고급바이오정보학

5강. Genetic Population Analysis : Penalized Logistic Analysis

출처 : 방송통신대학교 바이오정보·통계학과

목차

1) Genetic Association Studies

- 주요 목적

- 어떤 질병이나 특성에 관련된 원인 유전자 또는 유전자 지역들을 식별을 위해서..
  - 암, 백혈병, 비만 등…
  - 혈압, 콜레스테롤 수치, 지질단백질, 체질량 지수(BMI)..

- Genome-wide Association Study (GWAS) <- 최근 활발히 연구되는 분야로

- 질병 또는 특성과 관련된 일반적인 유전자 변이를 식별

- Single nucleotide polymorphisms(SNP)

- 일반적인 165개의 인간의 복잡한 질병과 특성은 1,200 개 이상의 유전자 변이와 연관됨.(Lander, 2011 Nature)

2) Type of Genetic Data

- 유전자/유전체 데이터 <= x변수 또는 feature
  - 단일 염기 다형성(SNP)
  - Microarray 유전자 발현 데이터
  - DNA methylation 데이터
  - 유전자 복제수 변이
  - 차세대 유전자 서열 데이터 (한사람의 유전자 전체 데이터)
- 표현형 데이터 <= y변수 또는 목표변수
  - 양적 특성
  - 사례-대조군 데이터
  - Matched case-control data : 한 사람의 암세포와 정상세포의 데이터
  - 생존 시간

3) Statistical Challenges

- High-dimensional genetic/genomic data

-고차원의 문제(변수 개수가 너무 많음)

-유전자(변수)수가 샘플수보다 많음.

-유전자들간의 상호 상관이 높음.

-소수의 유전자만이 표현형에 영향을 줌.(Sparisty of true singals)

- Statistical Analysis

-개별 유전자 Test 시 Multiple testing problem이 발생 => 8강, 9강에서 다룸.

-Regularization Procedures가 필요

  - Tuning parameter selection problem

1) Penalized Likelihood for Regression

- Penalized likelihood

$$Q_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + p_\lambda(\boldsymbol{\beta})$$

  - $l(\boldsymbol{\beta})$ is a log-likelihood
  - $\boldsymbol{\beta}$ is a parameter of interest (regression coefficients)
  - λ is a tuning parameter for sparsity
  - $p_\lambda(\boldsymbol{\beta})$ is a penalty function

- Data for regression model

- X : 유전자 데이터

- Y : 표현형

2) Linear Regression Model

- Analysis of quantitative outcomes : blood pressure, BMI, …
- The phenotype outcome follows a Normal distribution

$$Y_i \sim N(\beta_0 + X_i\ \beta, \sigma^2), \qquad i = 1, 2, \ldots, n$$

  - β_0 is intercept parameter
  - β is a vector of p-dimensional regression coefficients
  - σ^2 is a constant variance
  - $p_\lambda(\beta)$ is a penalty function
- Penalized likelihood using least squared loss

$$Q_\lambda(\beta_0, \beta) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \beta_0 - X_i\beta)^2 + p_\lambda(\beta)$$

3) Penalized Likelihood for Regression

- Analysis of binary outcomes : case-control outcomes

-Presence or absence of a disease

-$\boldsymbol{Yi} = \mathbf{1}$ for cases and $\boldsymbol{Yi} = \mathbf{0}$ for controls

- Penalized likelihood using logistic regression

$$Q_\lambda(\boldsymbol{\beta_0}, \boldsymbol{\beta}) = -\boldsymbol{l}(\boldsymbol{\beta_0}, \boldsymbol{\beta}) + \boldsymbol{p}_\lambda(\boldsymbol{\beta})$$

$$\boldsymbol{l}(\boldsymbol{\beta_0}, \boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} [\boldsymbol{Y_i} \log p_i(\boldsymbol{\beta_0}, \boldsymbol{\beta}) + (\mathbf{1} - \boldsymbol{Y_i} \log(1 - p_i(\boldsymbol{\beta_0}, \boldsymbol{\beta}))]$$

and

$$\boldsymbol{p}_i(\boldsymbol{\beta_0}, \boldsymbol{\beta}) = \frac{\boldsymbol{e}^{\boldsymbol{\beta_0} + \boldsymbol{X_i}\ \boldsymbol{\beta}}}{\mathbf{1} + \boldsymbol{e}^{\boldsymbol{\beta_0} + \boldsymbol{X_i}\ \boldsymbol{\beta}}}$$

