STAT 601 Final Project

Min Yang, Richard Yang, and Yezhou Li

June 10, 2019

1 Major findings

- We define 2 main response variables, Y_1 and $Y2$ choosing from 35 genes, where a gene named ***PBX1*** Y_1 and ***PCA1*** of those 35 genes to be Y_2.
- We use ***BIC stepwise*** procedure and ***LASSO*** to perform variable selection which choose around 60 covariates from 598 covatiates. Models chosen by these two methods have R^2 close to 0.99 and adjusted R^2 is close to R^2. Moreover, diagnose plots indicate that these models satisfy assumptions. We then use ***PLS*** to find those components who cumulatively explain 98% variance. Amount of components extracted varies from 5 to 8. PLS constructs model whose R^2 is close to 0.98 and diagnose plots satisfy assumptions.
- We conclude that it is difficult to find global optimizer of the first GMC method. The ***CG*** method of R function ***optim*** gives us a local optimizer where most b_is are of scale 10^{-2}, and only a few are of scale 10^{-1}. This procedure takes 3 hours to complete. Therefore we do not have enough time to choose proper λ_1, λ_2.

  The second GMC methods is even harder to find global optimizer. CG method gives us a local optimizer where most b_is are of scale 10^{-1}. This procedure is extremely time-consuming as well. Thus, it is challenging to determine the optimal λ.
- In the logistic regression, we use Y_1 as response variables, and find it show certain pattern which is convenient for us to binarize it. So we use that pattern to fit logistic regression model, and find that the coefficients in logistic model are a little smaller than linear regression model, which means less information should be explained by the convariates.

2 Procedures

We find that several genes(rows) in the data set have identical names(second column: A_Desc), while they are significantly different in terms of other column, indicating that every cell in this data set has at least two values for one single gene. In order to handle this absurd situation, we append "_2" to gene name whenever necessary.

2.1 Define the response variables

First, we regress one gene on the rest of 35 genes, choosing main response variables ($Y1$) by comparing models' R^2. In this case, all possible models have the same amount of covariates. Meanwhile AIC, BIC are monotone with respect to R^2 when covariate number is fixed. Therefore, it is sufficient to choose the best response variable by comparing R^2. Therefore, we conclude that our main response variable ($Y1$) is PBX1 with $R^2 = 0.9543$. And the best model is shown in the Table1.

Second, we calculate the principal components for all 35 variables and choose the first principal components as our response variable $Y2$ (namely, PCA1) whose proportion of variance is about 39.52% among all 35 principle compoents.

2.2 Linear regression models

We try all the methods and models learned from STAT 601 to deal with the data and choose three best regression models to be our final models: AIC forward selection, Lasso regression, and partial least squared regression (PLS).

2.2.1 BIC Stepwise

$$BIC = -2l(\hat{\beta}|y) + 2log(n)p$$

As a matter of fact, AIC tends to choose model much too complicated. In this case, we experiment on set e and Y1, finding that the AIC stepwise procedure does not stop until covariate number p exceeds n. Therefore, we conclude that it is unreasonable to perform variable selection according to AIC.

Instead of AIC, we use BIC to perform variable selection given that it tends to choose simpler model because sample size is above 200. We find that BIC choose models with 30 70 covariates, which is smaller than the ones of Lasso in following section. All models chosen by BIC have R^2 close to 0.99. Meanwhile, diagnose plots indicate that all models satisfy model assumptions.

2.2.2 Lasso Regression

$$\hat{\beta}^{lasso} = argmin\{\sum_{k=1}^{N}(Y_{ik} - \beta_0 - \sum_{j=1}^{p} x_{kj}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

We use R function ***glmnet*** to find the path of $\hat{\beta}^{lasso}$ against its L_1 norm and ***cv.glmnet*** to choose the best model by cross validation.

2.2.3 Partial Least Squares Regression (PLS)

PLS is the eigenvalue problem for $X'YY'X$ From the ratio of variance explained, we choose the number of principal components can explain $Y_i(i = 1, 2)$ more than 98%

2.3 GMC related methods

2.3.1 First method

$$\hat{\beta} = \arg \max_{b_i, i=1,\ldots,p} T_1 = \arg \max_{b_i, i=1,\ldots,p} \frac{Var(g(x))}{Var(g(x), Var(e)} - \lambda_1 |corr(g(x), e)| - \lambda_2 \sum_{i=1}^{p} |b_i|$$

We introduce lasso penalty to perform variable selection and regularization simultaneously. We also introduce correlation penalty to rule out models with large residuals. In theory, given λ_1 and λ_2, we should find that the global maximizer $\hat{\beta} = (\hat{b_1}, ..., \hat{b_p})$ only has several non-zero $\hat{b_i}$s. In practice, we use R function ***optim*** to find $\hat{\beta}$, which only guarantees that the returned *par* is a local maximizer. However, in this case T_1 has so many local maximizers that it is difficult to find the global maximizer $\hat{\beta}$. Moreover, given that p is around 600, it is unrealistic to perform grid search.

Plus as the order of polynomial k increases, we have to adjust the value of λ_i to make sure that the penalty term works well. and because of too many coefficients, we use the solver of CG to deal with the large scale optimization problem, which works well in our problem.

2.3.2 Second method

$$T_2 = GMC(Y|g(x)) - \sum_{i=1}^{p} \lambda |b_i|$$

Similarly, we use the function in given package to calculate the GMC and define our own function to add regulation term, and then optimize it to get coefficient $\hat{\beta}$, and use similar way to select variables according to the general value of b_i. In above procedure, We try several initial values to make sure that the algorithm isn't stuck in local optima.

