

# Generalized Group Elastic Net for Predictive Biomarker Identification

Wenxuan Deng<sup>1</sup>

<sup>1</sup>Department of Biostatistics, Yale University,

**Predictive biomarker identification is a significant problem when targeting patient subpopulation who gets an enhanced benefit under treatment. This paper proposed a new method, Predictive Effects Net (PEN), based on group lasso and special hierarchical structure for figuring out predictive effects. The new approach takes predictive biomarkers as interaction effects between treatment and biomarkers. To show PEN has an supreme performance, this paper shows simulations in different scenerios and comparisons with several other variable selection methods.**

## 1 Introduction

Prognostic biomarkers and predictive biomarkers.

Group lasso (1)

Elastic net (2) adaptive weights for elastic net (3)

Hierarchical Group lasso for interactions (4)

Overlapping group lasso (5) (6) (7)

Sparse Group Lasso (8) (9)

Structured group lasso (10) Group lasso for logistic regression (11)

Other variable selection methods:

GUIDE: a regression tree (12) (13)

SIS: screening (14) (15)

SIR: (16) (17)

Stepwise selection: (18)

## 2 Methods

### 2.1 Model

$$Y = X_0\beta_0 + X_T\beta_\tau + X_1\beta_1 + X_T \otimes X_1\beta_2 + \epsilon \quad (1)$$

Where  $X_0$  is the baseline variables,  $X_T$  is the treatment variable,  $X_1$  is the high dimensional design matrix of genes, i.e. gene expression levels, SNP and mutations, and  $X_T \otimes X_1$  is the interaction between genes and treatment.  $\beta = (\beta_0, \beta_\tau, \beta_1, \beta_2)$  is the corresponding coefficients.  $\epsilon$  is random error.

Let

$$X = [X_0, X_T, X_1^{(1)}, \dots, X_m^{(1)}, X_T X_1^{(1)}, \dots, X_T X_1^{(m)}] \quad (2)$$

and

$$\beta = [\beta_0, \beta_\tau, \beta_1^{(1)}, \dots, \beta_1^{(m)}, \beta_2^{(1)}, \dots, \beta_2^{(m)}] \quad (3)$$

For each gene  $l$ , its prognostic and predictive design matrix is denoted as  $X^{(l)} = [X_1^{(l)}, X_2^{(l)}]$  where  $X_2^{(l)} = X_T X_1^{(l)}$  and its corresponding coefficients are  $\beta^{(l)} = [\beta_1^{(l)}, \beta_2^{(l)}]$

## 2.2 Loss Function

We used group lasso and elastic net for variables selection when  $n \ll p$ , and assumed the hierarchical relationship between prognostic biomarkers and predictive biomarkers, that is the predictive biomarkers should be a prognostic biomarkers. The loss function is

$$\min_{\beta} f(\beta|Y, X_0, X_T, X_1) + g(\beta) \quad (4)$$

$$g(\beta) = \lambda_1 \sum_i \phi_i |\beta_2^{(i)}| + \lambda_1 \sum_i \psi_i \sqrt{(\beta_1^{(1)})^2 + (\beta_2^{(1)})^2} + \lambda_2 (\|\beta_1\|_2^2 + \|\beta_2\|_2^2) \quad (5)$$

Where  $\beta = (\beta_0, \beta_\tau, \beta_1, \beta_2)$  is the parameter, and  $f(\beta|Y, X_0, X_T, X_1)$  is  $L_2$  loss function. When the model is the ordinary linear model, the  $L_2$  loss function is  $\|Y - (X_0\beta_0 + X_T\beta_\tau + X_1\beta_1 + X_T \otimes X_1\beta_2)\|^2$ . Penalty function  $g(\beta)$  can construct a complex hierarchical selection of  $\beta_1$  and  $\beta_2$ , that nonzero  $\beta_2$  is a sufficient but not necessary condition for nonzero  $\beta_1$ . The contour plot for a pair of  $\beta_1$  and  $\beta_2$  is shown in Figure 1.  $\lambda_1$  and  $\lambda_2$  are regularization parameters.

## 2.3 Criterion and Adaptive Weights

### 2.3.1 KKT conditions

KKT (19)

For group  $\hat{\beta}^{(l)}$ , the KKT condition is

$$X^{(l)T}(Y - X^T\hat{\beta}) = \lambda_1 \phi_l \begin{bmatrix} 0 \\ v \end{bmatrix} + \lambda_1 \psi_l u + \frac{1}{2} \lambda_2 \hat{\beta}^{(l)} \quad (6)$$

where

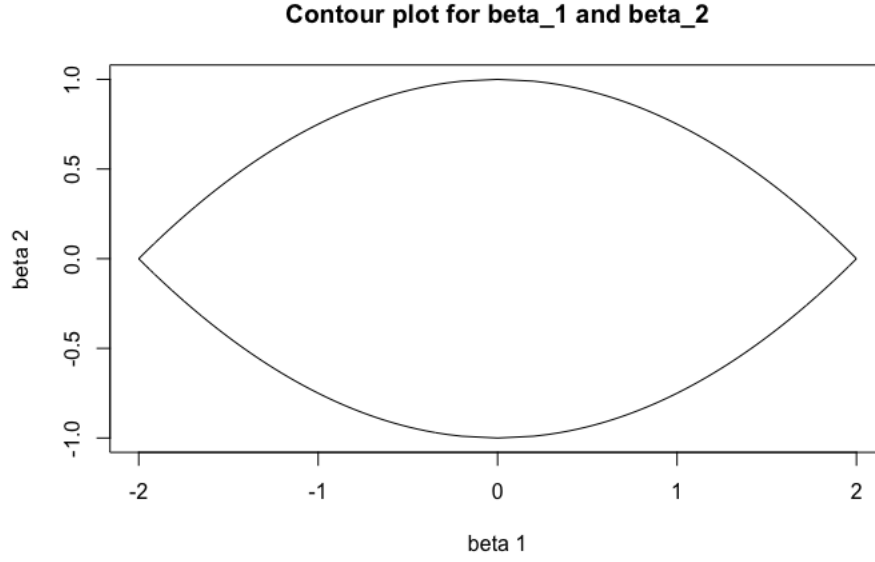


Figure 1: Geometrical interpretation of penalty function

$$v = \begin{cases} \text{sign}(\hat{\beta}_2^{(l)}) & \text{if } \hat{\beta}_2^{(l)} \neq 0 \\ \in \{v : |v|_1 \leq 1\} & \text{if } \hat{\beta}_2^{(l)} = 0 \end{cases} \quad (7)$$

$$u = \begin{cases} \hat{\beta}^{(l)} / \|\hat{\beta}^{(l)}\|_2 & \text{if } \hat{\beta}^{(l)} \neq 0 \\ \in \{u : \|u\|_2 \leq 1\} & \text{if } \hat{\beta}^{(l)} = 0 \end{cases} \quad (8)$$