4) Conditional Logistic Regression Model

- Analysis of matched binary outcomes : matched case-control outcomes

– Presence or absence of a disease for matched individuals

– $\boldsymbol{Yi} = \mathbf{1}$ for cases and $\boldsymbol{Yi} = \mathbf{0}$ for controls

– $\boldsymbol{\delta i} \in \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{K}$: the stratum of the $\boldsymbol{i}$-th individual

- Penalized likelihood using conditional logistic likelihood

- 한사람의 암세포 샘플은 1개 이고, 정상세포 샘플은 여러 개 일때

$$\boldsymbol{Q_\lambda(\beta)} = -\log \boldsymbol{L(\beta | n_1, \ldots, n_K)} + \boldsymbol{p_\lambda(\beta)}$$

$$\boldsymbol{L(\beta)} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{K}} \frac{\boldsymbol{exp}(\sum_{\boldsymbol{j \in \Delta_k}} \boldsymbol{X_i \beta Y_j})}{\sum_{\boldsymbol{j \in \Delta_k}} \boldsymbol{exp}(\boldsymbol{X_i \beta})}$$

with $\boldsymbol{\Delta_k} = \{\boldsymbol{j} \leq \boldsymbol{n}: \boldsymbol{\delta_j} = \boldsymbol{k}\}$ consists of $\boldsymbol{m} + \mathbf{1}$ indicies for $1:\boldsymbol{m}$ matched design.

5) Penalty Functions: Shrinkage Estimate

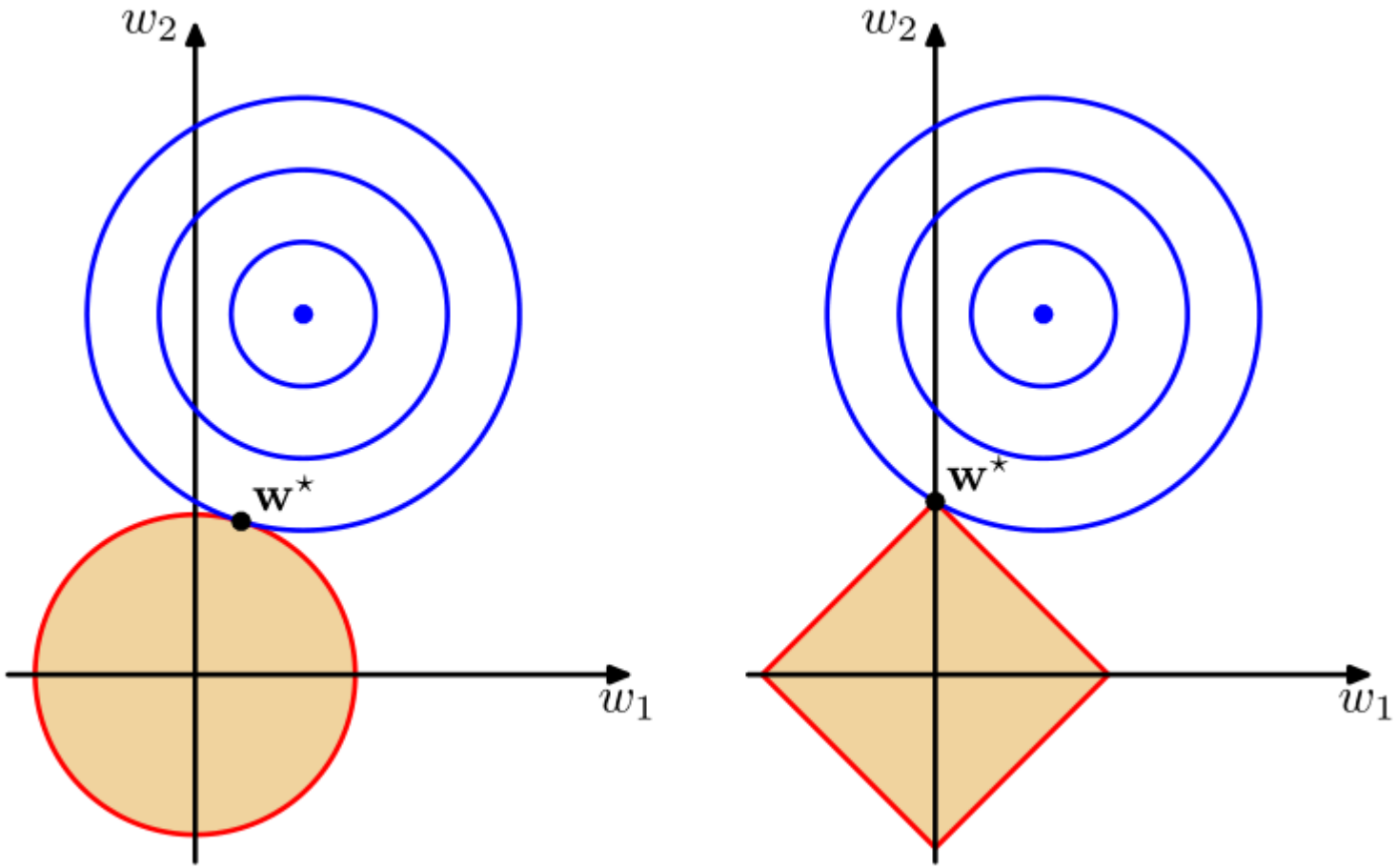
- Lasso (Least Absolute Shrinkage and Selection Operator)
  - l_1 norm penalty induces sparsity
  - Most of regression coefficients β are exactly to zero.

