelihood:

$$\max_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \{y_i (X_i \beta) - \log(1 + \exp(X_i \beta))\} - \lambda \sum_{j=0}^{p} |\beta_j|$$

  for some $\lambda \geq 0$

- Coordinate-wise descent with weighted update:

$$\tilde{\beta}_j^{lasso}(\lambda) \leftarrow \frac{S(\sum_{i=1}^{n} \omega_i x_{i,j} (y_i - \tilde{y}_i^{(-j)}), \lambda)}{\sum_{i=1}^{n} \omega_i x_{i,j}^2}$$

where $\tilde{y}_i^{(-j)} = \sum_{k \neq j} x_{i,k} \tilde{\beta}_k$ and $S(\hat{\beta}, \lambda) = sign(\hat{\beta})(|\hat{\beta}| - \lambda)_+$

Result

Estimation Path and Cross Validation for LASSO

Figure 2: A path of solutions with a sequence of descendi

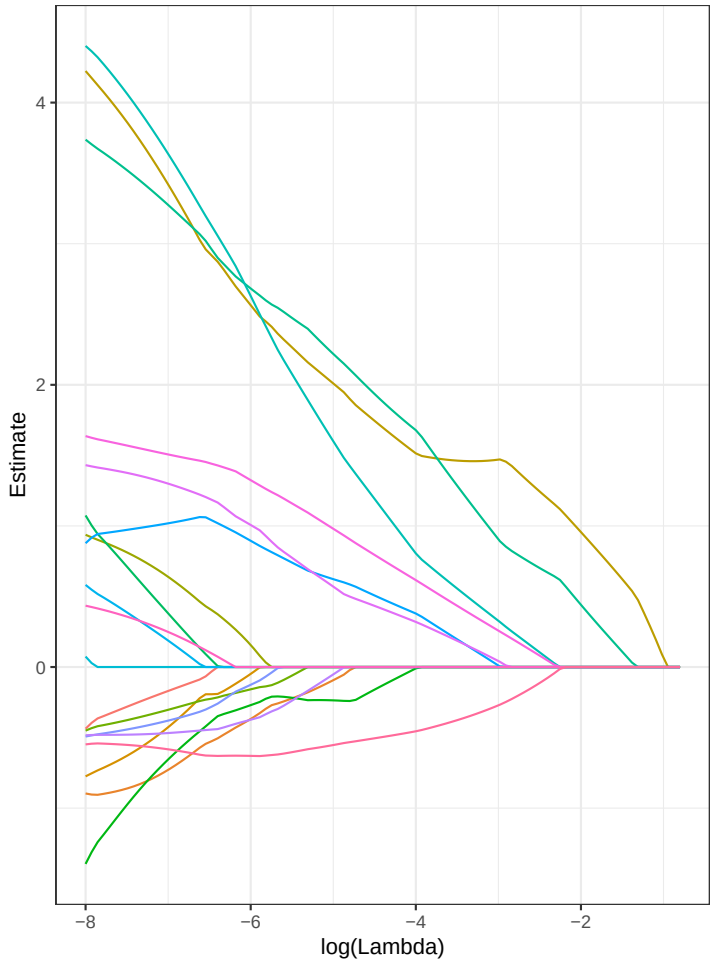
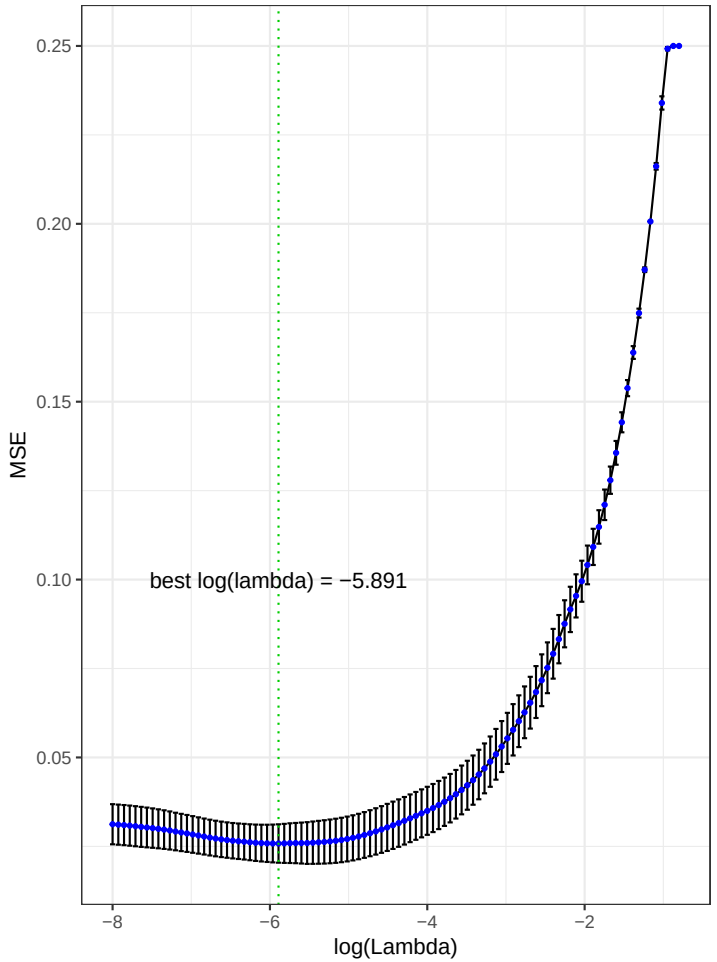