- We now investigate when  $\hat{\beta}^{(l)} = 0$  satisfies KKT condition 6. We propose the following condition

$$S(X_1^{(l)T} r_{(-l)}, 0)^2 + S(X_2^{(l)T} r_{(-l)}, \lambda_1 \phi_l)^2 \leq \lambda_1^2 \psi_l^2 \quad (9)$$

where

$$S(z, a) = \text{sign}(z)(|z| - a)_+ \quad (10)$$

and

$$r_{(-l)} = Y - X^{(-l)T} \hat{\beta}^{(-l)} \quad (11)$$

Under this condition, we can find

$$u = \begin{bmatrix} \frac{X_1^{(l)} r_{(-l)}}{\lambda_1 \psi_l} \\ \frac{S(X_2^{(l)T} r_{(-l)}, \lambda_1 \phi_l)}{\lambda_1 \psi_l} \end{bmatrix} \quad (12)$$

$$v = \begin{bmatrix} 0 \\ \frac{X^{(l)T} r_{(-l)} - S(X_2^{(-l)T} r_{(-l)}, \lambda_1 \phi_l)}{\lambda_1 \psi_l} \end{bmatrix} \quad (13)$$

such that  $\|u\|_2 \leq 1$  and  $|v|_\infty \leq 1$ . By simple algebra, the subgradient equation 6 was satisfied with  $\hat{\beta}^{(l)} = 0$

- if KKT condition holds with  $\hat{\beta}^{(l)} \neq 0$  but  $\hat{\beta}_2^{(l)} = 0$ , the KKT condition can be reformulated as following

$$X^{(l)T} (Y - X^{(-I(l))T} \hat{\beta}^{(-I(l))}) = \lambda_1 \phi_l \begin{bmatrix} 0 \\ v \end{bmatrix} + \lambda_1 \psi_l \frac{\hat{\beta}^{(l)}}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2} \lambda_2 \hat{\beta}^{(l)} \quad (14)$$

To satisfy  $\hat{\beta}_2^{(l)} = 0$  and KKT condtion, we should have

$$X_2^{(l)T} r_{(-I(l))} \leq \lambda_1 \phi_l \quad (15)$$

where  $r_{(-I(l))} = Y - X^{(-I(l))T} \hat{\beta}^{(-I(l))}$  and  $I(l)$  is the interaction effect nested in biomarker  $l$ th group.

- if KKT condition holds with  $\hat{\beta}^{(l)} \neq 0$  as well as  $\hat{\beta}_2^{(l)} \neq 0$ , the KKT condition is reformulated as

$$X^{(l)T} (Y - X^{(-I(l))T} \hat{\beta}^{(-I(l))}) = \lambda_1 \phi_l \begin{bmatrix} 0 \\ \text{sign}(\hat{\beta}_2^{(l)}) \end{bmatrix} + \lambda_1 \psi_l \frac{\hat{\beta}^{(l)}}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2} \lambda_2 \hat{\beta}^{(l)} \quad (16)$$

Then we get

$$\begin{aligned}\hat{\beta}_2^{(l)} &= \frac{X_2^{(l)T} (Y - X^T \hat{\beta}^{(-I(l))}) - \lambda_1 \phi_l \text{sign}(\hat{\beta}_2^{(l)})}{X_2^{(l)T} X_2^{(l)} + \frac{\lambda_1 \psi_l}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2} \lambda_2} \\ &= \frac{S(X_2^{(l)T} r_{(-I(l))}, \lambda_1 \phi_l)}{X_2^{(l)T} X_2^{(l)} + \frac{\lambda_1 \psi_l}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2} \lambda_2}\end{aligned}\tag{17}$$

$$\hat{\beta}_1^{(l)} = \frac{X_2^{(l)T} r_{(-G(l))}}{X_1^{(l)T} X_1^{(l)} + \frac{\lambda_1 \psi_l}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2} \lambda_2}\tag{18}$$

$$\hat{\beta}_0 = \frac{X_0^T r_{(-0)}}{X_0^T X_0}\tag{19}$$

$$\hat{\beta}_\tau = \frac{X_T^T r_{(-T)}}{X_T^T X_T}\tag{20}$$

### 2.3.2 Adaptive Weights

To give each biomarker equal probability to be prognostic and predictive, we define adaptive weights via a null model that the residual  $\epsilon = r_{(-I(l))}$  is a normal random error where  $\epsilon \sim N(0, \sigma^2)$  when  $\hat{\beta}_2^{(l)} = 0$ . Let

$$\begin{aligned}\|X_2^{(l)T} r_{(-I(l))}\|_2 &= \lambda_1 \phi_l \\ E[(X_2^{(l)T} r_{(-I(l))})^2] &= \lambda_1^2 \phi_l^2\end{aligned}\tag{21}$$

Thus, we can get  $\lambda_1^2 \phi_l^2 = \text{Var}(X_2^{(l)T} r_{(-I(l))})$  and

$$\phi_l \propto \|X_2^{(l)}\|_2\tag{22}$$

Since  $\lambda_1$  is regularization parameter, we define  $\phi_l = \|X_2^{(l)}\|_2$  without loss generality.

On the other hand, based on inequality 9 and results from formula 21, we let

$$\mathbb{E}[S(X_1^{(l)T} r_{(-l)}, 0)^2 + S(X_2^{(l)T} r_{(-l)}, \lambda_1 \phi_l)^2] = \lambda_1^2 \psi_l^2 \quad (23)$$

and assume  $r_{(-l)} = \epsilon \sim N(0, \sigma^2)$  if  $\beta^{(l)} = 0$ , thus  $\epsilon_1 = X_2^{(l)T} r_{(-l)} \sim N(0, \lambda_1^2 \phi_l^2)$  and  $\epsilon_0 = \epsilon_1 / \lambda_1 \phi_l \sim N(0, 1)$ .

