实例：新加坡眼科数据

---毕业论文中期汇报

高明

2018年5月

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Binomial Distribution

Suppose $y_i \sim Bin(1, p_i)$, then $\mu_i \triangleq E(y_i) = p_i$. Let $x_i^T \beta = g(\mu_i) \triangleq ln\frac{p_i}{1-p_i}$.
Hence,

$$p_i = P(y_i = 1|x_i) = \frac{1}{1 + e^{-x_i^T \beta}} \tag{10}$$

$$\begin{aligned} Q(\beta) &= \frac{1}{n}\sum_{i=1}^{n} y_i ln p_i + (1 - y_i) ln(1 - p_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i (x_i^T \beta) - ln(1 + e^{x_i^T \beta}) \end{aligned}$$

LASSO Family

LASSO:

$$\tilde{\beta} = argmin\{\frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda|\beta|_1\} \triangleq argmin\{Q(\beta) + \phi(\beta)\} \qquad (5)$$

GRPL:

we divide $\{1...m\}$ into J group $\{G_1, ..., G_J\}$. $\#\{G_j\} \triangleq p_j$. $\beta_{G_j} = (\beta_k)_{k \in G_j}$

$$\phi(\beta) = \lambda\sum_{j=1}^{J}\sqrt{p_j}|\beta_{G_j}|_2 \qquad (6)$$

SGRPL:

$$\phi(\beta) = \lambda\{(1-\alpha)\sum_{j=1}^{J}\sqrt{p_j}|\beta_{G_j}|_2 + \alpha|\beta|_1]\} \qquad (7)$$

Model Selection: CV

In K-fold cross-validation, the original sample is randomly partitioned into K equal sized subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K-1$ subsamples are used as training data.

for each $k = 1, ..., K$, fit the model with parameter λ to the other $K-1$ parts, giving $\tilde{\beta}^{-k}(\lambda)$ and compute its loss $LOSS_k(\lambda)$ in predicting the k^{th} part. This gives the cross-validation error

$$CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} LOSS_k(\lambda) \tag{11}$$

$$\lambda^* = argminCV(\lambda) \tag{12}$$

CV_dev & CV_ME

- Deviance.

$$Dev_k(\lambda) = \frac{-2}{n/k} Loglik(\tilde{\beta}^{-k}(\lambda)) \tag{20}$$

Deviance is inverse ratio to Log likelihood function, which is a measure of goodness of fit. Usually, deviance is obtained by log-likelihood ratio which contains the saturated model. However, since the principal use is in the form of the difference of the deviances of two models, this confusion in definition is unimportant. We use deviance on the left-out data with size n/k.

- Misclassification Error (ME).

$$ME = \frac{1}{n/k}\{\#_i(p_i > 0.5 \; \& \; y_i = 0) + \#_i(p_i < 0.5 \; \& \; y_i = 1)\} \tag{21}$$

ME is directly perceived through the sense. We use ME on the left-out data. ME can be treat as discrete type of Deviance.

Model selection: IC

7 AIC

$$AIC(\lambda) = -2lnl(\hat{\beta}(\lambda)) + 2\nu(\lambda) \tag{13}$$

8 BIC

$$BIC(\lambda) = -2lnl(\hat{\beta}(\lambda)) + \nu(\lambda)lnn \tag{14}$$

Here,$\nu(\lambda) = df(\lambda)$

9 EBIC

$$EBIC_\gamma(\lambda) = -2lnl(\hat{\beta}(\lambda)) + \nu(\lambda)lnn + 2\gamma\nu(\lambda)lnp \tag{15}$$

Theorem 1 *Suppose λ_0 is the true model. Under some mild conditions with $n \to \infty$, we have*

$$P\{minEBIC_\gamma(\lambda) \leq EBIC_\gamma(\lambda_0)\} \to 0 \tag{16}$$

Model Evaluation

$$p_i = P(\hat{y}_i = 1|x_i) = \frac{1}{1 + e^{-x_i^T \hat{\beta}}}$$

Where x_i is a sample in the test dataset, $\hat{y}_i$ is the prediction of y_i.
We need to select a threshold c, $0 \leq c \leq 1$. Then we forcast $\hat{y}_i = I(p_i > c)$.

$$CCR = P(y = \hat{y})$$

Receiver Operating Characteristic (ROC) curve summarizes the models performance by evaluating the tradeoffs between true positive rate (TPR, sensitivity) and false positive rate (FPR, 1-specificity), where $FPR = P(\hat{y} > c|y = 0)$ and $TPR = P(\hat{y} > c|y = 1)$.

AUC is the area under ROC curve. It is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example.(Fawcett, 2006)

$$AUC = \int TPR(c)\, \mathrm{d}FPR(c) = P(\hat{y}|_{y=1} > \hat{y}|_{y=0}) \tag{34}$$

Singapore Eye Study Database

- 3000 people' s 300 indexes
- basic information (age, height, …)
- blood data (glucose, cholesterol, …)
- eye data (myopia, blindness, sphere, …)
- eye disease (cataract, …)
- self-information (education, job, smoke, income)
- main disease(heart attack, stroke, hypertension, diabetes, …)

```
x                 3353 obs. of 314 variables
   sno : Factor w/ 3353 levels "CS30498","CS3755
   czmi : int 0 0 0 0 0 0 0 0 0 1 ...
   gender : int 2 2 2 1 1 2 1 1 2 1 ...
   age : num 63.1 75.4 69.4 57.6 62.4 ...
   agegp : int 3 4 3 2 3 1 2 2 3 1 ...
   agegp2 : int 3 4 3 2 3 1 2 2 3 1 ...
   agegp3 : int 5 8 6 4 5 2 3 4 6 2 ...
   bpsys_f : num 152 182 132 121 176 ...
   bpdia_f : num 79.5 98 63.5 72.5 104.5 ...
   bppul_f : num 80.5 82.3 58.5 58 62 ...
   pulse_press : num 72.5 84.5 68 48.5 71.5 62 5
   map : num 103.7 126.2 86.2 88.7 128.3 ...
   htcm : num 155 160 150 176 168 ...
   wtkg : num 61.9 64 58.3 73.3 62.3 67.5 90.7 8
   bmi : num 25.8 24.8 25.7 23.7 22.2 ...
   BMI_cat : int 3 2 3 2 2 3 3 3 2 4 ...
   anti_ht : int 1 0 1 0 0 0 1 1 1 0 ...
   anti_chol : int 1 0 1 0 0 0 0 0 0 0 ...
   anti_dm : int 0 0 1 0 0 0 0 0 0 0 ...
   drugs_others : int 1 0 0 0 0 1 0 1 1 0 ...
   drugs_unknown : int 0 0 0 0 0 0 0 0 0 0 ...
   dm5 : int 0 0 1 0 0 0 0 0 0 0 ...
   dm4 : int NA 0 1 0 0 0 0 0 NA 0 ...
   hypertension : int 1 1 1 0 1 0 1 1 1 0 ...
   hypertension3 : int 3 1 2 0 1 0 3 2 2 0 ...
   hyperlipidaemia1 : int 1 1 1 1 1 1 0 0 NA 0 .
   blood_data : int 0 1 1 1 1 1 1 1 0 1 ...
```