Figure 3: Lasso regression by 5 fold cross validation



Model Comparison-Prediction Performance

Table 1: The comparison of performance for estimation algorithms and models

	GLM package	Newton Raphson	Gradient Decent	Logistic Lasso	Lasso package
iteration times	NA	12	1001	100	NA
MSE	0.02	0.02	0.02	0.02	0.02

[a] Dataset: Breast Cancer Diagnosis

Model Comparison-Coefficient

Table 2: The comparison of performance for estimation algorithms and models

	GLM package	Newton Raphson	Gradient Decent	Logistic Lasso	Lasso package
radius_mean	4.43	4.43	3.18	2.63	2.71
texture_mean	1.89	1.89	1.34	1.29	1.37
smoothness_mean	0.78	0.78	0.47	0.00	0.00
compactness_mean	-1.14	-1.14	-0.59	0.00	0.00
symmetry_mean	-0.63	-0.63	-0.44	-0.10	-0.14
fractal_dimension_mean	-0.66	-0.66	-0.72	-0.14	-0.21
radius_se	5.13	5.13	3.28	2.50	2.58
texture_se	0.59	0.59	0.46	0.00	0.00
smoothness_se	1.10	1.10	0.77	0.00	0.00
compactness_se	-0.80	-0.80	-0.68	-0.33	-0.38
concavity_se	1.24	1.24	0.88	0.08	0.19
concave.points_se	-1.11	-1.11	-0.80	0.00	0.00
symmetry_se	-0.53	-0.53	-0.39	-0.36	-0.42
fractal_dimension_se	-2.73	-2.73	-1.55	-0.25	-0.31
smoothness_worst	0.31	0.31	0.31	0.86	0.92

Conclusion

Conclusion and Discussion

- The results of our methods are compared to the the same parameter estimation as R's built-in packages
- Newton-Raphson has the convincing estimation and it converged quickly
- Gradient decent method showed similar estimation as Newton-Raphson method but it was less efficient
- For Pathwise Coordinate descent with LASSO logistic, according to the result of 5 fold cross validation and estimation result, the λ with the lowest MSE and it shrunk six parameters to zero, which is comparable to the result by R's built-in packages.