$$\begin{aligned} \mathbb{E}[S(X_2^{(l)T} r_{(-l)}, \lambda_1 \phi_l)^2] &= \mathbb{E}[\|\epsilon_1\|_2^2 \mathbb{1}_{|\epsilon_1| > \lambda_1 \phi_l}] - 2\lambda_1 \phi_l \mathbb{E}[|\epsilon_1| \mathbb{1}_{|\epsilon_1| > \lambda_1 \phi_l}] \\ &\quad + \lambda_1^2 \phi_l^2 \mathbb{E}[\mathbb{1}_{|\epsilon_1| > \lambda_1 \phi_l}] \end{aligned} \quad (24)$$

$$\begin{aligned} \mathbb{E}[\|\epsilon_1\|_2^2 \mathbb{1}_{|\epsilon_1| > \lambda_1 \phi_l}] &= \mathbb{E}[\|\epsilon_1\|_2^2 (1 - \mathbb{1}_{|\epsilon_1| \leq \lambda_1 \phi_l})] \\ &= \lambda_1^2 \phi_l^2 (1 - \mathbb{E}[\|\epsilon_0\|_2^2 \mathbb{1}_{|\epsilon_0| \leq 1}]) \\ &\approx \lambda_1^2 \phi_l^2 (1 - (0.68 - 2 \frac{1}{\sqrt{2\pi}} \exp(-0.5))) \\ &= (0.32 + \sqrt{\frac{2}{\pi}} \exp(-0.5)) \lambda_1^2 \phi_l^2 \end{aligned} \quad (25)$$

$$\begin{aligned} \mathbb{E}[|\epsilon_1| \mathbb{1}_{|\epsilon_1| > \lambda_1 \phi_l}] &= \lambda_1 \phi_l \mathbb{E}[|\epsilon_0| \mathbb{1}_{|\epsilon_0| > 1}] \\ &= \lambda_1 \phi_l (\mathbb{E}[|\epsilon_0|] - \mathbb{E}[|\epsilon_0| \mathbb{1}_{|\epsilon_0| \leq 1}]) \\ &= \lambda_1 \phi_l (\sqrt{\frac{2}{\pi}} - \sqrt{\frac{2}{\pi}} (1 - \exp(-0.5))) \\ &= \sqrt{\frac{2}{\pi}} \exp(-0.5) \lambda_1 \phi_l \end{aligned} \quad (26)$$

$$\mathbb{E}[\mathbb{1}_{|\epsilon_1| > \lambda_1 \phi_l}] = \mathbb{P}(|\epsilon_0| > 1) \approx 0.32 \quad (27)$$

Take equations 25 - 27 to equation 24, we can get

$$\mathbb{E}[S(X_2^{(l)2} r_{(-l)}, \lambda_1 \phi_l)^2] \approx (0.64 - \sqrt{\frac{2}{\pi}} \exp(-0.5)) \lambda_1^2 \phi_l^2 \quad (28)$$

Insert results of equation (23) into (18), we define

$$\psi_l = \sqrt{\|X_1^{(l)}\|_2 + \{0.64 - \sqrt{\frac{2}{\pi}} \exp(-0.5)\} \|X_2^{(l)}\|_2} \quad (29)$$

such that

$$\lambda_1^2 \psi_l^2 = \lambda_1^2 [\|X_1^{(l)}\|_2 + \{0.64 - \sqrt{\frac{2}{\pi}} \exp(-0.5)\} \|X_2^{(l)}\|_2] \quad (30)$$

## 2.4 Algorithms

To optimize loss function 4, we use proximal algorithm since penalty function 5 is not differential everywhere (20). Our algorithm also implements fast iterative shrinkage-thresholding algorithm with backtracking, adaptive restart for rippling behavior, and adaptive stepwise of cyclic Barzilai-Borwein spectral approach to accelerate convergence (21) (22) (23).

Let

$$Q_{\tau,g}(t, u) = \lambda_1 \phi_l |t_2|_1 + \lambda_1 \psi_l \|t\|_2 + \frac{1}{2\tau} \|t - u\|_2^2 \quad (31)$$

then the proximal operator is defined as

$$\tilde{t} = \arg \min_t Q_{\tau,g}(t, u) \quad (32)$$

For convenience, we denote  $P_{\tau,g}(u) = \tilde{t}$

To get  $P_{\tau,g}(u)$ , we propose the following lemma, which is generalized from fast Overlapping group lasso method (20).



**Lemma 2.1.** *Define proximal operator*

$$\pi_{\lambda_2}^{\lambda_1}(u) = \arg \min_{t \in \mathbb{R}^2} \{g_{\lambda_2}^{\lambda_1}(t) \equiv \frac{1}{2\tau} \|t - u\|_2^2 + \lambda_1 |t_2|_1 + \lambda_2 \|t\|_2\} \quad (33)$$

*The the following equation holds*

$$\pi_{\lambda_2}^{\lambda_1}(u) = \pi_{\lambda_2}^0(v) \quad (34)$$

where

$$\begin{aligned} v_1 &= u_1 \\ v_2 &= \text{sign}(u_2) \max\{|u_2|_1 - \lambda_1, 0\} \\ \pi_{\lambda_2}^0(v) &= \arg \min_{t \in \mathbb{R}^2} \{h_{\lambda_2}(t) \equiv \frac{1}{2\tau} \|t - v\|_2^2 + \lambda_2 \|t\|_2\} \end{aligned} \quad (35)$$

*Proof.* Assume  $x^* = \pi_{\lambda_2}^0(v)$  and  $\phi_{\lambda_2}^{\lambda_1}(x^*) = \lambda_1 |x_2^*|_1 + \lambda_2 \|x^*\|_2$ . Then

$$0 \in \partial h_{\lambda_2}(x^*) = x^* - v + \partial \phi_{\lambda_2}^0(x^*) \quad (36)$$

$$\partial g_{\lambda_2}^{\lambda_1}(x^*) = x^* - u + \partial \phi_{\lambda_2}^{\lambda_1}(x^*) \quad (37)$$

Because we have  $-v + \partial \phi_{\lambda_2}^0(x^*) \in -u + \partial \phi_{\lambda_2}^{\lambda_1}(x^*)$ , the above equations imply that  $0 \in \partial g_{\lambda_2}^{\lambda_1}(x^*)$ .

□

Therefore,

$$\begin{aligned} P_{\tau,g}(u) &= \left(1 - \frac{\lambda_2}{\|u\|_2}\right)_+ v \\ v_1 &= u_1 \\ v_2 &= \text{sign}(u_1) \max\{|u_2| - \lambda_1\} \end{aligned} \quad (38)$$

Based on equation 38, the algorithm framework is shown in Algorithm 1.

initialization  $\theta_0 = 0$  or warm start from previous run,  $\tau_0 = 0.1$ , stepsize  $\eta = 0.5$ ;  
**while**  $i \leq k$  **do**  
     $u_i = \theta_{i-1} - \tau_i \nabla f(\theta_{i-1})$  Find the smallest nonnegative integers  $s_i$  such that with  
     $\tau_i = \eta^{s_{i-1}} \tau_{i-1}$ ,  $(f + g)(P_{\tau_i, g}(u_i)) \leq Q_{\tau_i, g}(P_{\tau_i, g}(u_i), u_i)$ ;  
    Then, we compute  $t_i = P_{\tau_i, g}(u_i)$  And accelerate the computation by setting **if**  
     $f(\theta_i) + g(\theta_i) > f(\theta_{i-1}) + g(\theta_{i-1})$  **then**  
    |  $\rho_i = 1$   
    **else**  
    |  $\rho_i = \frac{1 + \sqrt{1 + 4\rho_{i-1}^2}}{2}$   
    **end**  
     $\theta_i = t_i + (\frac{\rho_{i-1}-1}{\rho_i})(t_i - t_{i-1})$  and find  $\tau_{i+1}$  that  $\tau_{i+1}I$  can mimic the Hessian  $\nabla^2 f(\theta_i)$   
**end**