2.4 Logistic regression models

For response variable Y1 We make a scatter plot for $Y1$ and find $Y1$ is clustering in two parts and the boundary is about $Y_i = 0.5$, so we convert $Y1$ to dichotomized observations according to its clustering.

$$Y_i = \begin{cases} 1, & Y_{i,j} > 0.5 \\ 0, & Otherwise \end{cases} \qquad i = 1, 2; j = 1, 2, ..., n_i$$

For response variable Y2 From the scatter plot of $Y2$, we find the $Y2$ is evenly distributed and there is not clustering phenomenon, which means the information entropy for binary response is maximized when $P(Y=1)=0.5$.

$$Y_i = \begin{cases} 1, & Y_{i,j} > Y_{i,(n/2)} \\ 0, & Otherwise \end{cases} \qquad i = 1,2; j = 1,2,...,n_i$$

Then we use the same three models chosen from the "Linear regression models" step to fit the logistic response variables and compare the result with the linear regression models—Can logistic regression explains the major deviation of our response variables and if the variables have similar significance pattern with OLS.

3 Analysis

3.1 Linear regression models

The three major ways learned from STAT601 to deal with the situation that the number of variable is larger than the number of observation is

- Subset selection among models(by AIC, BIC, and Mallow's CP);
- Shrinkage methods with some constrains on the parameters (Ridge Regression and Lasso Regression);
- Derived principal components of variables (PCR and PLS);

Therefore, our three models come from each of these three classes. All explicit models are in the last pages for tables and plots. There is not big difference from choosing AIC and BIC, but BIC tends to we choose a small model. Considering the large number of variables, we choose BIC to be the standard. And choosing forward selection instead of backward one is because we just want a small subset of the large variable range, so forward selection is more efficient.

As for the selection between Ridge regression and Lasso regression, we think with the constrain: $\sum_{j=1}^{p} |\beta_j| \leq s$, some of the β_j will decrease to zero so it will be more clear to find the related variables and express the model. So we represent Lasso regression model in our report and.

Last but not least, since the the PCR just considers the variance of X and PLS considers the covariance of X and Y, which mean it consider more information than PCR. Thus, we choose the PLS to fit our model.

3.2 GMC related methods

Just like the regular lasso regression, the first GMC method uses penalty of covariance and lasso to maximize T_1 in terms of b_i, but we find it tricky to find proper hyper-parameter

λ_j to perfectly shrink some of the coefficients to 0, and leave rest coefficients unchanged, so we just find the most intuitive way to handle it, that is to subjectively find some threshold for b_i and select corresponding features.

3.3 Logistic regression models

According to the distribution of Y_1 and Y_2, we know that we can use different way to binarize them, which can make it's easier to fit logistic regression model. The fitted parameters in logistic regression models are smaller than that of linear regression models in general, when we use them to fit the same dataset.

4 Future work

GMC is the most fundamental measurement based on the definition of regression, which can reflect the variance explained by algorithm. If we can find the true global optima, it would be the best way to select variables according to response variables, no matter what the regression function is.

However, the GMC function doesn't have very ideal properties because it's not a convex (concave) function and has lots of local optima, which makes it really easy to get stuck. Sometimes it's impossible for us to try every possible initial values and compare the results.

In conclusion, I think the future work should come up with some specific optimization or approximation methods for GMC model selection, which can guarantee both the computation efficiency and the accuracy. And then it would have its own advantage over all other algorithm.

5 Limitation

In our model, we don't have the outcome expected – lots of coefficients shrink to 0, that's partly because of the heavy computation cost and lack of advance devices. So we cannot try many values of hyper-parameter and initial value, and we just subjectively find some thresholds to select model.

6 Results and graph

6.1 Define the response variables

```
Call:
lm(formula = Y1 ~ TRIM58 + TNFSF4 + PSMB10 + COX6B1 + TUBA6 +
    FRAT1 + LEPR + HOXA9 + PF4 + RPS24 + PIM1 + BACE2 + LRIG1 +
    PCNX + BRAF, data = gene1.value)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82926 -0.17644 -0.02649  0.17278  1.10872

Call:
lm(formula = figure2[, 23] ~ ., data = figure2[, -23])

Residuals:
     Min       1Q   Median       3Q      Max
-1.13848 -0.22975 -0.02012  0.21964  1.12460
```