Preprosessing

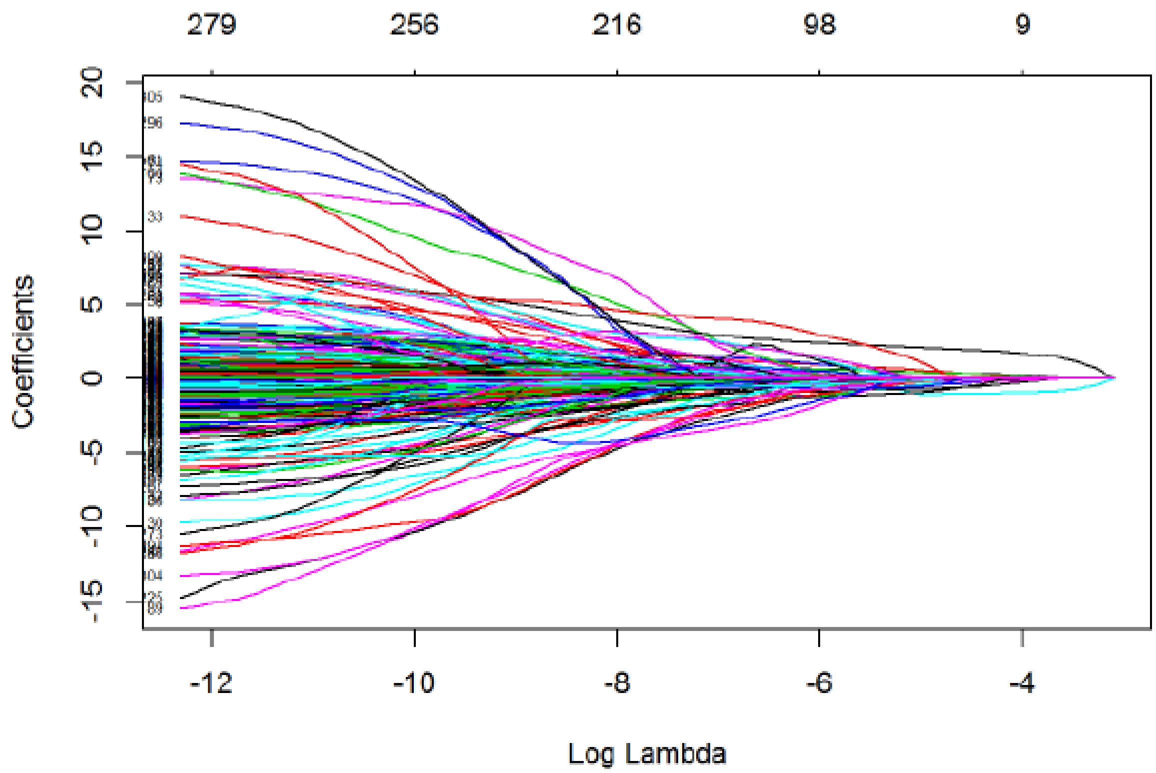
- Delete columns and rows which have too many NA values
- Assort variables less than 6 different values as factors and others as continuous predictors
- fill in the missing values with their mode and median separately
- we only focus on heart attack (variable ″mi″) as the output. A binary factor: 0/1, 1:person do not suffer from heart attack
- One Hot Encoding to turn all the factors into dummy variables
- Divide the whole dataset randomly into training and test part

Dimension

- Dimension of training input:(2949, 339)
- dimension of test input:(327, 339)
- length of training output:(2949)
- length of test output:(327)

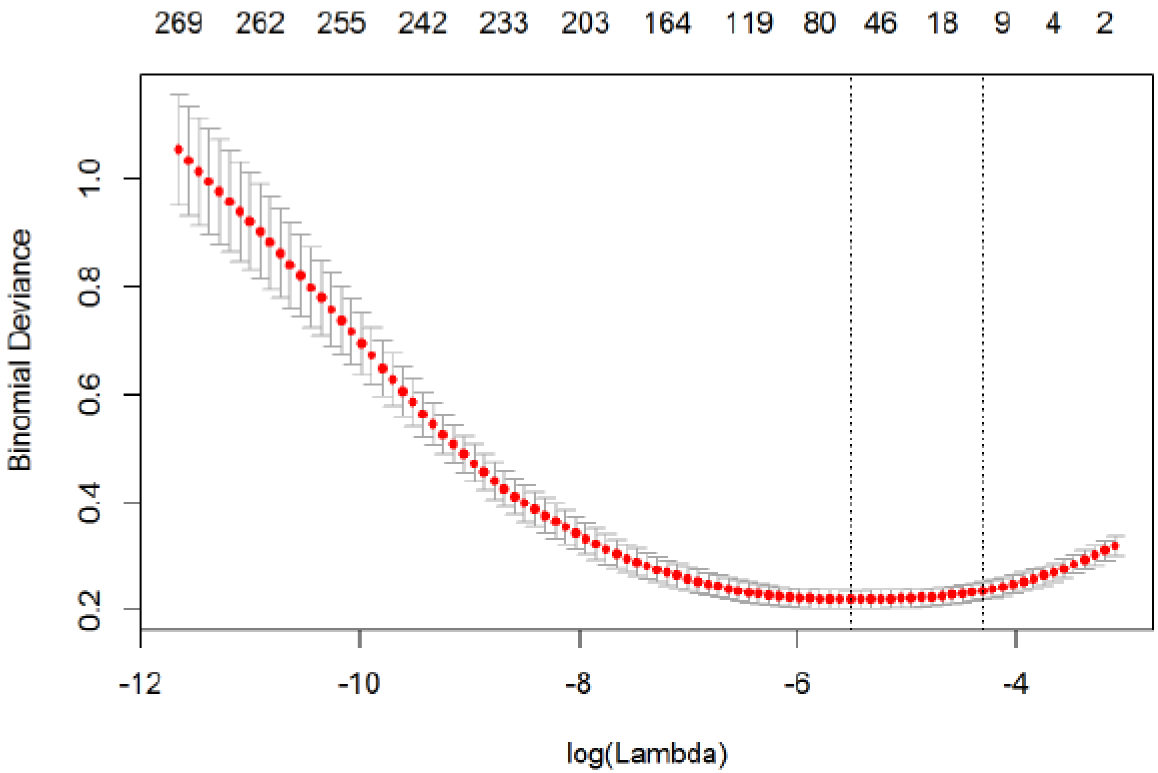
	czmi1	gender2	age	agegp2	agegp3	agegp4	agegp22	agegp23	agegp24	agegp25	agegp3	bpsys_f
887	1	1	58.55441478	1	0	0	1	0	0	0	4	99.5
1248	0	0	75.71800137	0	0	1	0	0	1	0	8	106
1918	0	0	54.00136893	1	0	0	1	0	0	0	3	107
3047	0	1	73.10335387	0	0	1	0	0	1	0	7	139.5
673	0	1	50.41204654	1	0	0	1	0	0	0	3	115.5
3013	0	0	78.65023956	0	0	1	0	0	1	0	8	180
3166	0	1	47.75633128	0	0	0	0	0	0	0	2	136.5
2211	1	0	66.69130732	0	1	0	0	1	0	0	6	136
2103	0	0	53.60711841	1	0	0	1	0	0	0	3	142
208	1	1	49.10882957	0	0	0	0	0	0	0	2	105
686	0	1	78.74606434	0	0	1	0	0	1	0	8	184
588	1	0	47.13483915	0	0	0	0	0	0	0	2	108
2296	1	1	58.88843258	1	0	0	1	0	0	0	4	135
1283	0	0	63.88227242	0	1	0	0	1	0	0	5	153

LASSO



279
256
216
98
9
20
15
10
5
0
-5
-10
-15
Coefficients
-12
-10
-8
-6
-4
Log Lambda

CV: dev

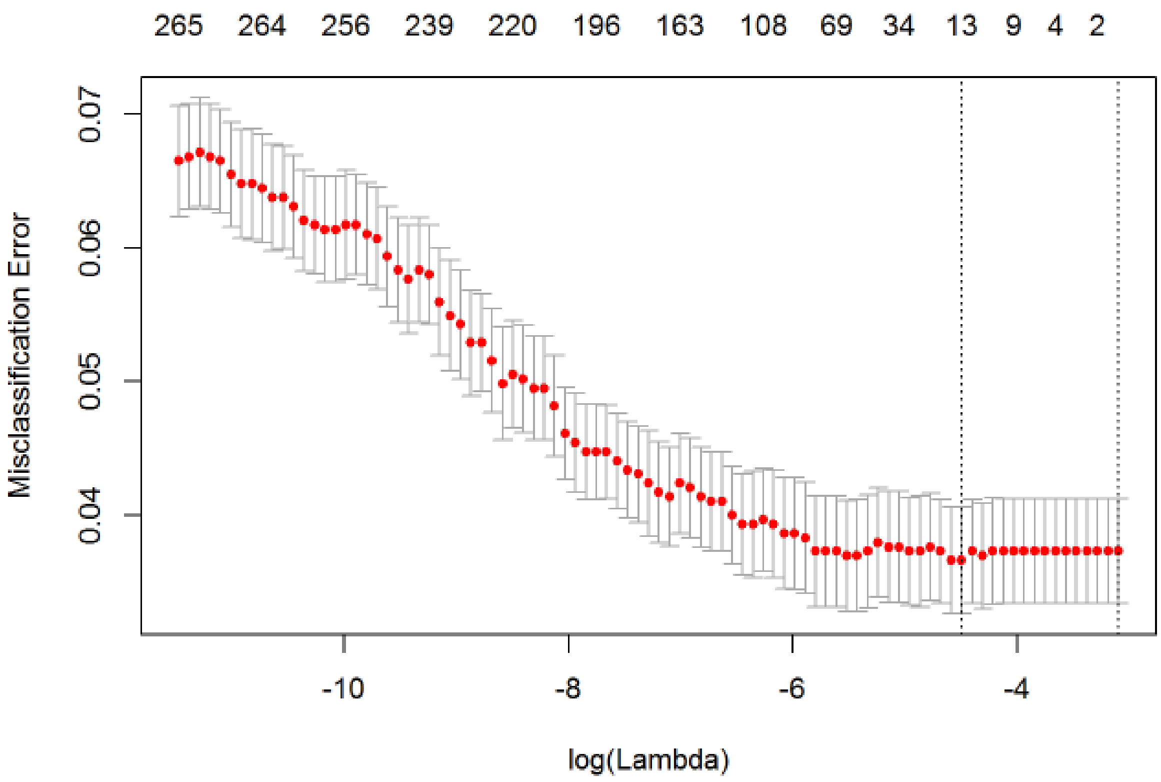


269 262 255 242 233 203 164 119 80 46 18 9 4 2
Binomial Deviance
0.2
0.4
0.6
0.8
1.0
-12
-10
-8
-6
-4
log(Lambda)