**Algorithm 1:** Patient Subgroup Identification Group Lasso Algorithm

## 2.5 Cross Validation and Regularization Parameter

Appropriate regularization parameters,  $\lambda_1$  and  $\lambda_2$ , are critical for variable selection. Previous lasso methods tend to use smallest Mean Error Square(MSE) for optimal regularization parameters. However, in this method, we will not use MSE anymore since that will result in overfitting although the model is simplified. So we used an arbitrary regularization parameters to select the top covariates. But in the future, we will develop an AIC-like approach that can balance the MSE and the model size of predictive effects.

## 3 Experiments

We conducted several experiments to inspect how PEN will perform under different simulation setup. Small sample size is very common in real clinical trial dataset. That results in small  $n$  and big  $p$ , where  $n$  is sample size and  $p$  is dimension. However, most of previous approaches used to set  $n$  as approximately 1000, which is too big to mimic a real clinical trial. In our experiments, we always assume sample size is as small as 100, i.e.  $n = 100$ .

On the other hand, the design matrix contains 5 baseline covariates, 1 treatment covariate

and  $p$  biomarkers, where  $p$  is ranged from  $\frac{n}{2}$  to  $2n$ . Experiments with different  $p$  can help us identify how the ratio  $n$  and  $p$  will change the variable selection.

We also conducted different proportions of nonzero (1-sparsity) prognostic and predictive effects from 10% to 40% and 5% to 20%, respectively. Since PEN has a strong assumption of hierarchical structure between prognostic and predictive effects, the sparsity of predictive biomarker is always bigger than the sparsity of prognostic biomarkers, due to the reason that a biomarker is prognostic before it is predictive.

To test different signal to noise ratio (SNR), PEN did variable selection on different SNR, which is defined as  $SNR = \frac{Var(X\beta)}{Var(\epsilon)}$ .

Since the correlation among SNPs and biomarkers from the same pathway is common, we also assume a blockwise AR(1) correlation structure with  $\rho = .3$ , where the sizes of blocks are sampling from multinomial distribution where the mean for block size is 5.

	Proportion	SNR	Dimension	SNP
# Predictive biomarkers	5%, 10%, 15%, 20%	10%	10%	10%
# Biomarkers	100	100	50, 100, 200	100
SNR	10	1, 5, 10, 20, 100	10	10
Covariate Type	N(0,1)	N(0,1)	N(0,1)	N(0,1) for baseline and treatment covariates; Binom(2,0.5) for genomics covariates

Table 1: Summary of Simulation Setup

All experiments of PEN ("glasso" in tables and figures) were compared with other standard variable selection methods: General Elastic Net without penalizing baseline and treatment variables (Lasso), Bayesian Model Averaging (BMA), Stepwise Variable Selection by likelihood (step), Iterative Sure Independent Screening (SIS) and Random Forest.

Table 1 shows the summary of different simulation setups we conducted. All cases were run

100 times with fixed seed 1001-1100. The coefficients for baseline and treatment covariates are sampled from gaussian distribution while the coefficients for genomics covariates are constants and randomly picked up from  $\pm 3$  and  $\pm 5$ .

**Proportion of Nonzero Predictive Effects** The proportion of nonzero predictive effects indicates the number of biomarkers which are relate with outcomes. Although PEN was applied on the simulation with 20% nonzero predictive effects, real datasets usually has much higher sparsity. Figure 2 shows a comprehensive analysis with different predictive effects sparsities. We inspected the results from several metrics: the difference between true and estimated parameters, sum square of errors (SSE), positive predictive value (PPV), false negative rate (FNR) and model size for only predictive biomarkers. Final goal for PEN is to enhance the variable selection accuracy only for predictive biomarkers, so PPV and FNR for predictive effects are two mosrt important metrics. .

From Figure 2 and 3, we observe that PEN always gets almost highest PPV in all different scenerios in the comparisons with other five methods. SIS can achieve a bit higher PPV when proportions are 15% and 20%. But it is still not reliable due to two reasons. Firstly, only proportion of  $< 10\%$  is practical in real datasets. Cases of 15% and bigger imply too much significant biomarkers. Secondly, SIS has extremely high FNR whatever the proportion is. That is due to the limited number of top covariates SIS selects. As shown in Figure 3 (c), the estimated predictive biomarker model size of SIS are significantly below the ground truth and close to axis. The other observation is that PEN also has a slightly underestimated predictive effect model size. That is also the key reason why PEN does not achieve the lowest FNR. That is why our next step is to develop a new stop criterion. Because of the arbitrary values of regularization parameters, PEN tends to select fewer predictive biomarkers. Our future stop criterion should guarantee an unbiased predictive biomarker model size.

**SNR** Small SNR indicates large noise, e.g. half of the outcome variation can be explained

by the noise when SNR is as small as 1. The comprehensive data analysis and visualization of PPV, FNR and predictive biomarker model size curves are shown in Figure 4, 5 and 6. Similarly, PEN always gets the best performances in terms of PPV and FNR.

**Dimension** In this session, PEN was applied on difference cases when the number of biomarkers are 50, 100, and 200, respectively. PEN is unable to deal with ultra high-dimensional data so far with small sample size, i.e.  $n = 100$ . PEN even does not have a very promising performance even with moderate high-dimensional data, i.e.  $p = 2n = 200$ , although it is the most reliable method compared with other methods. The problem of underestimated model size still exists for PEN based on the results of Figure 7 (c).

**Categorical Covariates** For genetics data when covariates are SNPs, binomial distribution instead of gaussian distribution can better represent the distribution of biomarkers. Shown in Figure 8, PEN has an overwhelming performance in the comparisons with the other five methods and its FNR is approximate 0.75, which is similar with the other approaches. That indicates PEN is still the most reliable one when the covariates are SNPs.