Figure 1: best model with response variable to be PBX1

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.655e-10  2.530e-02   0.000  1.00000
`102`        5.875e-02  5.434e-02   1.081  0.28115
`474`        3.902e-03  7.209e-02   0.054  0.95690
`506`        1.017e-01  8.076e-02   1.260  0.20951
`558`        9.246e-02  1.026e-01   0.901  0.36860
`990`        4.532e-01  9.708e-02   4.668 6.00e-06 ***
`1301`      -1.611e-01  1.135e-01  -1.419  0.15774
`1328`       3.995e-02  4.016e-02   0.995  0.32122
`1458`      -1.714e-02  6.456e-02  -0.266  0.79088
`1850`      -4.627e-02  2.984e-02  -1.550  0.12286
`1881`      -1.103e-01  8.790e-02  -1.255  0.21121
`2074`      -5.356e-03  1.057e-01  -0.051  0.95964
`2285`      -5.769e-02  7.194e-02  -0.802  0.42366
`2435`      -7.492e-02  4.563e-02  -1.642  0.10241
`2967`       2.970e-01  5.083e-02   5.844 2.42e-08 ***
`3087`      -1.320e-01  4.955e-02  -2.664  0.00844 **
`3550`      -4.670e-01  1.084e-01  -4.307 2.75e-05 ***
`3565`       2.590e-02  6.395e-02   0.405  0.68600
`4064`       5.445e-02  1.083e-01   0.503  0.61570
`5285`       1.981e-02  6.066e-02   0.327  0.74439
`5595`       4.674e-02  6.348e-02   0.736  0.46252
`5617`      -1.880e-02  4.098e-02  -0.459  0.64705
`5909`       3.478e-02  5.427e-02   0.641  0.52241
`6147`       4.420e-02  7.992e-02   0.553  0.58097
`6457`       1.324e-03  2.866e-02   0.046  0.96321
`6488`      -3.760e-02  1.324e-01  -0.284  0.77681
`6942`       2.252e-01  8.975e-02   2.510  0.01299 *
`6983`       4.779e-02  5.267e-02   0.907  0.36544
`7431`       1.713e-01  9.259e-02   1.850  0.06595 .
`7948`       5.039e-01  4.841e-02  10.409  < 2e-16 ***
`7995`       2.288e-02  7.687e-02   0.298  0.76631
`8166`      -1.207e-01  9.309e-02  -1.296  0.19653
`8333`       9.575e-02  5.624e-02   1.703  0.09042 .
`8567`      -8.034e-02  6.715e-02  -1.197  0.23310
`8870`       4.254e-02  4.886e-02   0.871  0.38511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3676 on 176 degrees of freedom
Multiple R-squared:  0.9543,	Adjusted R-squared:  0.9454
F-statistic:   108 on 34 and 176 DF,  p-value: < 2.2e-16
```

Figure 2: best model with response variable to be PBX1

6.2 linear regression models

6.2.1 BIC Forward

1) Dataset e

Response variable is Y1

```
Call:
lm(formula = Y1 ~ `2549` + `2967` + `2682` + `2561` + `2720` +
    `2824` + `2403` + `2594` + `2521` + `2424` + `2606` + `2489` +
    `2598` + `2658` + `2749` + `2809` + `2957` + `2615` + `2426` +
    `2969` + `2975` + `2510` + `2619` + `2689` + `2477` + `2647` +
    `2768` + `2986` + `2778` + `2918` + `2680` + `2631` + `2613` +
    `2707` + `2726` + `2400` + `2865` + `2928` + `2432` + `2740` +
    `2700` + `2651`, data = gene.e.Y1.value)
```

Figure 3: Summary table of BIC selection result 1

```
Residual standard error: 0.1718 on 168 degrees of freedom
Multiple R-squared:  0.9905,    Adjusted R-squared:  0.9881
F-statistic: 415.5 on 42 and 168 DF,  p-value: < 2.2e-16
```

Figure 4: Summary table of BIC selection result 2

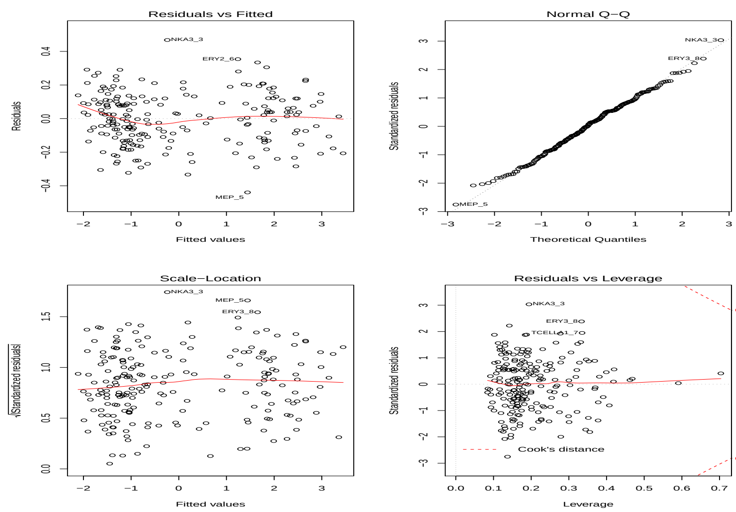

Figure 5: Diagnosis plot

Response variable is Y2

```
lm(formula = Y2 ~ `2621` + `2486` + `2785` + `2958` + `2407` +
    `2775` + `2896` + `2674` + `2943` + `2568` + `2655` + `2482` +
    `2592` + `2723` + `2758` + `2934` + `2682` + `2697` + `2765` +
    `2900` + `2545` + `2964` + `2837` + `2537` + `2869` + `2442` +
    `2576` + `2530` + `2656` + `2826` + `2740` + `2402` + `2550` +
    `2903` + `2975` + `2753` + `2969` + `2645` + `2478` + `2819` +
    `2774` + `2763` + `2549` + `2738` + `2616` + `2578` + `2547` +
    `2847` + `2959` + `2973` + `2712` + `2963` + `2786` + `2743` +
    `2779` + `2703` + `2666` + `2646` + `2814` + `2669`, data = gene.e.Y2.value)
```

Figure 6: Summary table of BIC selection result 1