CV: dev

- The best lambda is 0.0040135.
- There are 63 no-zero variables.
- Correct classification rate of training data: 0.9677857
- Area under curve of training data: 0.9367191
- Correct classification rate of test data: 0.9602446
- Area under curve of test data: 0.8455043

CV：ME



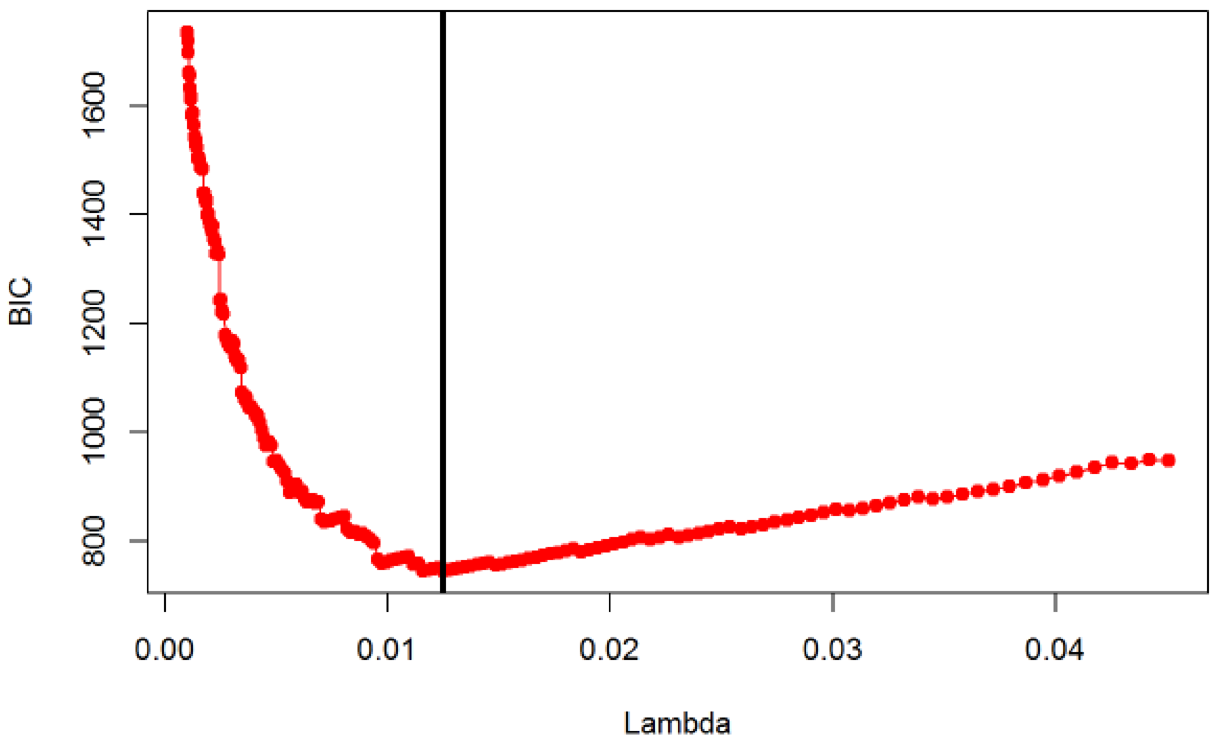
265 264 256 239 220 196 163 108 69 34 13 9 4 2
Misclassification Error
0.07
0.06
0.05
0.04
-10
-8
-6
-4
log(Lambda)

CV : ME

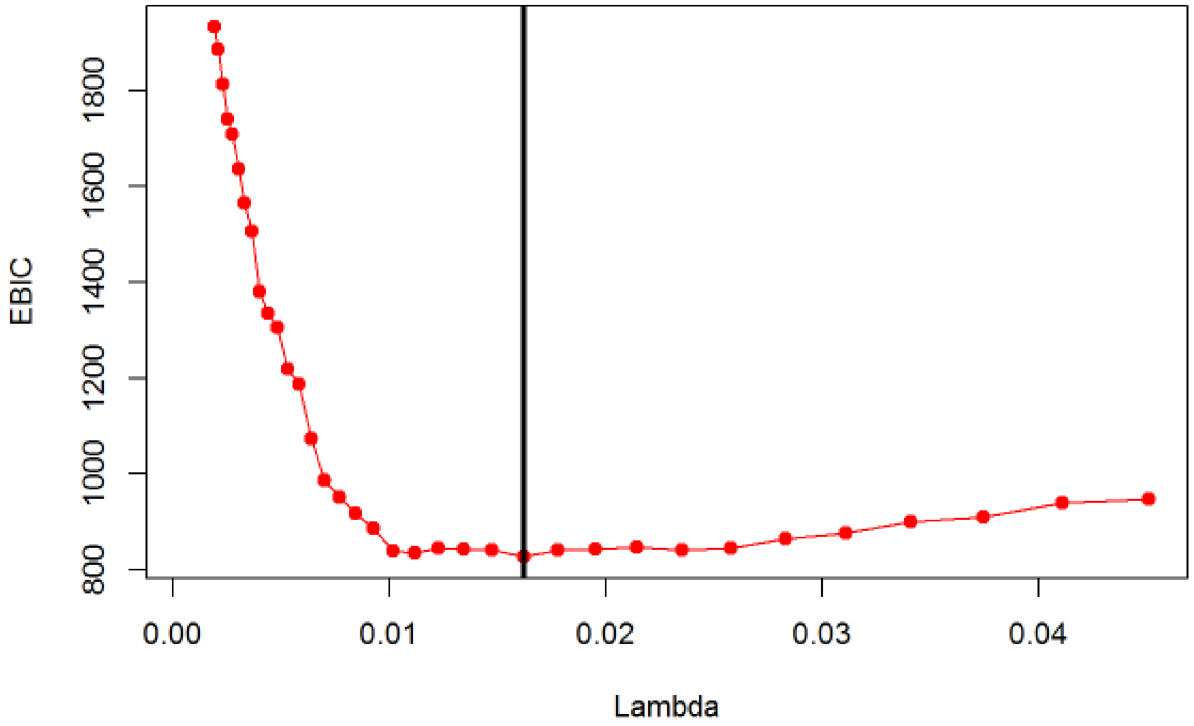
- The best lambda is 0.0111678.
- There are 14 no-zero variables.
- Correct classification rate of training data: 0.9637165
- Area under curve of training data: 0.9105655
- Correct classification rate of test data: 0.9633028
- Area under curve of test data: 0.8523505

$$
\begin{aligned}
ln\frac{p}{1-p} = {} & (-0.87) + (0.32) * gender2 + (-0.23) * agegp25 \\
& + (-0.19) * anti\_ht1 + (-0.97) * anti\_chol1 + (-0.49) * drugs\_others1 \\
& + (-0.07) * hypertension1 + (0.35) * chol + (0.01) * GFR\_EPI \\
& + (-0.16) * bvalogr\_USA2 + (0.33) * smkyn2 + (-0.19) * smk\_cat3 \\
& + (1.94) * ang2 + (-0.47) * R\_retino\_cat2
\end{aligned}
$$