$$p_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$$

- Ridge regression estimate
  - l_2 norm penalty induces smoothness
  - Regression coefficients for correlated variables are shrinkged to each other

$$p_\lambda(\beta) = \lambda \sum_{j=1}^{p} \beta_j{}^2$$

5) Penalty Functions: Shrinkage Estimate



6) Elastic Net Regularization

- Elastic net regularization procedure
  - It combines l_1 norm penalty and l_2 norm penalty

$$p_\lambda(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2$$

  - l_1 part generate a sparse model
  - l_2 part encourages grouping effects for correlated variables
  - No limit on the number of selected variables
  - If $\lambda_1 = 0$, the elastic net reduces to ridge
  - If $\lambda_2 = 0$, the elastic net reduces to lasso

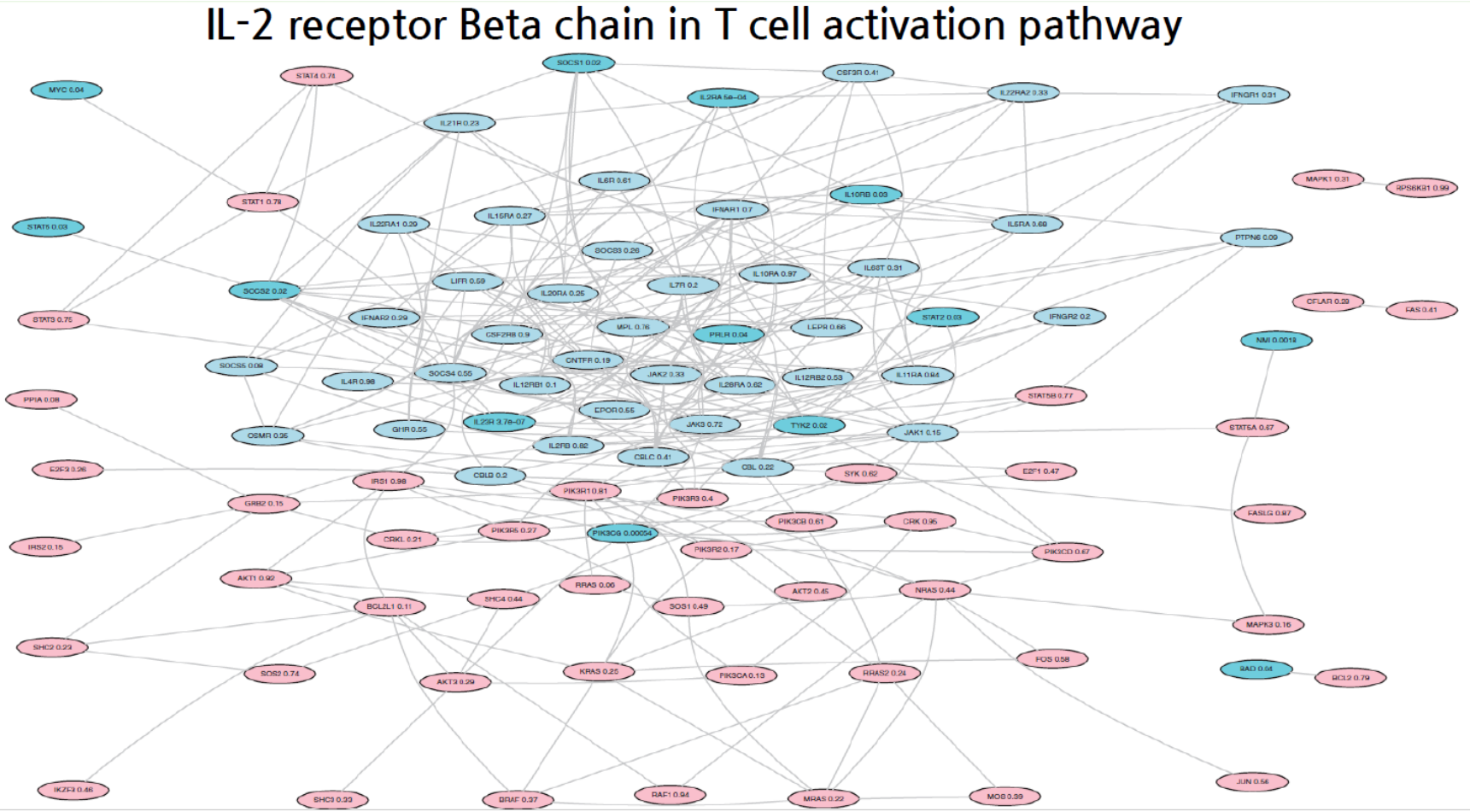