```
Residual standard error: 0.2263 on 150 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9971
F-statistic:  1187 on 60 and 150 DF,  p-value: < 2.2e-16
```

Figure 7: Summary table of BIC selection result 2

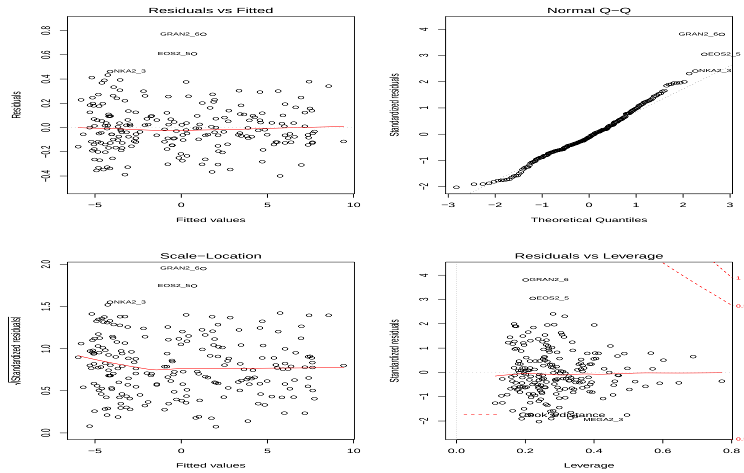

Figure 8: Diagnosis plot

2) dataset k
Response variable is Y1

```
lm(formula = Y1 ~ `5624` + `5786` + `5966` + `5755` + `5857` +
    `5757` + `5900` + `5388` + `5385` + `5753` + `5475` + `5861` +
    `5944` + `5416` + `5676` + `5711` + `5713` + `5476` + `5774` +
    `5844` + `5826` + `5912` + `5835` + `5733` + `5899` + `5978` +
    `5819` + `5911` + `5712` + `5797` + `5937` + `5427`, data = gene.f.Y1.value)
```

Figure 9: Summary table of BIC selection result 1

```
Residual standard error: 0.18 on 178 degrees of freedom
Multiple R-squared:  0.9889,    Adjusted R-squared:  0.9869
F-statistic: 496.1 on 32 and 178 DF,  p-value: < 2.2e-16
```

Figure 10: Summary table of BIC selection result 2

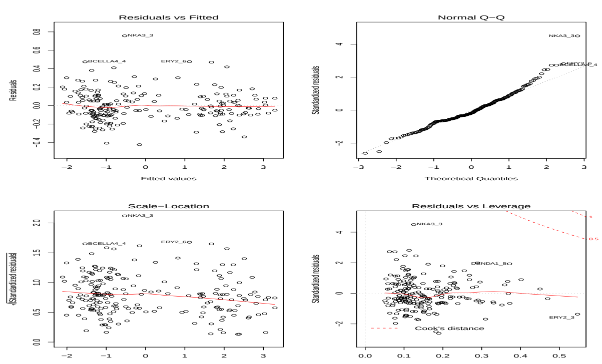

Figure 11: Diagnosis plot

Response variable is Y2

```
lm(formula = Y2 ~ `5786` + `5868` + `5879` + `5930` + `5448` +
    `5546` + `5877` + `5504` + `5833` + `5732` + `5782` + `5407` +
    `5452` + `5585` + `5395` + `5628` + `5518` + `5658` + `5735` +
    `5711` + `5808` + `5648` + `5603` + `5532` + `5684` + `5446` +
    `5733` + `5932` + `5408` + `5745` + `5766` + `5439` + `5654` +
    `5761` + `5872` + `5507` + `5561` + `5875` + `5899` + `5695` +
    `5640` + `5873` + `5884` + `5716` + `5861` + `5960` + `5927` +
    `5753` + `5689` + `5604` + `5483` + `5775` + `5424` + `5749` +
    `5597` + `5492` + `5923` + `5737` + `5686` + `5637` + `5838` +
    `5672` + `5823` + `5611` + `5568` + `5537` + `5382` + `5432` +
    `5460` + `5784` + `5613` + `5803` + `5724`, data = gene.f.Y2.value)
```

Figure 12: Summary table of BIC selection result 1

```
Residual standard error: 0.1951 on 137 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9978
F-statistic:  1314 on 73 and 137 DF,  p-value: < 2.2e-16
```

Figure 13: Summary table of BIC selection result 2

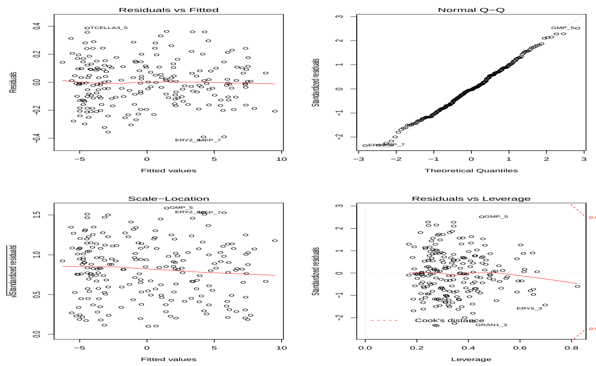

Figure 14: Diagnosis plot

6.2.2 Lasso

1) dataset e
Response variable is Y1

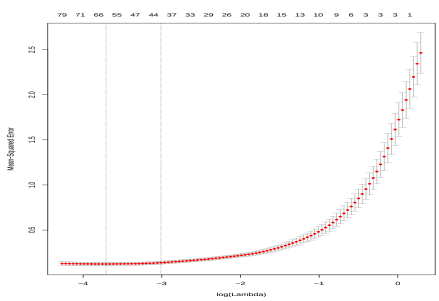

Figure 15: Choosing λ

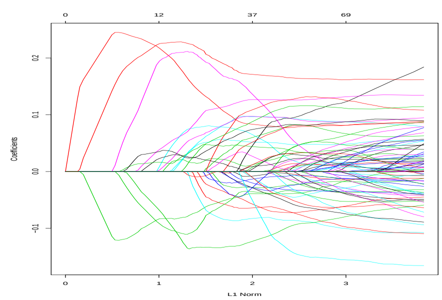

Figure 16: The trace of coefficients