## 4 Discussion

In this project, we proposed a new method called PEN based on hierarchical group lasso and elastic net. The special penalty term design of PEN lets it have a hierarchical structure between prognostic and predictive effects, which could enhance the accuracy of predictive biomarker identification. Simulations on different scenerios have shown PEN is a reliable approach compared with other popular methods such as Elastic Net, SIS, and random forest. However, PEN can be further improved for a better performance. Firstly, a new stop criterion is needed for an unbiased predictive biomarker model size. To solve this problem, it is reasonable to combine prediction error or likelihood with the number of selected predictive biomarkers. Secondly, the hierarchical assumption is too strong and may not hold for some biomarkers. So we need to

conduct more experiments without such relationship. Thirdly, it is better to check whether our estimations satisfy KKT conditions or not. That will help us to understand the results. Finally, adaptive weights could be improved for better estimations.

## References

1. Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
2. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.
3. Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.
4. Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.
5. Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.
6. Daniel Percival et al. Theoretical properties of the overlapping groups lasso. *Electronic Journal of Statistics*, 6:269–288, 2012.
7. Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

8. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.
9. Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
10. Peng Zhao, Guilherme Rocha, Bin Yu, et al. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.
11. Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.
12. Wei-Yin Loh, Xu He, and Michael Man. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine*, 34(11):1818–1833, 2015.
13. Wei-Yin Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386, 2002.
14. Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.
15. Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of machine learning research*, 10(Sep):2013–2038, 2009.
16. Bo Jiang and Jun S Liu. Sliced inverse regression with variable selection and interaction detection. *arXiv preprint arXiv:1304.4056*, 652, 2013.

17. Yang Li and Jun S Liu. Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association*, pages 1–16, 2018.
18. Alan J Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, pages 389–425, 1984.
19. Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.
20. Jun Liu and Jieping Ye. Fast overlapping group lasso. *arXiv preprint arXiv:1009.0306*, 2010.
21. Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
22. Brendan O’donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.
23. Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.



	L2	L1	SSE	PPV all	FNR all	PPV prog	PPV pred	FNR prog	FNR pred	num pred
<i>glasso</i>	20.094	109.746	417.007	0.779	0.61	0.742	0.92	0.568	0.694	3.32
<i>lasso</i>	20.264	112.605	415.909	0.728	0.636	0.771	0.681	0.64	0.628	5.61
<i>Stepwise</i>	20.139	109.993	414.014	0.806	0.597	0.845	0.757	0.6	0.591	5.54
<i>SIS</i>	22.399	125.715	434.41	0.736	0.853	0.785	0.654	0.854	0.851	2.3
<i>Random Forest</i>	24.626	153.615	436.781	0.237	0.86	0.297	0.14	0.833	0.915	6.33
<i>BMA</i>	27.834	183.012	436.781	0.155	0.891	NaN	0.155	1	0.674	21

(a) proportion = 5%

	L2	L1	SSE	PPV all	FNR all	PPV prog	PPV pred	FNR prog	FNR pred	num pred
<i>glasso</i>	16.569	93.874	353.888	0.641	0.359	0.63	0.688	0.288	0.502	7.36
<i>lasso</i>	17.89	106.679	341.056	0.569	0.431	0.663	0.458	0.44	0.411	12.98
<i>Stepwise</i>	23.341	150.19	312.545	0.454	0.546	0.557	0.333	0.544	0.552	13.56
<i>SIS</i>	22.866	129.041	427.122	0.747	0.874	0.769	0.657	0.876	0.871	1.99
<i>Random Forest</i>	24.375	159.322	411.028	0.318	0.648	0.373	0.242	0.637	0.67	13.7
<i>BMA</i>	38.888	279.686	411.028	0.116	0.884	NaN	0.116	1	0.653	30

(b) proportion = 10%

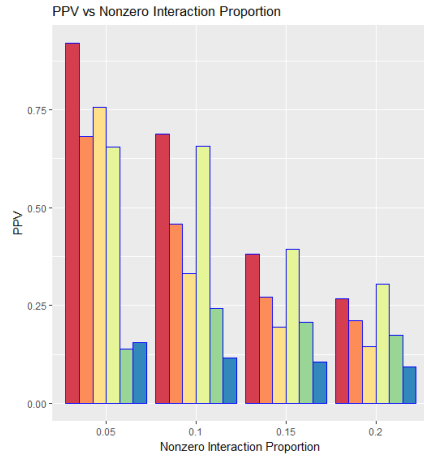
	L2	L1	SSE	PPV all	FNR all	PPV prog	PPV pred	FNR prog	FNR pred	num pred
<i>glasso</i>	18.472	117.706	345.981	0.444	0.333	0.466	0.381	0.231	0.538	12.13
<i>lasso</i>	20.114	133.778	340.223	0.375	0.437	0.451	0.272	0.418	0.475	19.34
<i>Stepwise</i>	27.439	195.657	338.083	0.271	0.594	0.34	0.194	0.594	0.592	21.13
<i>SIS</i>	23.459	134.417	436.151	0.492	0.92	0.528	0.394	0.922	0.917	2.03
<i>Random Forest</i>	26.293	184.904	388.71	0.284	0.545	0.345	0.208	0.539	0.557	21.33
<i>BMA</i>	36.442	260.38	388.71	0.105	0.895	NaN	0.105	1	0.685	30

(c) proportion = 15%

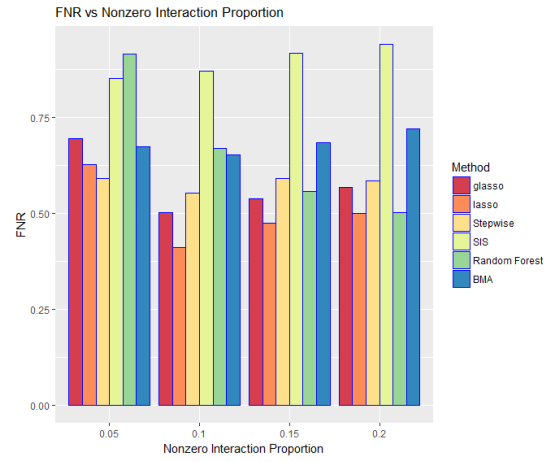
	L2	L1	SSE	PPV all	FNR all	PPV prog	PPV pred	FNR prog	FNR pred	num pred
<i>glasso</i>	20.774	143.908	325.852	0.371	0.258	0.409	0.267	0.103	0.568	16.26
<i>lasso</i>	20.465	143.565	322.515	0.357	0.286	0.452	0.212	0.179	0.499	23.76
<i>Stepwise</i>	30.438	233.134	334.807	0.207	0.586	0.263	0.146	0.587	0.584	28.49
<i>SIS</i>	23.765	137.029	440.872	0.407	0.943	0.433	0.305	0.944	0.941	1.78
<i>Random Forest</i>	31.202	239.743	367.004	0.23	0.517	0.276	0.174	0.524	0.503	28.53
<i>BMA</i>	35.273	251.139	367.004	0.093	0.907	NaN	0.093	1	0.721	30