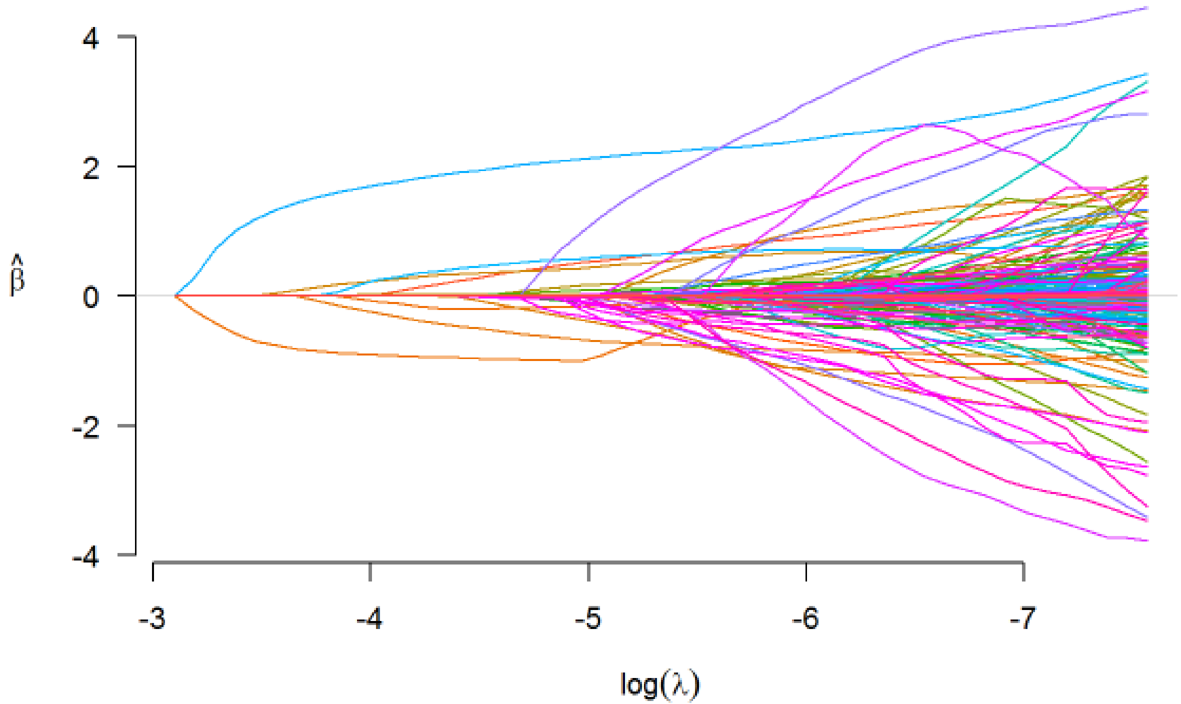
BIC



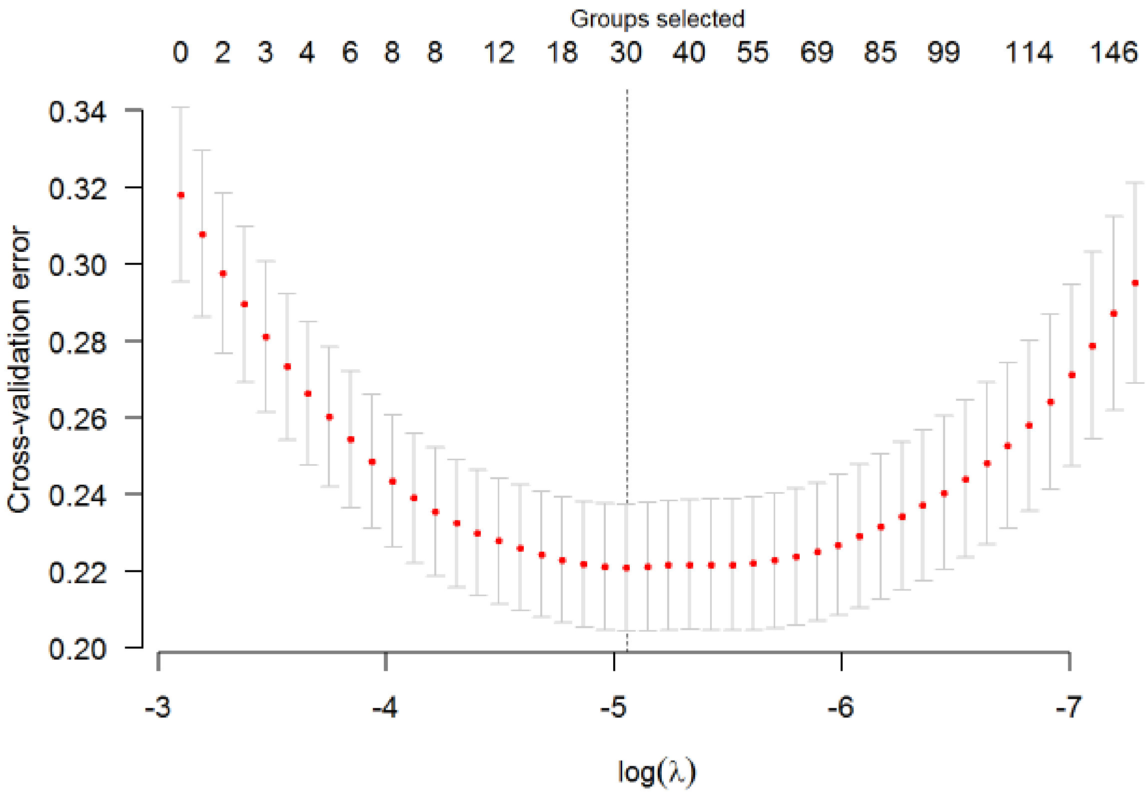
EBIC



Grp LASSO

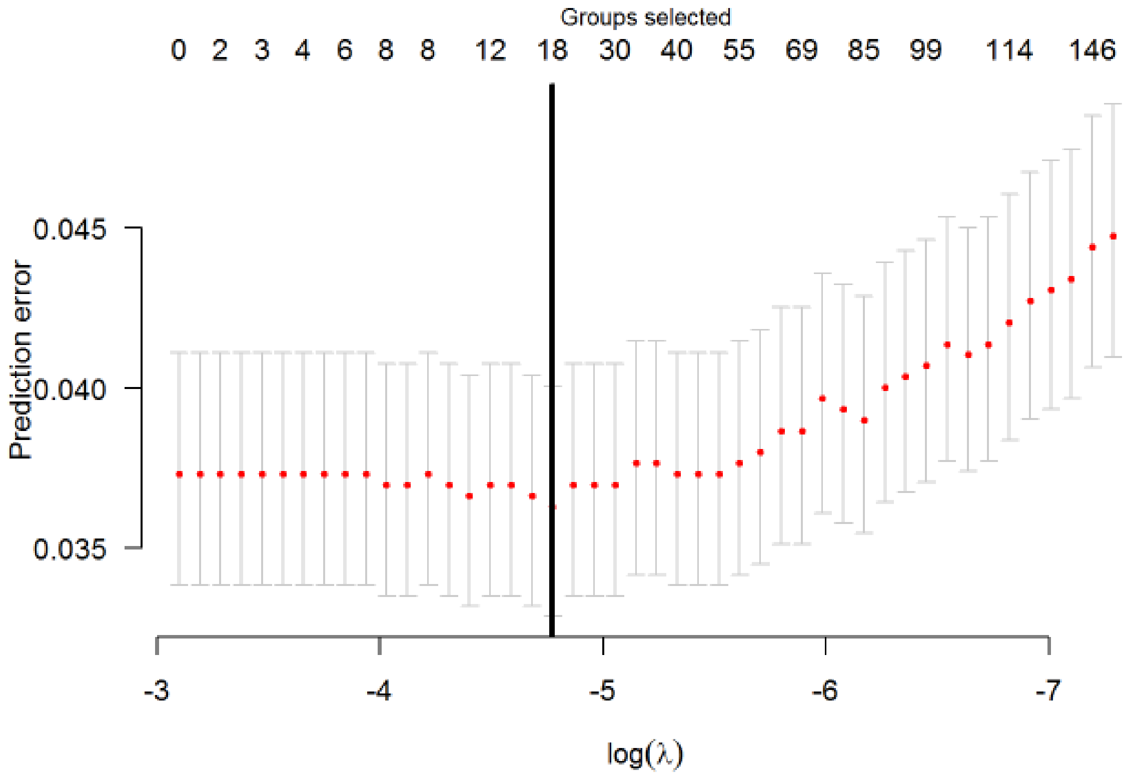


CV: dev



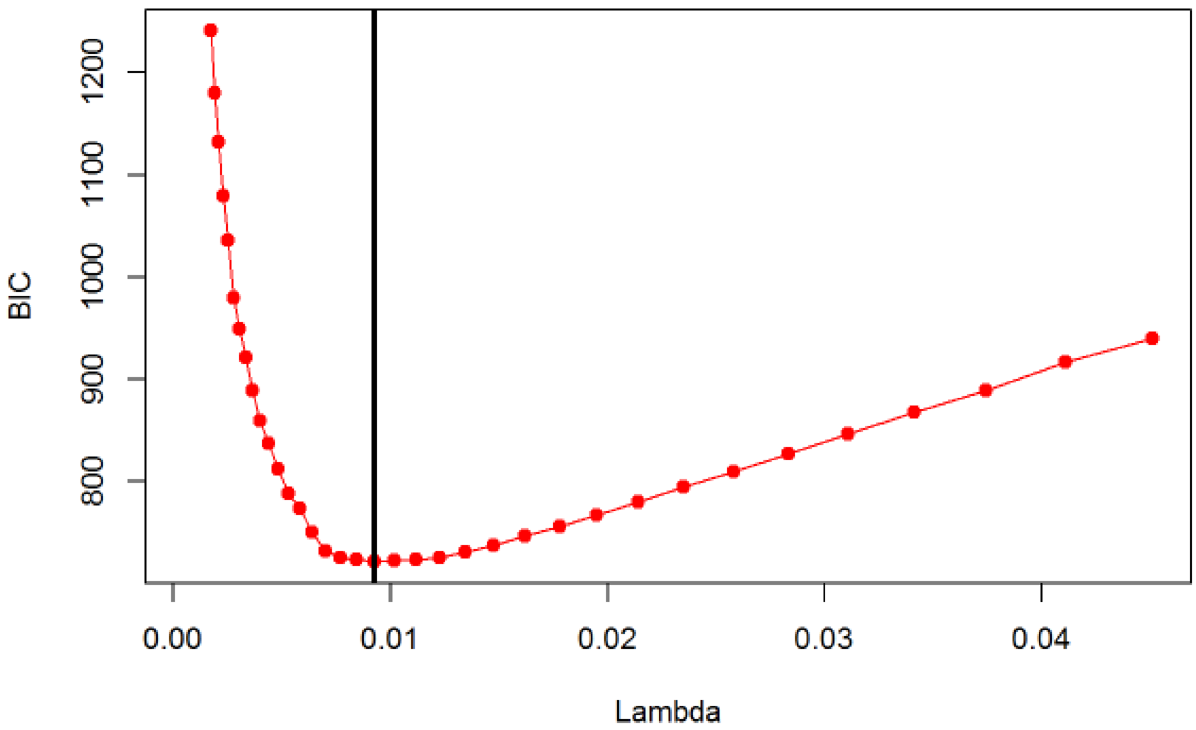
Groups selected
0 2 3 4 6 8 8 12 18 30 40 55 69 85 99 114 146