```
Call:  glmnet(x = m1, y = Y1, alpha = 1, lambda = 0.2)

     Df   %Dev Lambda
[1,] 18 0.9092    0.2
```

Figure 17: Lasso effect

```
> lasso1$beta[,1][lasso1$beta[,1]!=0]
       HOXA9         PIM1          PF4         JUNB         FBN1     SERPINI1        TRIM58       PSMB10
 0.067330739  0.059635194  0.082487057 -0.149634851  0.010102138  0.025946537  0.157410330 -0.002085385
       IRAK4        ASH2L       MAN2A2       TNFSF4        NEDD9        MYO5A         LEPR        SCARF1
-0.089582746  0.013877360  0.086401232  0.097646630 -0.147338687 -0.102629924  0.135788718  0.022899049
         WRB        TUBA6
 0.075705354  0.066112194
```

Figure 18: Nonzero coefficients

Using the similar result, we can implement different scenarios, but we omit the selection procedure, and just show the results.

Response variable is Y2

```
Call:  glmnet(x = m1, y = Y2, alpha = 1, lambda = 0.4)

     Df   %Dev Lambda
[1,] 18 0.9576    0.4
```

Figure 19: Lasso effect

```
> lasso2$beta[,1][lasso2$beta[,1]!=0]
      HOXA9        JUNB        FBN1   LOC646278      TRIM58       PDGFC         LCK       ASH2L 
0.271563234 -0.061204292 0.127433495 0.055133424 0.283482920 0.505480403 -0.105817468 0.285484665 
     MAN2A2      TNFSF4       EMID1     GADD45A      SCARF1        LPXN       TUBA6       FAIM3 
0.006611466 0.037618742 0.241074532 0.029339048 0.291369361 -0.291839373 0.177305091 -0.027481499 
     LHFPL2        EREG 
0.070535125 0.061509270 
```

Figure 20: Nonzero coefficients

2) dataset k

Response variable is Y1

```
Call:  glmnet(x = m2, y = Y1, alpha = 1, lambda = 0.05) 

     Df %Dev Lambda
[1,] 33 0.97   0.05
```

Figure 21: Lasso effect

```
> lasso1$beta[,1][lasso1$beta[,1]!=0]
      NAP1L3       PRKCB1         PTRF         NT5M        CECR1        CRHBP         AATK 
 0.035990169 -0.020252309  0.024081643  0.163145899 -0.008614332  0.021345945  0.017649625 
        OAS2        CTBP2      RPL23A           XK       ZNF37B       FCGR2B       SFRS2B 
-0.029852893  0.092651876 -0.014876451  0.133242678 -0.028559296 -0.008581891  0.019012115 
     MGC3032         NADK       CTDSPL        ITSN2       MBOAT2        DYRK3        SPARC 
 0.025019108 -0.253549297  0.239700710 -0.047341144  0.119098349  0.055827486  0.028224382 
       ABCF2         USP6          MAX       ADAM10         IGKC DKFZp667M2411     SLC24A3 
-0.023251184  0.032213272  0.102968120 -0.006820790 -0.008622690  0.099512868  0.028697546 
        CPA3         SNCA        IKZF3        NPTX2        TIMP3 
 0.048340935  0.087014445 -0.030685210  0.043881668  0.143072215 
```

Figure 22: Nonzero coefficients

Response variable is Y2

```
Call:  glmnet(x = m2, y = Y2, alpha = 1, lambda = 0.05) 

     Df   %Dev Lambda
[1,] 63 0.9885   0.05
```

Figure 23: Lasso effect

```
> lasso2$beta[,1][lasso2$beta[,1]!=0]
        GPR65          CHD7          DOK4        PRKCB1          PTRF          NT5M         CRHBP 
-0.0213864533 -0.0297287889  0.0879534458 -0.2891786798  0.2140470058  0.0421275313  0.2795203942 
        FTHP1         ABHD5          PAF1         VIPR1        ZBTB24          PISD       SUV420H1 
 0.1462051624  0.0054663491 -0.0368288728 -0.0752674461 -0.0652946532  0.0459133752 -0.0834119149 
          F2R         PDE6B          PSPH       SLC43A3         NUTF2      EPM2AIP1         ROBO1 
 0.0491385629  0.1631676646  0.1172119269  0.1637242341  0.2123199115  0.1031911147  0.0053080524 
           XK         STK39          GNAQ        ZNF37B       SMARCA1        CRNKL1         GSTA3 
 0.0719407372 -0.0534172806  0.0261146439 -0.0156296987  0.0697199363  0.0025541874 -0.1989981329 
        RASA3       HISPPD1         CD247          DPP4         MRLC2       MGC3032       TMEM118 
-0.0303546412  0.0249804442 -0.0570761623 -0.1935481946 -0.0113751153  0.0487206069 -0.2239349741 
      ALDH8A1        CTDSPL        KIF21B        MBOAT2       NADSYN1         SPARC         CSRP1 
-0.0643408605  0.3435563046 -0.0061637667  0.4037213963 -0.0280537019  0.0520579228  0.0329746309 
        ABCF2          USP6        ZNF200        ARMCX6         GLRX5         KLF10          HHEX 
-0.1019664391  0.1407780535 -0.0300846446  0.0625955556  0.0607811392  0.0178775337  0.1057495209 
        CARD9        CXYorf3       SLC24A3         DAPK1          CPA3           CAT          PLP2 
 0.1250228539 -0.1006037067  0.1547914800  0.3009829077  0.1819196913  0.1047243987 -0.0448127639 
         BCL3          SNCA         IKZF3           TKT         NPTX2          DRAM         DDIT3 
-0.1355108860  0.0596075996 -0.0990749632  0.1923541029  0.0710216088  0.2969865066 -0.0005998264 
```

Figure 24: Nonzero coefficients

6.2.3 PLS

1) dataset e
Response variable is Y1