7) Network-based Regularization

- Elastic net regularization including a Laplacian matrix : 이미 알고 있는 유전자들간의 관계를 반영
  - A Laplacian matrix L represents a complex network structure of genes

$$p_\lambda(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \beta^T L \beta$$

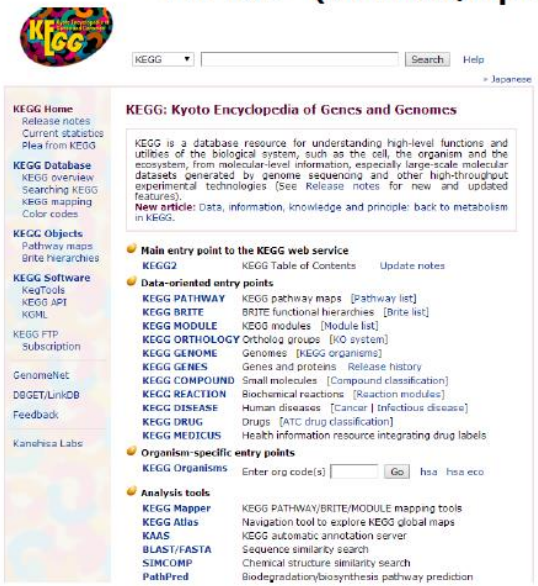
  - If $L = I$, it is exactly the same as elastic net
  - Smoothness w.r.t linked structure of the regression coefficients is induced.
  - Biological network information can be incorporated into regularization procedure.
    - Gene regulatory pathway
    - Metabolic pathway

8) Example of Genetic Pathway



9) Biological Network Database

- Biological network information is available at
  - KEGG (www.genome.jp/kegg/)
  - BioCarta (www.biocarta.com)
  - GenMAPP (www.genmapp.org)
  - HPRD (www.hprd.org)