0.34
0.32
0.30
0.28
0.26
0.24
0.22
0.20
Cross-validation error
-3 -4 -5 -6 -7
log(λ)

CV： ME

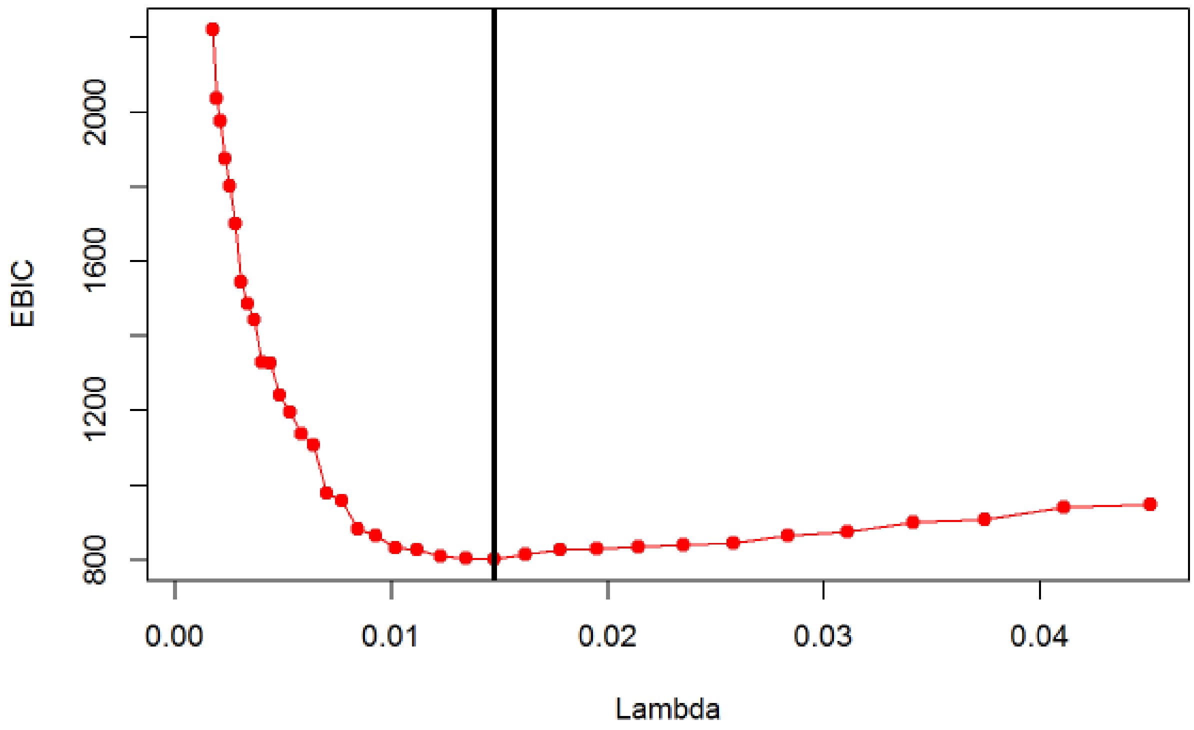


Groups selected
0 2 3 4 6 8 8 12 18 30 40 55 69 85 99 114 146
Prediction error
0.045
0.040
0.035
-3
-4
-5
-6
-7
log(λ)