```
TRAINING: % variance explained
    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
X     20.15    39.31    46.53    49.69    59.50    62.35    69.70    71.51    72.67     73.80     75.19
Y1    85.59    90.49    93.58    96.31    96.68    97.46    97.66    98.21    98.59     98.97     99.18
    12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps  21 comps
X      76.11     76.91      77.8     78.35     78.98     79.58     80.22     80.93     81.37     81.83
Y1     99.36     99.50      99.6     99.71     99.78     99.84     99.88     99.90     99.93     99.95
    22 comps  23 comps  24 comps  25 comps  26 comps  27 comps  28 comps  29 comps  30 comps  31 comps
X      82.18     82.62     83.03     83.38     83.94     84.40     84.72     85.11     85.53      85.9
Y1     99.97     99.98     99.98     99.99     99.99     99.99    100.00    100.00    100.00     100.0
    32 comps  33 comps  34 comps  35 comps  36 comps  37 comps  38 comps  39 comps  40 comps  41 comps
X      86.28     86.54     86.74     86.96     87.19     87.46     87.67     87.91     88.09     88.36
Y1    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00
    42 comps  43 comps  44 comps  45 comps  46 comps  47 comps  48 comps  49 comps  50 comps  51 comps
X      88.63      88.8     88.97     89.17     89.35     89.58     89.76     89.91     90.07     90.23
Y1    100.00     100.0    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00
    52 comps  53 comps  54 comps  55 comps  56 comps  57 comps  58 comps  59 comps  60 comps  61 comps
X       90.4     90.55     90.71     90.88     91.01     91.19     91.37     91.51     91.65     91.76
Y1     100.0    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00    100.00
    62 comps  63 comps  64 comps  65 comps  66 comps  67 comps  68 comps  69 comps  70 comps  71 comps
X       91.9     92.06     92.18     92.31     92.43     92.55     92.68     92.79      92.9     92.99
Y1     100.0    100.00    100.00    100.00    100.00    100.00    100.00    100.00     100.0    100.00
```

Figure 25: variance explained(We extract 8 components

```
Call:
lm(formula = Y1 ~ enx1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.91900 -0.14107  0.00237  0.13130  1.15756 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.133e-09  1.544e-02   0.000   1.0000    
enx1Comp 1   1.863e-01  2.047e-03  91.004  < 2e-16 ***
enx1Comp 2   5.024e-02  1.706e-03  29.447  < 2e-16 ***
enx1Comp 3   6.716e-02  3.946e-03  17.022  < 2e-16 ***
enx1Comp 4   8.604e-02  5.122e-03  16.796  < 2e-16 ***
enx1Comp 5  -6.674e-03  3.523e-03  -1.894   0.0596 .  
enx1Comp 6   3.678e-02  6.190e-03   5.941 1.22e-08 ***
enx1Comp 7   1.954e-02  4.430e-03   4.411 1.67e-05 ***
enx1Comp 8   5.655e-02  5.912e-03   9.565  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2243 on 202 degrees of freedom
Multiple R-squared:  0.9805,	Adjusted R-squared:  0.9797 
F-statistic:  1267 on 8 and 202 DF,  p-value: < 2.2e-16
```

Figure 26: Summary table

Response variable is Y2 Similarly, we use the same method, and just show the results.

```
Call:
lm(formula = Y2 ~ enx1)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.37217 -0.32271 -0.02504  0.32800  1.70524 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.845e-15  3.638e-02   0.000  1.00000    
enx1Comp 1  5.356e-01  4.931e-03 108.607  < 2e-16 ***
enx1Comp 2  1.199e-01  4.014e-03  29.879  < 2e-16 ***
enx1Comp 3  9.214e-02  6.116e-03  15.067  < 2e-16 ***
enx1Comp 4  2.910e-02  9.194e-03   3.165  0.00179 ** 
enx1Comp 5  1.388e-01  1.136e-02  12.223  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5285 on 205 degrees of freedom
Multiple R-squared:  0.9843,	Adjusted R-squared:  0.984 
F-statistic:  2576 on 5 and 205 DF,  p-value: < 2.2e-16
```

Figure 27: Summary table

2) dataset k
Response variable is Y1

```
Call:
lm(formula = Y1 ~ enx1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.58138 -0.12945 -0.00489  0.11982  1.05656 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.133e-09  1.397e-02   0.000        1    
enx1Comp 1   1.876e-01  1.736e-03 108.062  < 2e-16 ***
enx1Comp 2   6.862e-02  1.714e-03  40.041  < 2e-16 ***
enx1Comp 3   6.820e-02  2.571e-03  26.529  < 2e-16 ***
enx1Comp 4   6.686e-02  3.570e-03  18.730  < 2e-16 ***
enx1Comp 5   3.476e-02  3.897e-03   8.921 2.74e-16 ***
enx1Comp 6   3.293e-02  3.573e-03   9.217  < 2e-16 ***
enx1Comp 7   4.814e-02  4.955e-03   9.715  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2029 on 203 degrees of freedom
Multiple R-squared:  0.9839,	Adjusted R-squared:  0.9834 
F-statistic:  1776 on 7 and 203 DF,  p-value: < 2.2e-16
```

Figure 28: Summary table

Response variable is Y2