Project: Bootstrap

Background

Down Syndrome Data

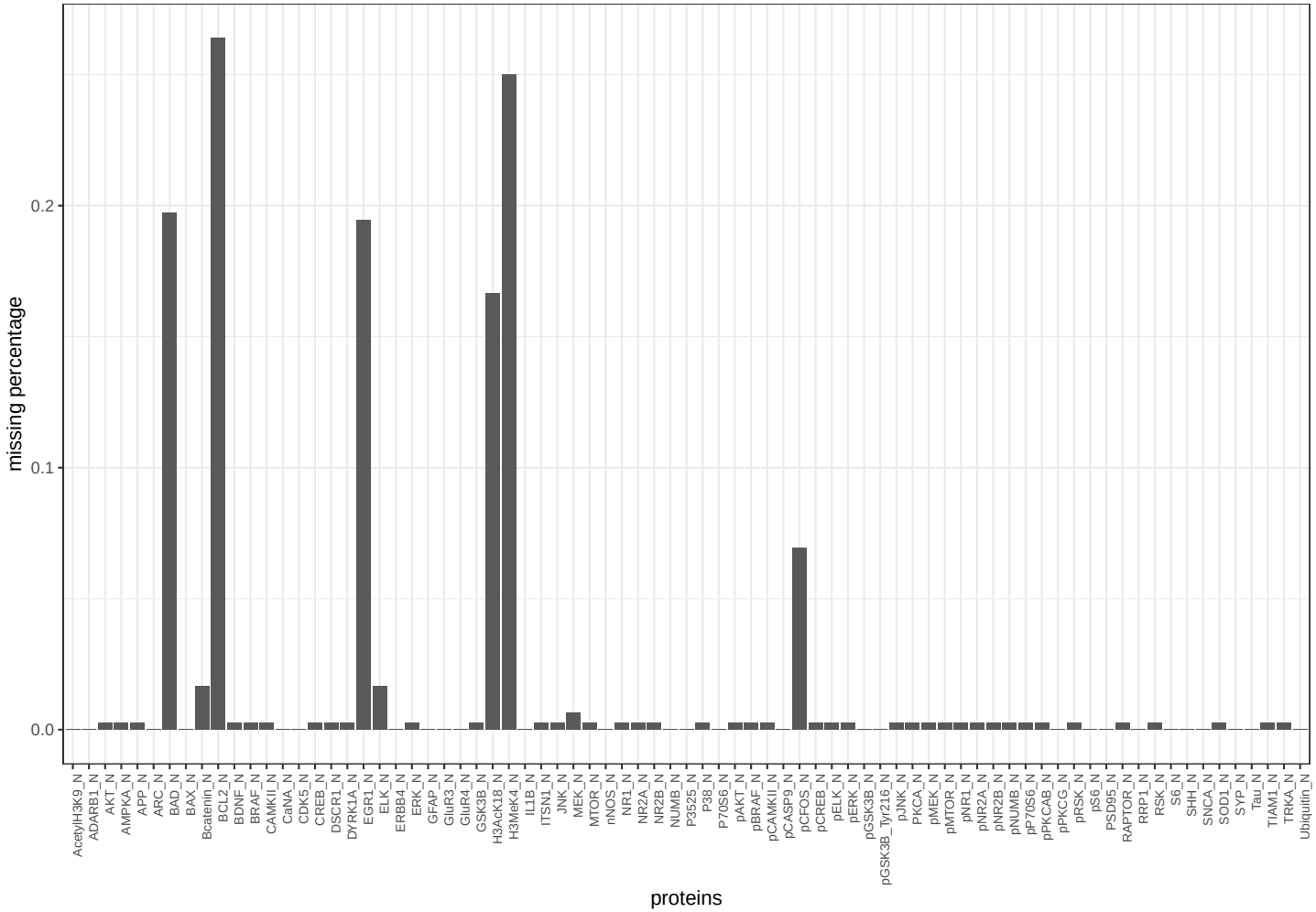
The data Down.csv consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex. It has 1080 rows and 79 columns. The first column MouseID identifies individual mice; The column 2-78 are values of expression levels of 77 proteins. Column 79 indicates whether the mouse is a control or has Down syndrome. The goal is to develop classification model based on the proteins expression levels.

Missingness

- **Variable Selection:** Delete variables with high missing rate ($\geq 20\%$)

Also due to the intrinsic correlation between individual proteins, it's impossible to apply normal regression methods to this dataset because of sigularity propblem. Instead, we choose regularized methods, LASSO, to be more specific.