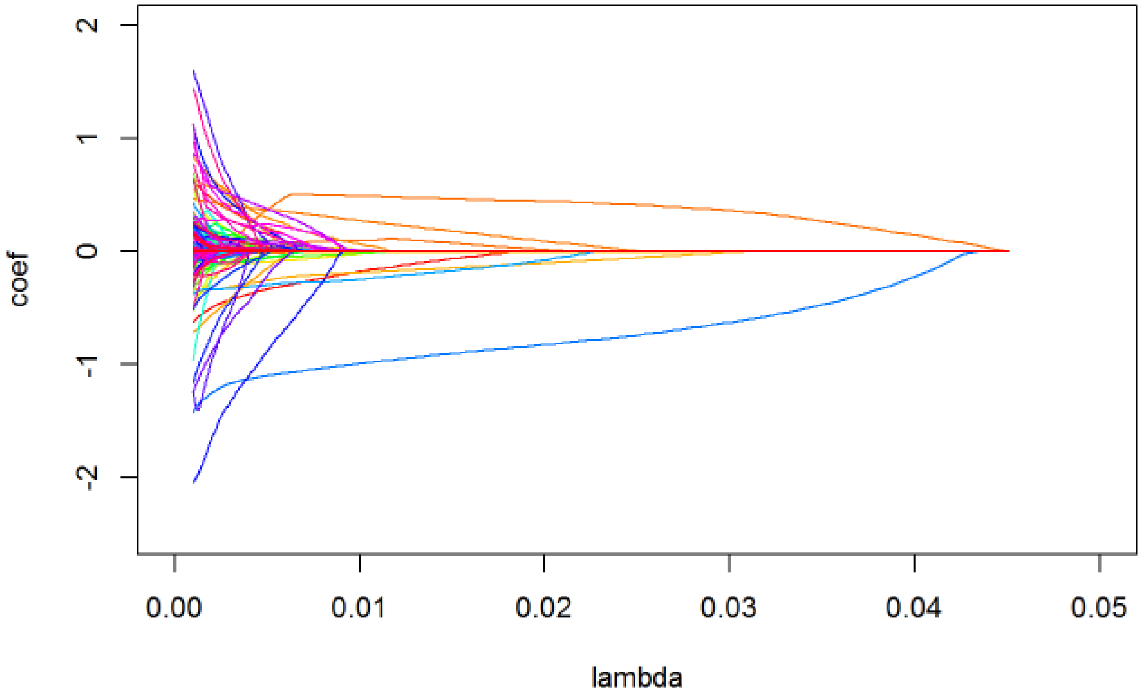
BIC



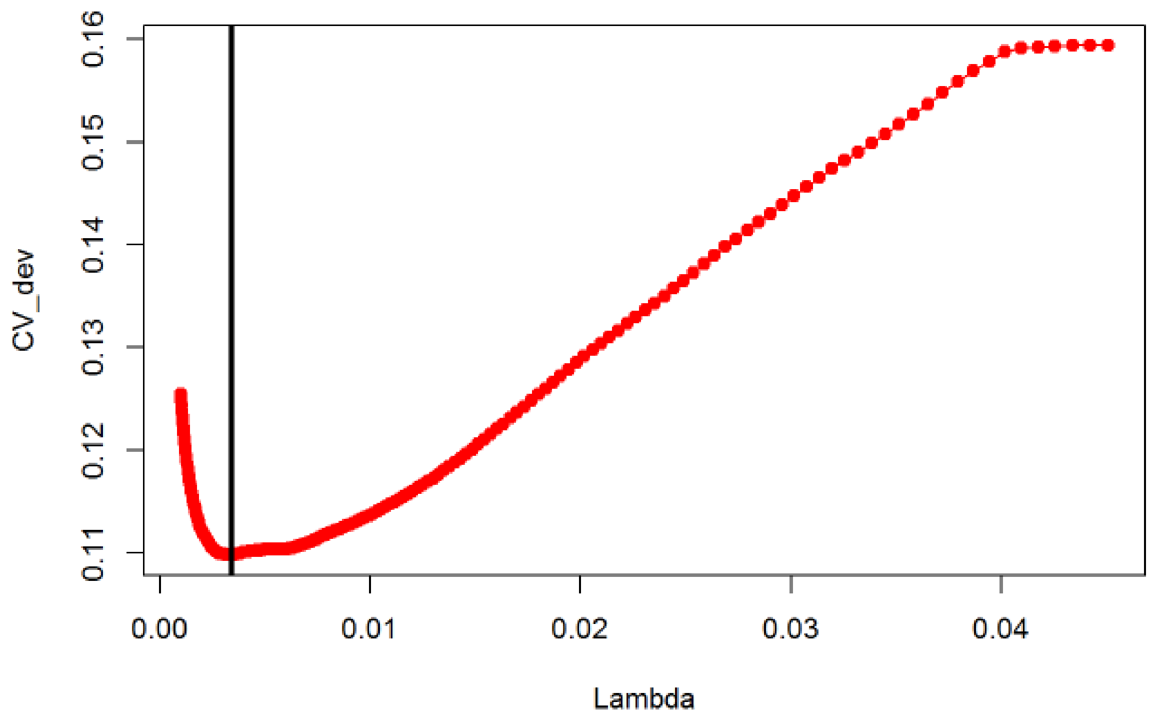
EBIC



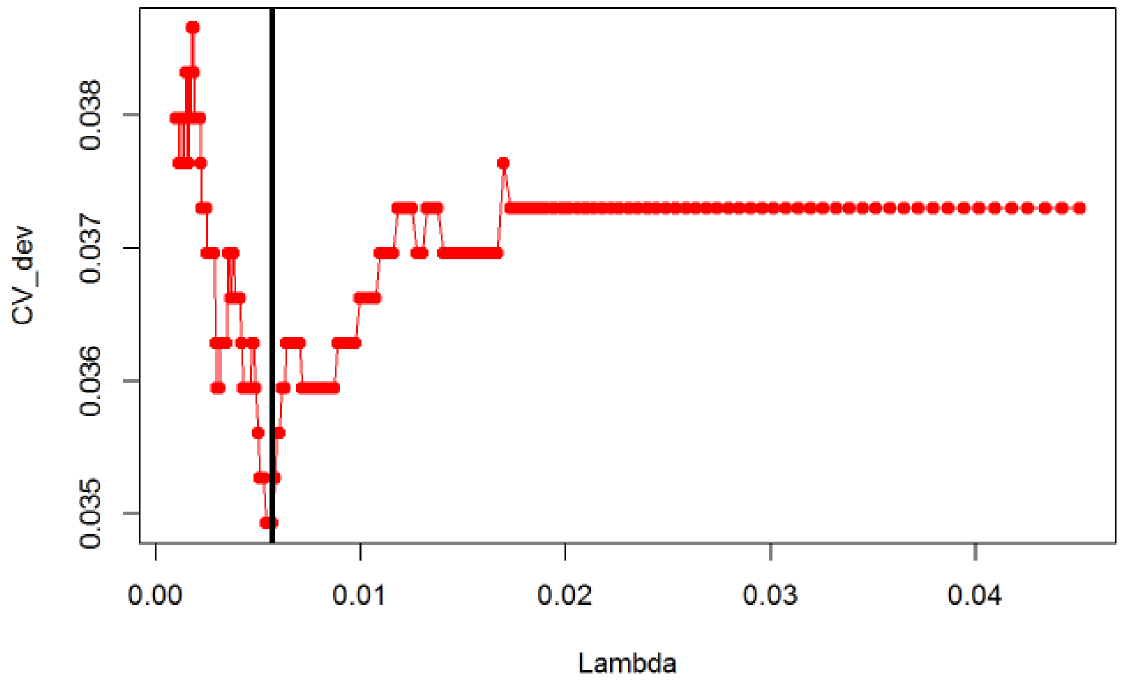
Sparse Grp LASSO



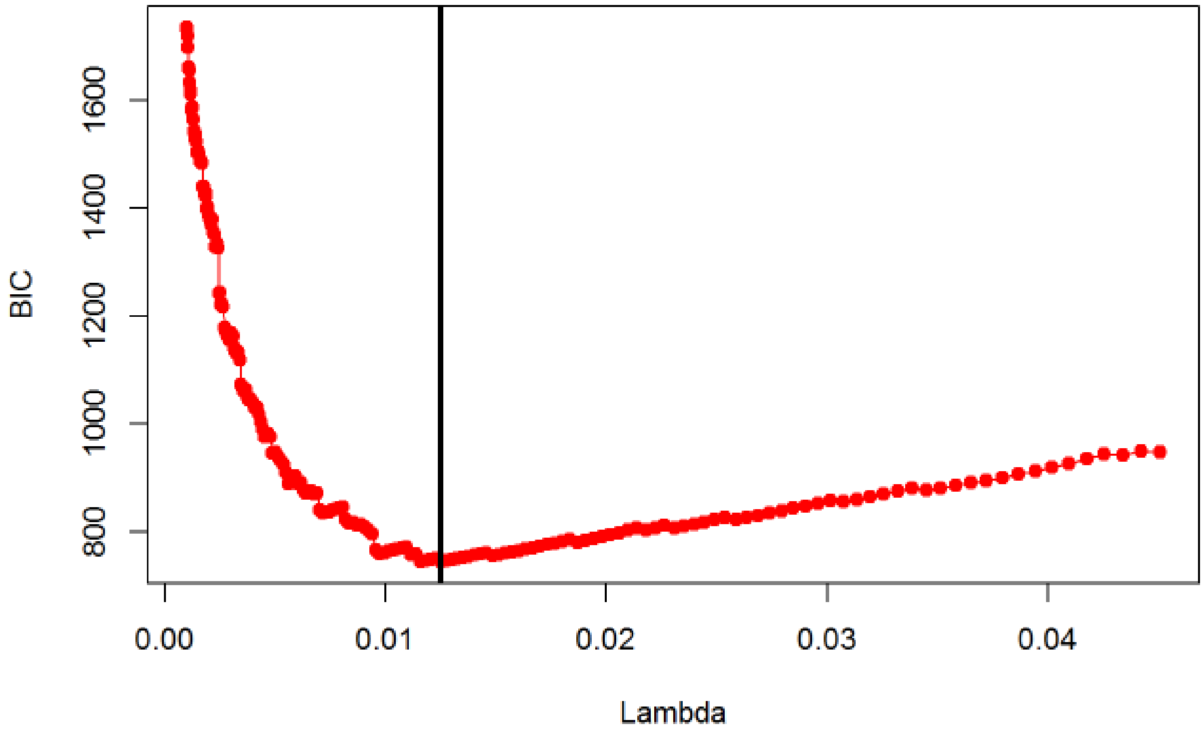
CV: dev



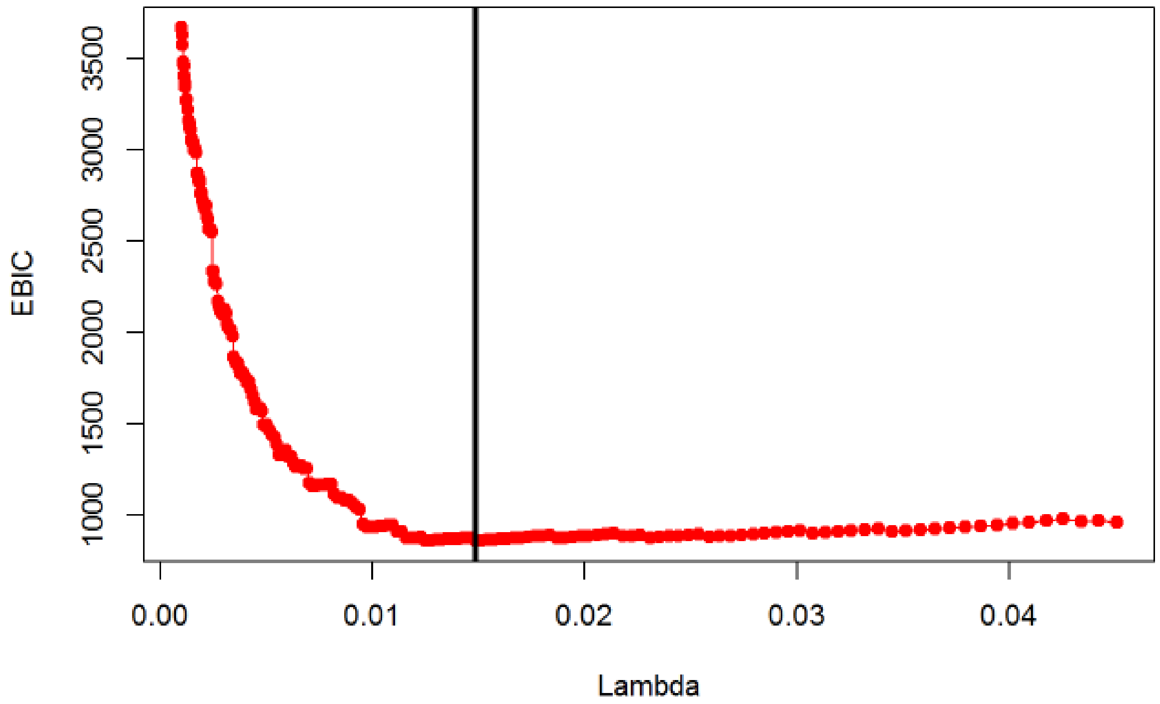
CV： ME



BIC



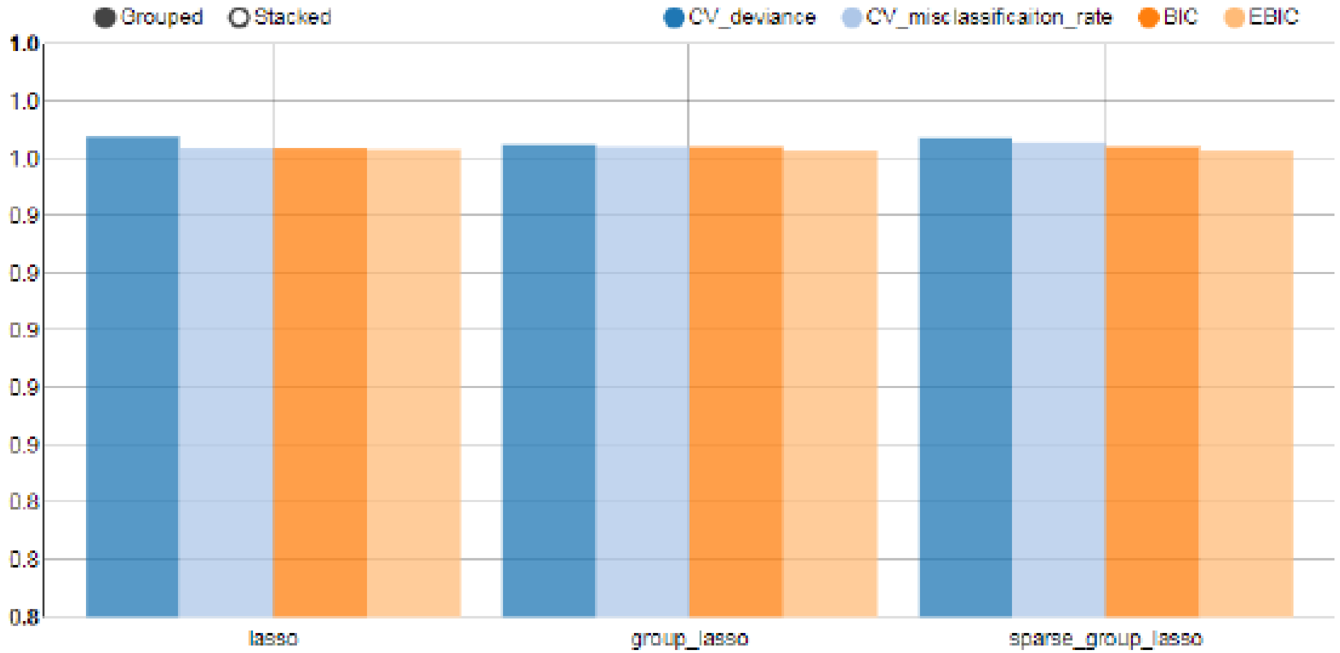
EBIC



Comparison

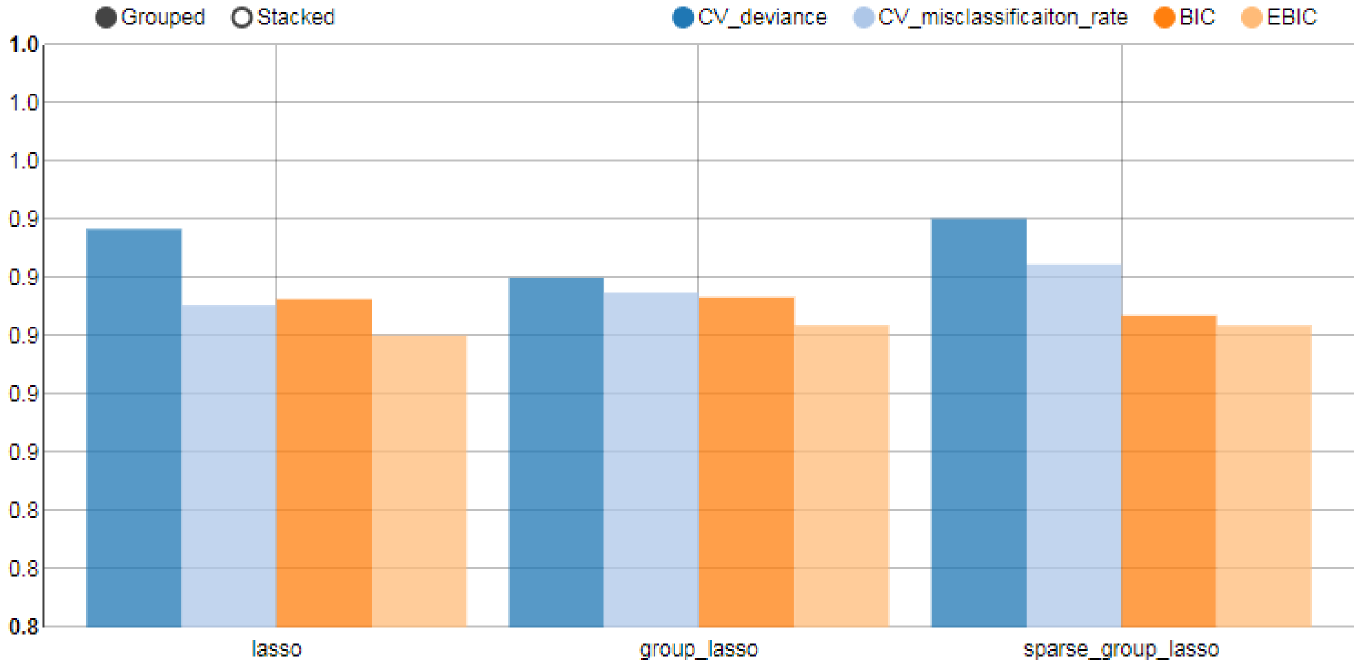
	lambda	num_non_zero	CCR_train	HUM_train	CCR_test	HUM_test
CV_dev_L	0.004	63	0.968	0.937	0.960	0.846
CV_ME_L	0.011	14	0.964	0.911	0.963	0.852
BIC_L	0.010	15	0.964	0.913	0.963	0.851
EBIC_L	0.016	10	0.963	0.900	0.963	0.846
CV_dev_GL	0.006	37	0.965	0.920	0.960	0.839
CV_ME_GL	0.008	19	0.964	0.915	0.963	0.832
BIC_GL	0.009	17	0.964	0.913	0.963	0.834
EBIC_GL	0.015	9	0.963	0.904	0.963	0.841
CV_dev_SGL	0.003	74	0.967	0.940	0.960	0.840
CV_ME_SGL	0.006	38	0.966	0.925	0.960	0.846
BIC_SGL	0.013	10	0.964	0.907	0.963	0.840
EBIC_SGL	0.015	9	0.963	0.904	0.963	0.841