```
Call:
lm(formula = Y2 ~ enx2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.34197 -0.36528 -0.05002  0.36660  2.25293 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.374e-16  4.253e-02   0.000        1    
enx2Comp 1  4.959e-01  5.326e-03  93.101  < 2e-16 ***
enx2Comp 2  1.485e-01  4.859e-03  30.560  < 2e-16 ***
enx2Comp 3  8.821e-02  6.171e-03  14.295  < 2e-16 ***
enx2Comp 4  1.242e-01  1.137e-02  10.928  < 2e-16 ***
enx2Comp 5  5.870e-02  8.835e-03   6.644  2.7e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6177 on 205 degrees of freedom
Multiple R-squared:  0.9786,	Adjusted R-squared:  0.9781 
F-statistic:  1875 on 5 and 205 DF,  p-value: < 2.2e-16
```

Figure 29: Summary table

6.3 GMC related models

Because the optimization is a little too time-consuming, we just use the first dataset(e) to show the general results.

6.3.1 First model

We choose the order of Polynomial is k=3, the $\lambda_1 = 0.5$, $\lambda_2 = 0.05$, and the target function attains value of approximately 0.76 when using random initial values from distribution $N(0.5, 0.1)$.

The following is our graph of coefficients.

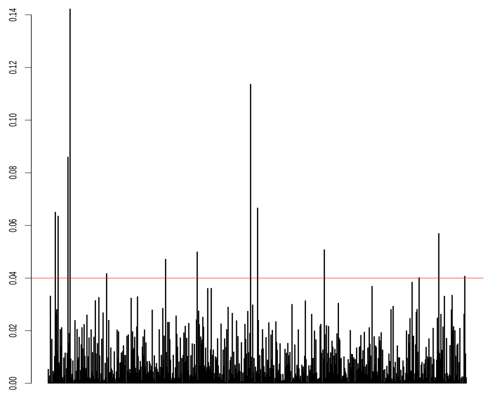

Figure 30: GMC coefficients using Y_1

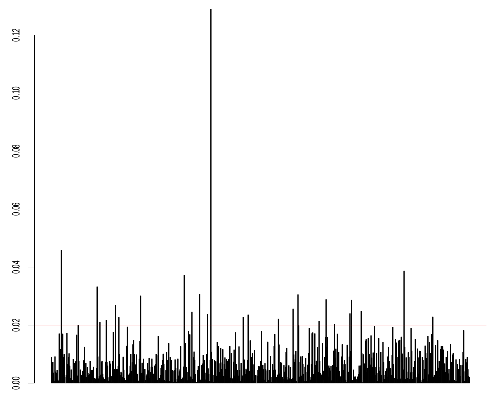

Figure 31: GMC coefficients using Y_2

6.3.2 Second model

And in second model, we found it's even harder to optimize, because the coefficients tend to remain unchanged in terms of their initial value.

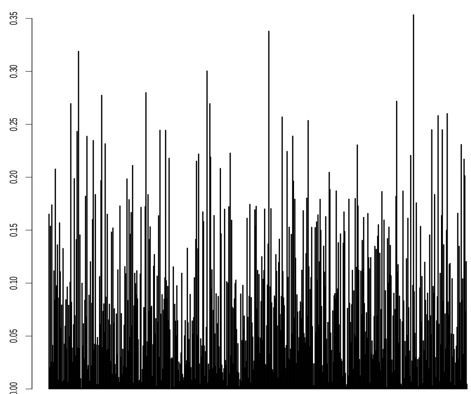

Figure 32: second coefficients using Y_2

And because of this problem, we cannot use it to implement variable selection.

6.4 Logistic regression models

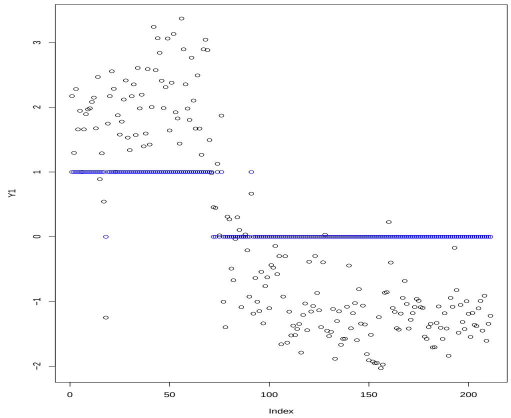

Figure 33: Changing Y_1 to dichotomized observations

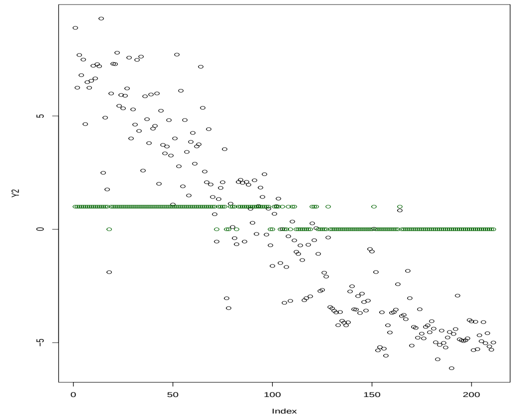

Figure 34: Changing Y_2 to dichotomized observations

And we just use the response variable Y1 and dataset e as an example to show our result due to the limited space

6.4.1 BIC Forward