Missingness



Methods

Pathwise Coordinate Descent with Regularized Logistic Regression

- Object: maximize the penalized log likelihood:

$$\max_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} \{ y_i (X_i \beta) - \log(1 + \exp(X_i \beta)) \} - \lambda \sum_{j=0}^{p} |\beta_j|$$

  for some $\lambda \geq 0$

- Coordinate-wise descent with weighted update:

$$\tilde{\beta}_j^{lasso}(\lambda) \leftarrow \frac{S(\sum_{i=1}^{n} \omega_i x_{i,j} (y_i - \tilde{y}_i^{(-j)}), \lambda)}{\sum_{i=1}^{n} \omega_i x_{i,j}^2}$$

where $\tilde{y}_i^{(-j)} = \sum_{k \neq j} x_{i,k} \tilde{\beta}_k$ and $S(\hat{\beta}, \lambda) = sign(\hat{\beta})(|\hat{\beta}| - \lambda)_+$

Smoothed Bootstrap Estimation and Inference

- First we need to prepare a couple of candidate models
- for each bootstrap in bootstrap with B times, select the best model and get estimates for the coefficient denoted as $t(y^*)$
- smooth $\hat{\mu} = t(y)$ by averaging over the bootstrap replications, defining

$$\tilde{\mu} = s(y) = \frac{1}{B} \sum_{i=1}^{B} t(y^*)$$

Smoothed Bootstrap Estimation and Inference

And in addition to the percentile confidence interval, the nonparametric delta-method estimate of standard deviation for $s(y)$ in the nonideal case is:

$$\tilde{sd}_B = [\sum_{i=1}^{n} \hat{cov}_j^2]^{1/2}$$

where

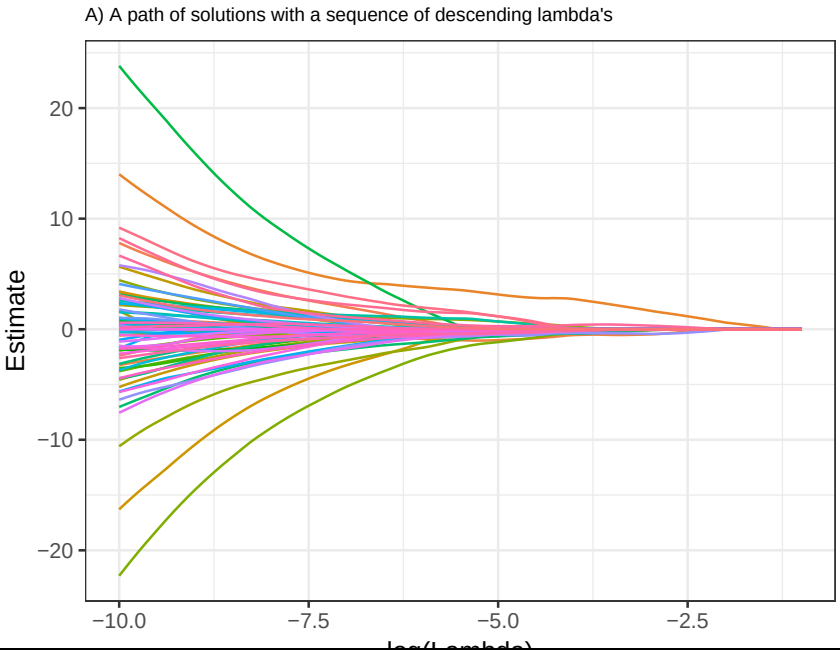
$$\hat{cov}_j = \sum_{i=1}^{B} (Y_{ij}^* - Y_{.j}^*)(t_i^* - t_.^*)/B$$