(d) proportion = 20%

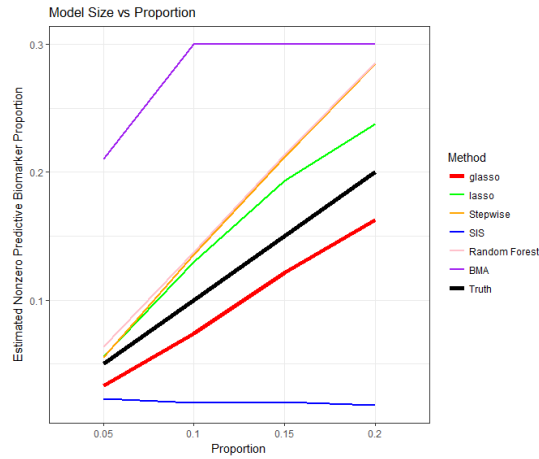
Figure 2: Tables for different proportion when fixing the number of biomarkers as 100, SNR as 10, and continuous covariate type. The proportions of nonzero predictive effects are 5%, 10%, 15%, and 20%. L2 = mean  $\|\hat{\beta} - \beta\|_2$ , L1 = mean  $\|\hat{\beta} - \beta\|_1$ , SSE = Sum Square of Errors, PPV = Positive Predictive Value, FNR = False Negative Rate, num = Model Size, all = across both prognostic and predictive biomarkers, prog = across only prognostic biomarkers, pred = across only predictive biomarkers



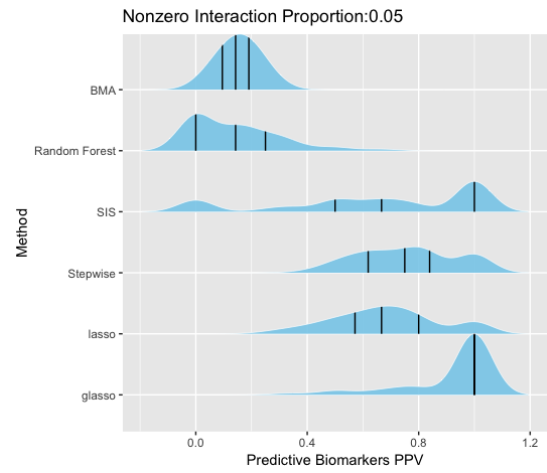
(a) Precision of Predictive Biomarkers Selection



(b) FNR of Predictive Biomarkers Selection



(c) Model Size of Predictive Biomarkers



(d) Distribution of predictive biomarker selection precision when proportion=5% over 100 Independent simulations.

Figure 3: performance of PEN with different proportions of nonzero predictive effects.

	<b>L2</b>	<b>L1</b>	<b>SSE</b>	<b>PPV all</b>	<b>FNR all</b>	<b>PPV prog</b>	<b>PPV pred</b>	<b>FNR prog</b>	<b>FNR pred</b>	<b>num pred</b>
<i>glasso</i>	29.203	185.008	458.579	0.401	0.599	0.403	0.398	0.534	0.727	6.81
<i>lasso</i>	29.662	190.869	460.737	0.343	0.657	0.432	0.245	0.658	0.654	14.19
<i>Stepwise</i>	30.925	203.484	463.265	0.266	0.734	0.334	0.195	0.742	0.719	14.36
<i>SIS</i>	24.602	140.732	485.276	0.516	0.936	0.59	0.423	0.937	0.933	1.73
<i>Random Forest</i>	32.225	215.563	458.783	0.258	0.71	0.334	0.168	0.694	0.74	15.48
<i>BMA</i>	33.314	229.804	458.783	0.101	0.899	NaN	0.101	1	0.696	30

(a) SNR = 1

	<b>L2</b>	<b>L1</b>	<b>SSE</b>	<b>PPV all</b>	<b>FNR all</b>	<b>PPV prog</b>	<b>PPV pred</b>	<b>FNR prog</b>	<b>FNR pred</b>	<b>num pred</b>
<i>glasso</i>	20.369	118.955	405.68	0.575	0.425	0.564	0.623	0.356	0.561	7.15
<i>lasso</i>	20.324	118.592	405.601	0.58	0.42	0.679	0.429	0.38	0.5	11.75
<i>Stepwise</i>	23.58	146.84	418.518	0.41	0.59	0.493	0.314	0.596	0.577	13.54
<i>SIS</i>	23.119	130.447	451.972	0.689	0.889	0.767	0.563	0.892	0.884	2.12
<i>Random Forest</i>	25.502	167.541	420.195	0.307	0.661	0.373	0.218	0.644	0.696	14.05
<i>BMA</i>	28.263	196.45	420.195	0.113	0.887	NaN	0.113	1	0.662	30

(b) SNR = 5

	<b>L2</b>	<b>L1</b>	<b>SSE</b>	<b>PPV all</b>	<b>FNR all</b>	<b>PPV prog</b>	<b>PPV pred</b>	<b>FNR prog</b>	<b>FNR pred</b>	<b>num pred</b>
<i>glasso</i>	16.569	93.874	353.888	0.641	0.359	0.63	0.688	0.288	0.502	7.36
<i>lasso</i>	17.89	106.679	341.056	0.569	0.431	0.663	0.458	0.44	0.411	12.98
<i>Stepwise</i>	23.341	150.19	312.545	0.454	0.546	0.557	0.333	0.544	0.552	13.56
<i>SIS</i>	22.866	129.041	427.122	0.747	0.874	0.769	0.657	0.876	0.871	1.99
<i>Random Forest</i>	24.375	159.322	411.028	0.318	0.648	0.373	0.242	0.637	0.67	13.7
<i>BMA</i>	38.888	279.686	411.028	0.116	0.884	NaN	0.116	1	0.653	30

(c) SNR = 10

Figure 4: Tables for different SNR when fixing the number of biomarkers as 100, proportion as 10%, and continuous covariate type. The proportions of nonzero predictive effects are 1, 2, 5, and 10.