CCR_train



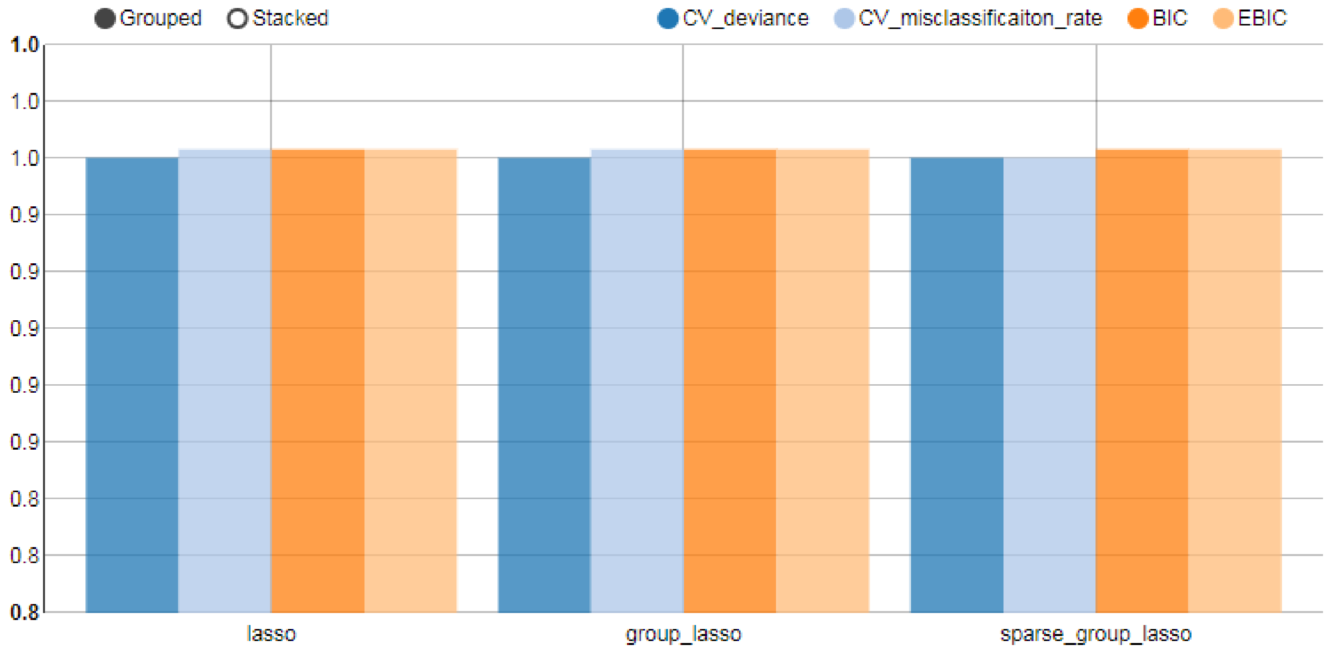
Grouped
Stacked
CV_deviance
CV_misclassificaiton_rate
BIC
EBIC
1.0
1.0
1.0
0.9
0.9
0.9
0.9
0.9
0.8
0.8
0.8
lasso
group_lasso
sparse_group_lasso

HUM_train



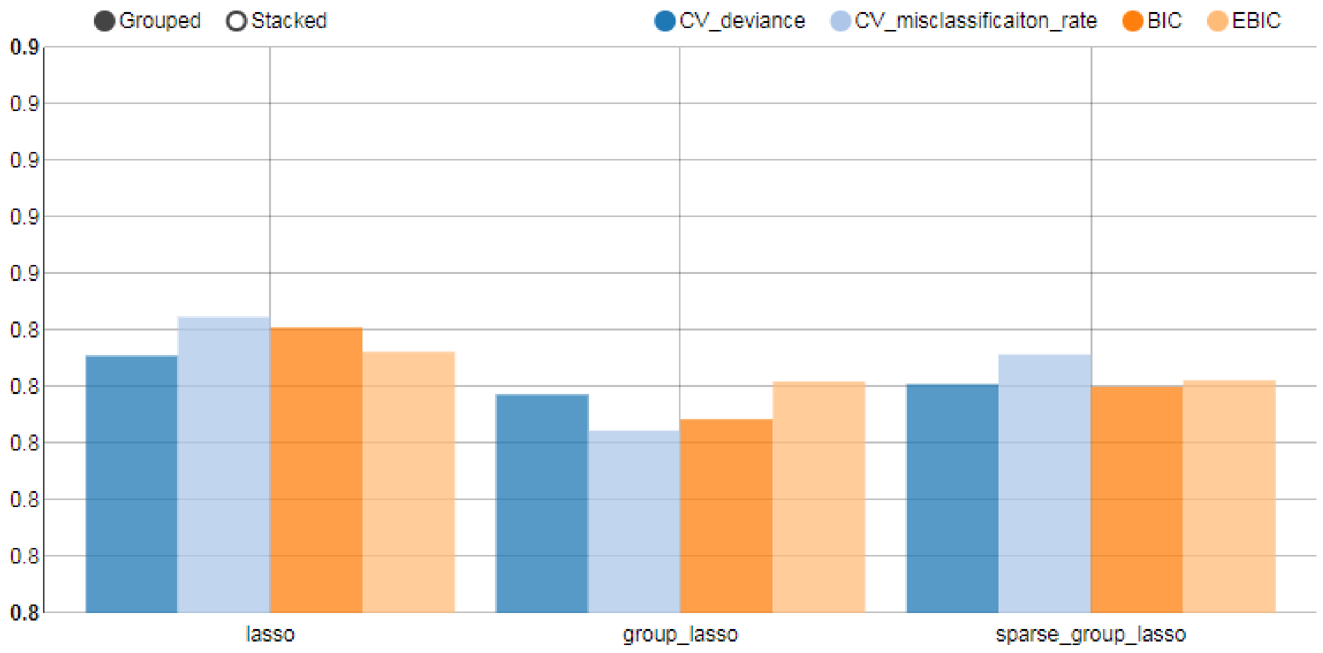
Grouped
Stacked
CV_deviance
CV_misclassificaiton_rate
BIC
EBIC
1.0
1.0
1.0
0.9
0.9
0.9
0.9
0.9
0.8
0.8
0.8
lasso
group_lasso
sparse_group_lasso

CCR_test



Grouped
Stacked
CV_deviance
CV_misclassificaiton_rate
BIC
EBIC
1.0
1.0
1.0
0.9
0.9
0.9
0.9
0.9
0.8
0.8
0.8
lasso
group_lasso
sparse_group_lasso

HUM_test



Grouped
Stacked
CV_deviance
CV_misclassificaiton_rate
BIC
EBIC
0.9
0.9
0.9
0.9
0.9
0.8
0.8
0.8
0.8
0.8
0.8
lasso
group_lasso
sparse_group_lasso

SGL with EBIC is the Best

- 9 predictors with an interception
- 0.841 AUC value

$$ln\frac{p}{1-p} = (-1.473) + (0.067) * gender2 + (-0.078) * anti\_ht1 + (-0.467) * anti\_chol1 + (-0.185) * drugs\_others1 + (0.146) * chol + (0.006) * GFR\_EPI + (0.172) * smkyn2 + (0.904) * ang2$$

	Variable's.code	Meaning	Type	Range
1	gender2	gender	binary	1:female
2	anti_ht1	Anti-hyperstensive drugs	binary	1:yes
3	anti_chol1	Anti-cholesterol drugs	binary	1:yes
4	drugs_others1	Drugs - Others	binary	1:yes
5	chol	Blood Total Cholesterol	continuous	2-14
6	GFR_EPI	Glomerular Filtration Rate (EPI)	continuous	3-300
7	smkyn2	Have you ever smoked?	binary	1:no
8	ang2	Angina (self-reported history)	binary	1:no

谢谢！