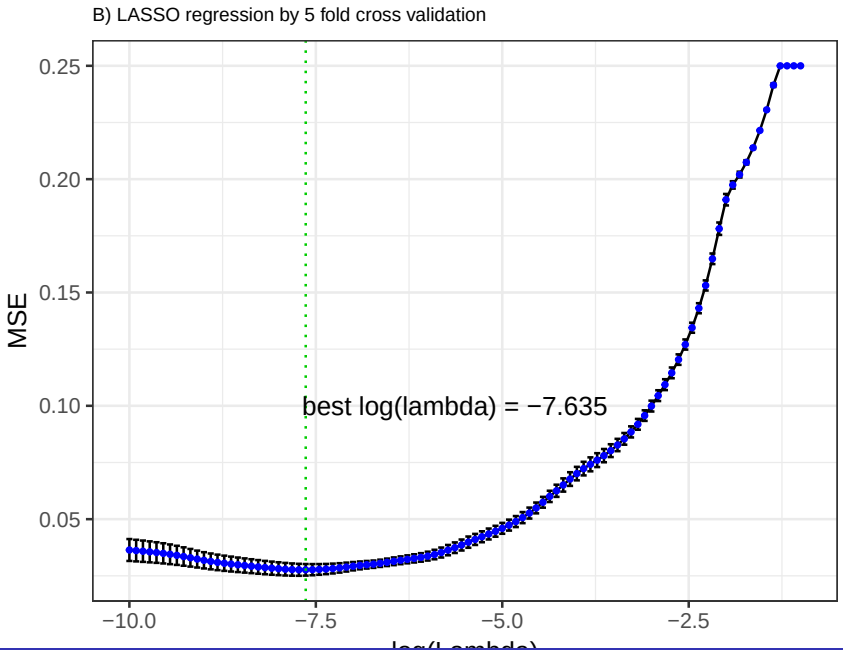
with $Y_{.j}^* = \sum_{i=1}^{B} Y_{ij}^*/B$ and $t_.^* = \sum_{i=1}^{B} t_i^*/B = s(y)$.

Result

Pathwise Coordinate Descent Logistic-LASSO

The left-hand side plot shows us that as the λ increases, all the variable estimates of parameters shrink accordingly since we penalize all the parameters. When $\lambda = 0$, the result is the same as least square method and when λ is too large, all the estimates of parameters shrink to 0. The right-hand side plot shows us the cross validation result for choosing the best λ.





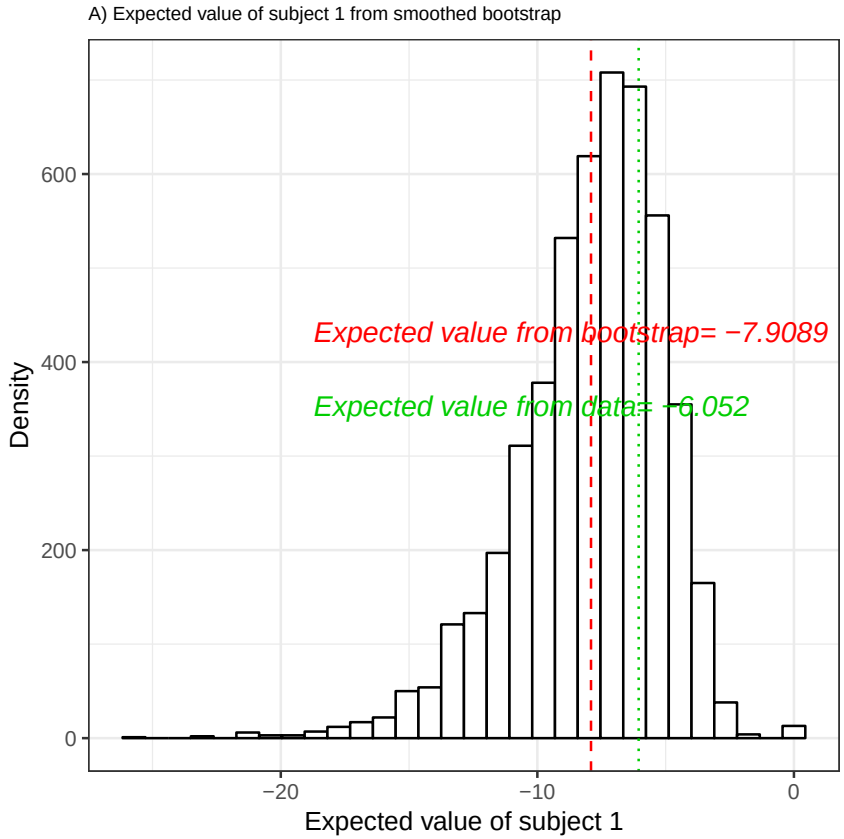
Model Selection Based on Smooth Bootstrap Estimation for Logistic-LASSO