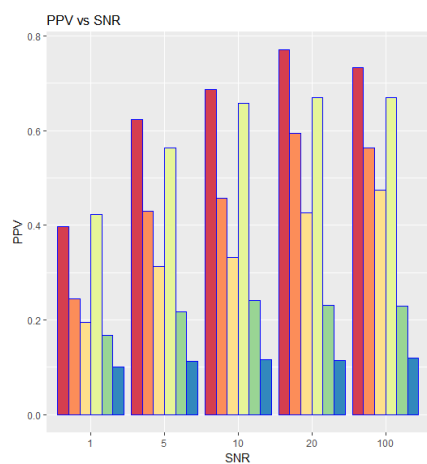
	<b>L2</b>	<b>L1</b>	<b>SSE</b>	<b>PPV all</b>	<b>FNR all</b>	<b>PPV prog</b>	<b>PPV pred</b>	<b>FNR prog</b>	<b>FNR pred</b>	<b>num pred</b>
<i>glasso</i>	21.375	119.795	402.598	0.317	0.704	0.695	0.771	0.456	0.599	5.34
<i>lasso</i>	21.819	123.663	406.793	0.292	0.741	0.764	0.595	0.544	0.535	8.04
<i>Stepwise</i>	19.56	114.27	391.578	0.536	0.464	0.625	0.427	0.462	0.468	12.77
<i>SIS</i>	22.591	126.135	440.827	0.767	0.862	0.818	0.67	0.866	0.852	2.09
<i>Random Forest</i>	23.955	155.791	407.54	0.327	0.64	0.395	0.231	0.618	0.684	13.64
<i>BMA</i>	27.266	188.004	407.54	0.115	0.885	NaN	0.115	1	0.654	30

(a) SNR = 20

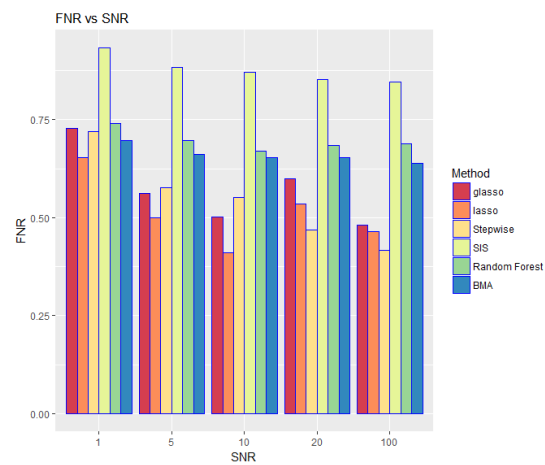
	<b>L2</b>	<b>L1</b>	<b>SSE</b>	<b>PPV all</b>	<b>FNR all</b>	<b>PPV prog</b>	<b>PPV pred</b>	<b>FNR prog</b>	<b>FNR pred</b>	<b>num pred</b>
<i>glasso</i>	15.856	84.172	364.794	0.695	0.305	0.683	0.734	0.218	0.481	7.09
<i>lasso</i>	15.112	78.803	359.894	0.73	0.27	0.812	0.563	0.173	0.465	9.62
<i>Stepwise</i>	18.135	103.565	377.448	0.591	0.409	0.686	0.474	0.405	0.417	12.68
<i>SIS</i>	22.477	126.001	437.874	0.743	0.856	0.781	0.67	0.86	0.847	2.29
<i>Random Forest</i>	23.437	151.734	402.414	0.33	0.638	0.4	0.23	0.612	0.689	13.57
<i>BMA</i>	26.617	182.727	402.414	0.12	0.88	NaN	0.12	1	0.639	30

(b) SNR = 100

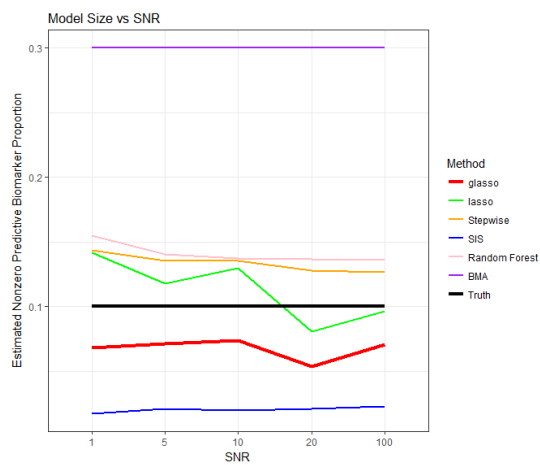
Figure 5: Tables for different SNR when fixing the number of biomarkers as 100, proportion as 10%, and continuous covariate type. The proportions of nonzero predictive effects are 20 and 100.