```
Call:
lm(formula = Y1.logi ~ JUNB + TUBA6 + LRMP + TNFSF4 + GIMAP4 +
    PIM1 + FGFR1OP + HOXA9 + HERC5 + OSBP2 + SMARCD2 + COL8A2 +
    ASH2L + BCL2A1 + GORASP2 + NFYA + GNPDA1 + FLJ20054 + IRAK4,
    data = gene1.value)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36983 -0.07275  0.00183  0.07101  0.69761

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.345972   0.009232  37.476  < 2e-16 ***
JUNB        -0.004776   0.022012  -0.217 0.828464
TUBA6        0.104104   0.017082   6.094 5.95e-09 ***
LRMP        -0.101636   0.015157  -6.705 2.21e-10 ***
TNFSF4       0.068107   0.009602   7.093 2.49e-11 ***
GIMAP4      -0.052809   0.008010  -6.593 4.11e-10 ***
PIM1         0.038422   0.012527   3.067 0.002474 **
FGFR1OP     -0.077671   0.016233  -4.785 3.42e-06 ***
HOXA9        0.049789   0.011002   4.525 1.06e-05 ***
HERC5       -0.023679   0.014105  -1.679 0.094833 .
OSBP2       -0.114407   0.018456  -6.199 3.43e-09 ***
SMARCD2     -0.155253   0.025373  -6.119 5.23e-09 ***
COL8A2      -0.259685   0.042695  -6.082 6.33e-09 ***
ASH2L        0.059607   0.024078   2.476 0.014174 *
BCL2A1      -0.042167   0.010857  -3.884 0.000142 ***
GORASP2     -0.132798   0.025346  -5.240 4.24e-07 ***
NFYA         0.132060   0.036424   3.626 0.000370 ***
GNPDA1       0.088064   0.020927   4.208 3.96e-05 ***
FLJ20054     0.056713   0.017535   3.234 0.001437 **
IRAK4       -0.064075   0.024146  -2.654 0.008635 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1341 on 191 degrees of freedom
Multiple R-squared:  0.9281,    Adjusted R-squared:  0.9209
F-statistic: 129.7 on 19 and 191 DF,  p-value: < 2.2e-16
```

Figure 35: Summary table of BIC selection result

6.4.2 Lasso

```
> lasso1.logi$beta[,1][lasso1.logi$beta[,1]!=0]
       CSH1 DKFZp667G2110        HOXA9         PIM1          PF4         IER3         PPIF
-0.014293221  -0.011417970  0.039583530  0.011620442  0.009754263 -0.011959168  0.004986487
     FAM82B           GGH        ACOX3         LRMP         GAS7         CYCS       PSMB10
-0.017138514   0.013972542 -0.017721199 -0.033946290 -0.003475088  0.009858345 -0.029844694
     IFNAR2         XRCC4         CCNK        PDGFC        FRAT1        IRAK4        ASH2L
-0.014545809  -0.003438885 -0.031394962  0.007626280 -0.023292591 -0.019368537  0.045345089
     MAN2A2        TNFSF4         NRGN       COX6B1         MEG3        EMID1        CD79B
 0.020629093   0.043235806  0.002230937  0.014770639  0.029377006  0.002792838 -0.015187162
      BICD1         BCKDK        MYO5A       BCL2A1      SMARCD2       GIMAP4       DNAJC3
 0.004117169  -0.007811375 -0.002569587 -0.014879144 -0.007847312 -0.002881258 -0.011623256
     COL8A2         RNF11         RPA4        BACE2        TUBD1          WRB         LPXN
-0.050804220   0.024419333 -0.005619270  0.016556699 -0.008966316  0.016234660 -0.025647963
      TUBA6        PTGER3         EREG         UGCG
 0.032152997   0.011344612  0.015473601 -0.002368188
```

Figure 36: Nonezero coefficients

```
Call:  glmnet(x = m1, y = Y1.logi, alpha = 1, lambda = 0.02)

     Df   %Dev Lambda
[1,] 46 0.9039   0.02
```

Figure 37: Lasso effect

6.4.3 PLS

```
Call:
lm(formula = Y1.logi ~ enx1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26397 -0.06987 -0.00471  0.05731  0.50803 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.3459716  0.0072409  47.780  < 2e-16 ***
enx1Comp 1  0.0536991  0.0009663  55.570  < 2e-16 ***
enx1Comp 2  0.0134160  0.0008541  15.707  < 2e-16 ***
enx1Comp 3  0.0151414  0.0014949  10.129  < 2e-16 ***
enx1Comp 4  0.0380894  0.0026291  14.488  < 2e-16 ***
enx1Comp 5  0.0008561  0.0017584   0.487    0.627    
enx1Comp 6  0.0142776  0.0027304   5.229 4.24e-07 ***
enx1Comp 7  0.0085631  0.0017336   4.940 1.64e-06 ***
enx1Comp 8  0.0254012  0.0034893   7.280 7.29e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1052 on 202 degrees of freedom
Multiple R-squared:  0.9532,    Adjusted R-squared:  0.9513 
F-statistic: 514.2 on 8 and 202 DF,  p-value: < 2.2e-16
```

Figure 38: Lasso effect