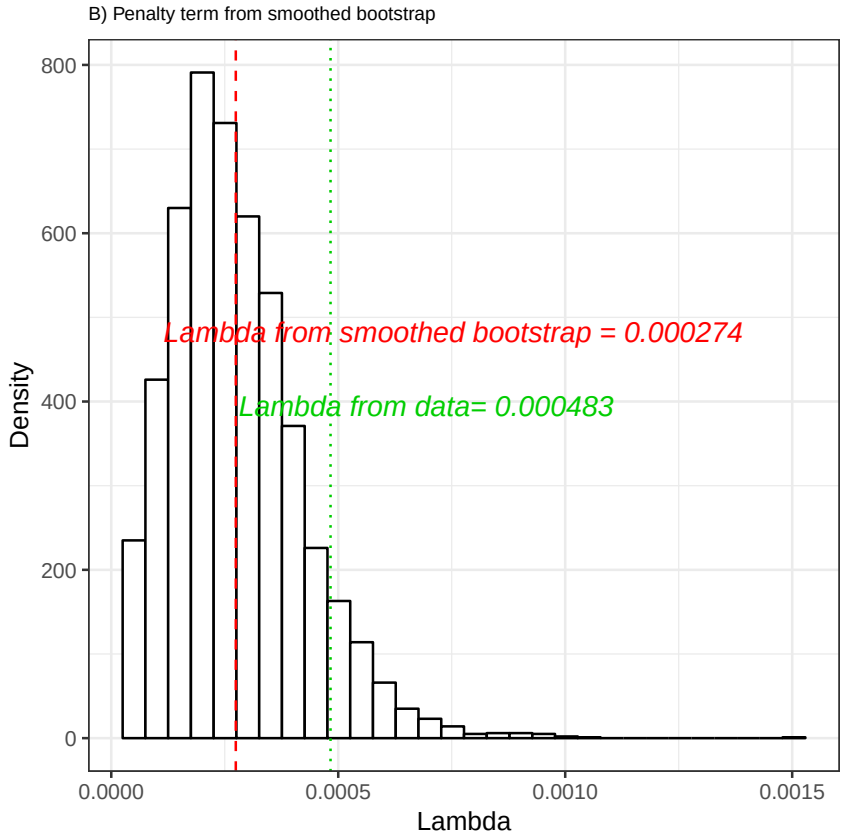
algorithm:

- bootstrap data from the original dataset
- do cross validation and select the best λ_i^* for each repetition
- calculate average $\lambda^* = \frac{1}{B}\sum_{i=1}^{B}\lambda_i^*$

We can see the discrepancy between results of PCD-LASSO and smooth bootstrap estimation for Logistic LASSO both in prediction and finding the best λ, the results of PCD-LASSO is deviated from the center of empirical distribution.

Lambda Selection Based on Smooth Bootstrap Estimation for Logistic LASSO





Cross Validation for Model Prediction Comparison

We used 10 fold cross-validation to compare two different models, one is with λ selected from data, the other is selected from the SBE. Table 1 shows us that while the Cross Validation MSE are similar between these two methods, smooth bootstrap estimation provides a more accurate classification result.

Table 3: The comparison of performance for two models

	Misclassification rate	Mean squred error
Penalty chosen by data	0.0353	0.0229
Penalty selected from smoothed bootstrap	0.0335	0.0216

[a] Dataset: Proteins expression levels of Down syndrome

Significant Random Variable Selection from Smooth Bootstrap Estimation

Table 2 & 3 provide the full results of smooth bootstrap estimation for logistic LASSO. Our identification criterions here are:

1. the chosen probability greater than 96%
2. smooth bootstrap estimation confidence interval excludes zero.