(a) Precision of Predictive Biomarkers Selection

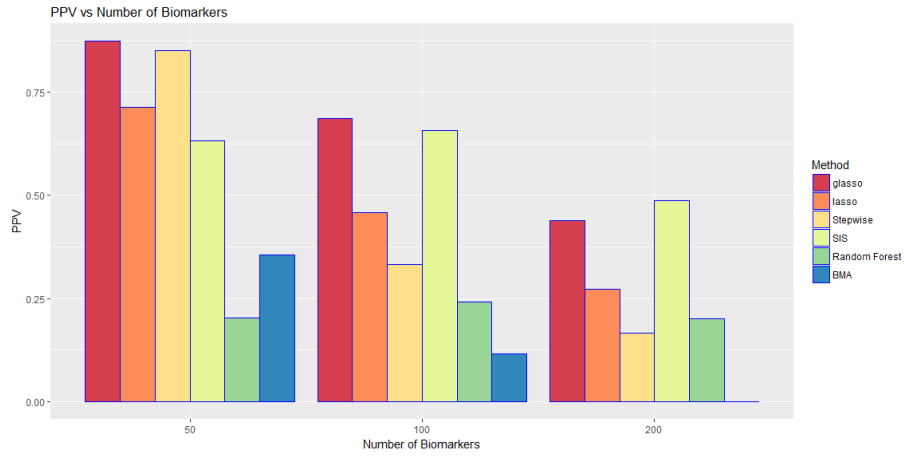


(b) FNR of Predictive Biomarkers Selection

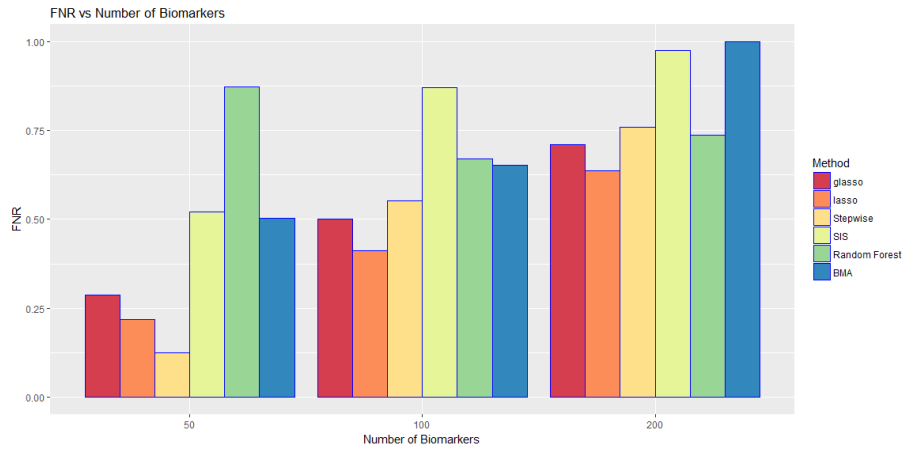


(c) Model Size of Predictive Biomarkers

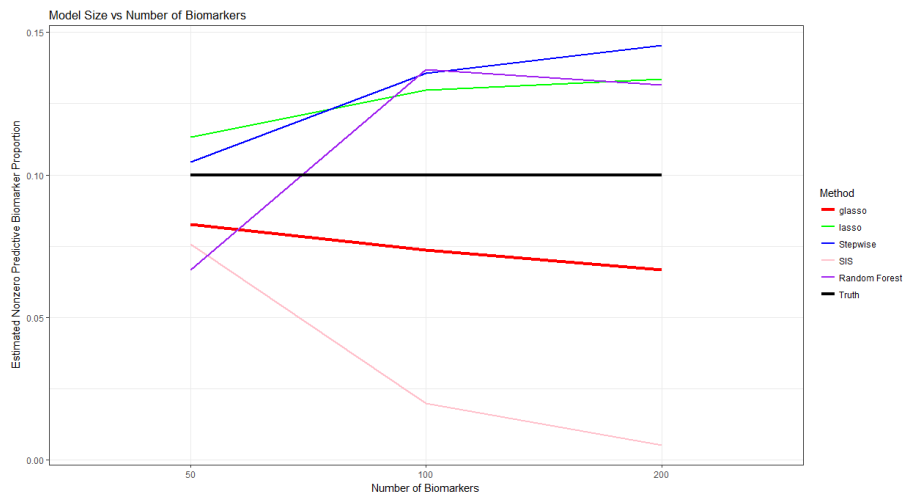
Figure 6: performance of PEN with different SNR.



(a) Precision of Predictive Biomarkers Selection

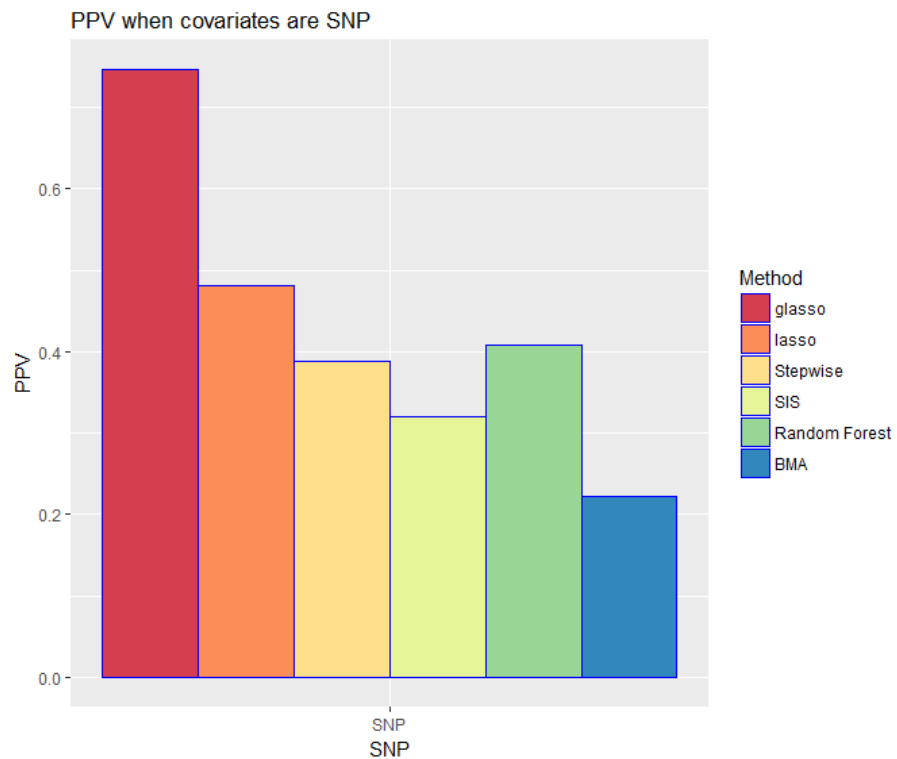


(b) FNR of Predictive Biomarkers Selection

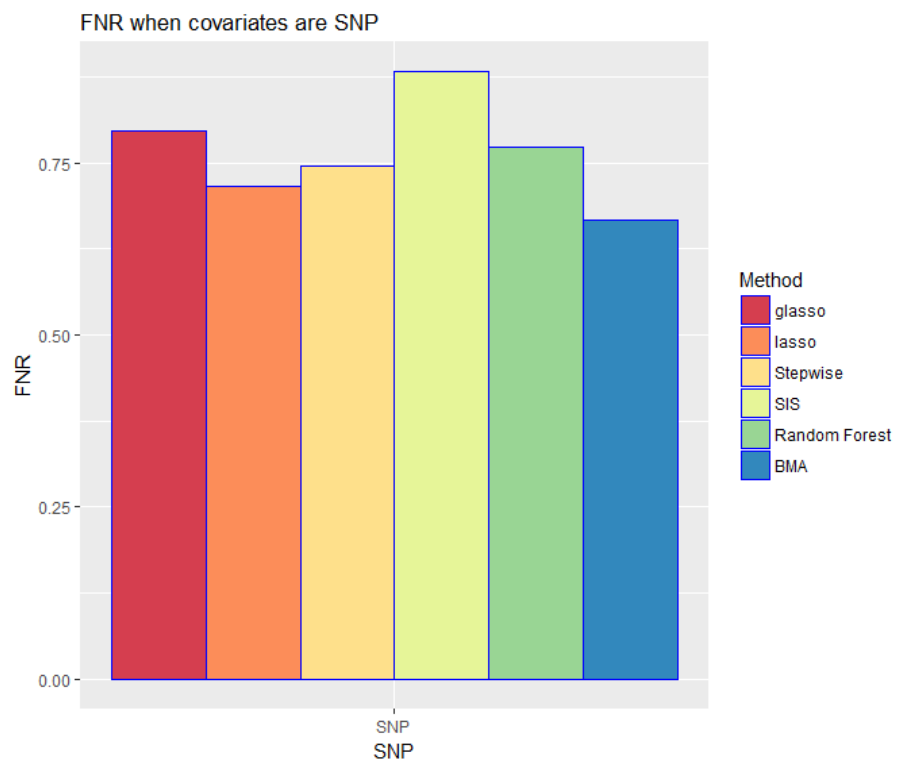


(c) Model Size of Predictive Biomarkers

Figure 7: performance of PEN with different numbers of biomarkers.



(a) Precision of Predictive Biomarkers Selection



(b) FNR of Predictive Biomarkers Selection