Based on that, we got 27 proteins that meets these two criterions (Table 4).

Significant Proteins

Table 4: Significant Proteins

	origin	prob	coef	sd	lower	upper	lower.new	upper.new
ITSN1_N	7.9040	1.00	9.6725	2.1862	4.4201	16.9149	5.3875	13.9575
pELK_N	-2.1788	0.99	-2.7419	1.1203	-6.1524	-0.4194	-4.9377	-0.5461
pNR1_N	-2.4536	0.96	-2.6193	1.2315	-5.9502	0.0000	-5.0330	-0.2056
pRSK_N	-2.4102	1.00	-2.8974	0.7656	-5.2917	-1.1120	-4.3980	-1.3968
AKT_N	2.7674	1.00	3.3722	0.8799	1.2074	6.0927	1.6476	5.0968
BRAF_N	-4.8560	1.00	-5.6899	1.7241	-10.9610	-1.8681	-9.0691	-2.3107
CAMKII_N	-1.5901	0.99	-2.3111	0.8889	-4.9441	-0.3901	-4.0533	-0.5689
CREB_N	-1.2469	0.98	-1.3510	0.5539	-2.9474	-0.0163	-2.4366	-0.2654
ELK_N	-3.6872	1.00	-4.6751	1.0545	-8.0746	-2.1847	-6.7419	-2.6083
ERK_N	-7.4243	1.00	-8.7471	1.6856	-14.5871	-4.7828	-12.0509	-5.4433
MEK_N	1.3308	0.98	1.6328	0.7152	0.0194	3.6599	0.2310	3.0346
TRKA_N	3.7756	1.00	5.5845	2.1337	2.0454	11.9326	1.4024	9.7666
APP_N	5.3514	1.00	7.8719	1.4402	5.0187	13.0828	5.0491	10.6947
MTOR_N	-2.3190	0.99	-2.8751	0.9748	-5.7528	-0.7263	-4.7857	-0.9645
DSCR1_N	1.2781	0.98	1.5514	0.6235	0.0022	3.3412	0.3293	2.7735

Non-significant Proteins

Table 5: Non-significant Proteins

	origin	prob	coef	sd	lower	upper	lower.new	upper.new
Intercept	0.0949	1.00	0.3143	0.3546	-0.4132	1.2144	-0.3807	1.0093
DYRK1A_N	0.0000	0.04	0.0310	0.1390	0.0000	0.2384	-0.2414	0.3034
BDNF_N	1.0655	0.86	1.0811	0.7540	0.0000	3.1864	-0.3967	2.5589
NR1_N	-0.9385	0.90	-2.0857	1.2323	-5.7961	0.0000	-4.5010	0.3296
NR2A_N	0.3742	0.54	0.4882	0.5696	0.0000	2.5152	-0.6282	1.6046
pAKT_N	0.6050	0.82	0.8438	0.6632	0.0000	2.7745	-0.4561	2.1437
pBRAF_N	0.0000	0.62	0.1922	0.4213	-0.7271	1.5147	-0.6335	1.0179
pCAMKII_N	-0.6756	0.93	-1.0979	0.6266	-2.7897	0.0000	-2.3260	0.1302
pCREB_N	1.2113	0.95	1.3620	0.6984	0.0000	3.3009	-0.0069	2.7309
pERK_N	-0.2493	0.54	-0.5806	0.5946	-2.7898	0.0000	-1.7460	0.5848
pJNK_N	-0.0067	0.46	-0.2080	0.3361	-1.4612	0.3040	-0.8668	0.4508
PKCA_N	1.1962	0.82	1.2604	0.9599	0.0000	4.0332	-0.6210	3.1418
pMEK_N	0.0000	0.54	-0.0939	0.5630	-1.8812	1.2504	-1.1974	1.0096
pNR2A_N	0.0000	0.28	-0.0891	0.2675	-1.1664	0.5175	-0.6134	0.4352
pNR2B_N	1.5568	0.90	1.8496	1.0888	0.0000	5.1242	-0.2844	3.9836

Conclusion