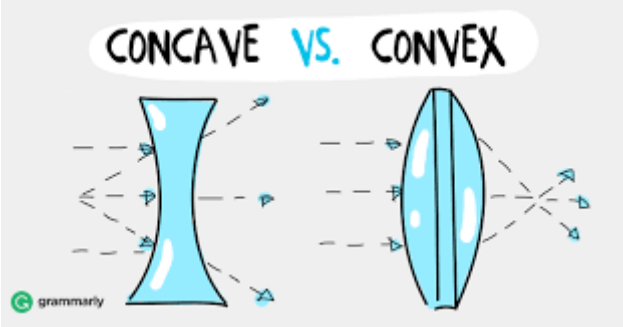
10) Convex Optimization

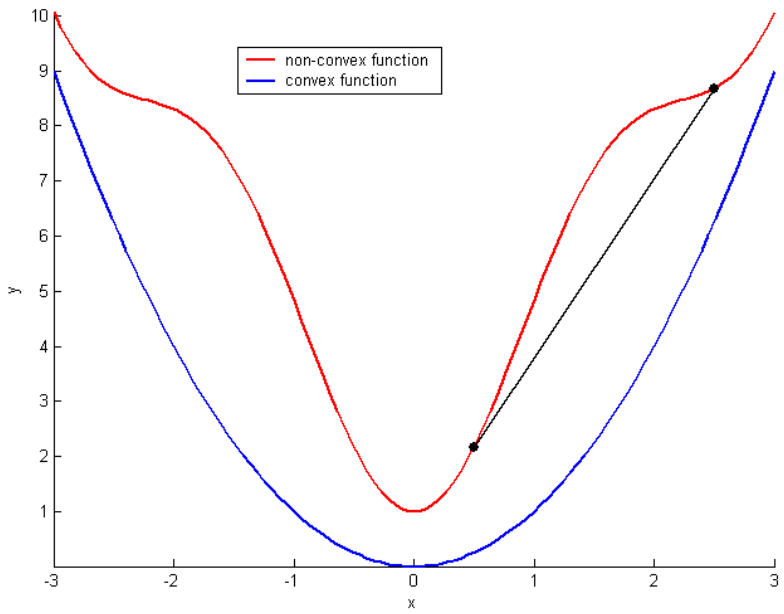
- Regression coefficients estimates
  - For given λ, we need to minimize the penalized likelihood

$$Q_\lambda(\beta) = -l(\beta) + p_\lambda(\beta)$$

  - The closed form solution to β does not exist.
  - The penalized likelihood is convex.
  - Computational algorithms for convex optimization can solve the equation.
  - Cyclic coordinate descent algorithm provide an efficient and fast solution for high-dimensional genomic data.
  - Statistical software such as an R package is publicly available.

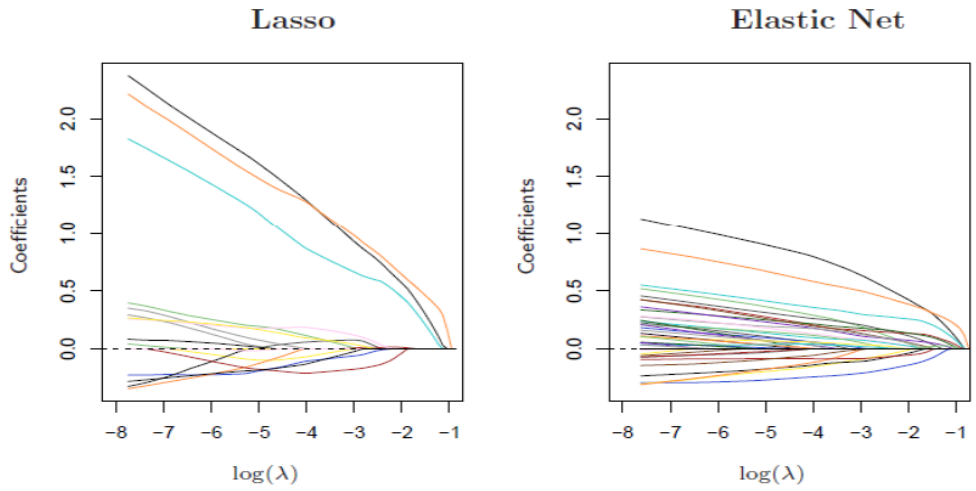
https://en.wikipedia.org/wiki/Coordinate_descent





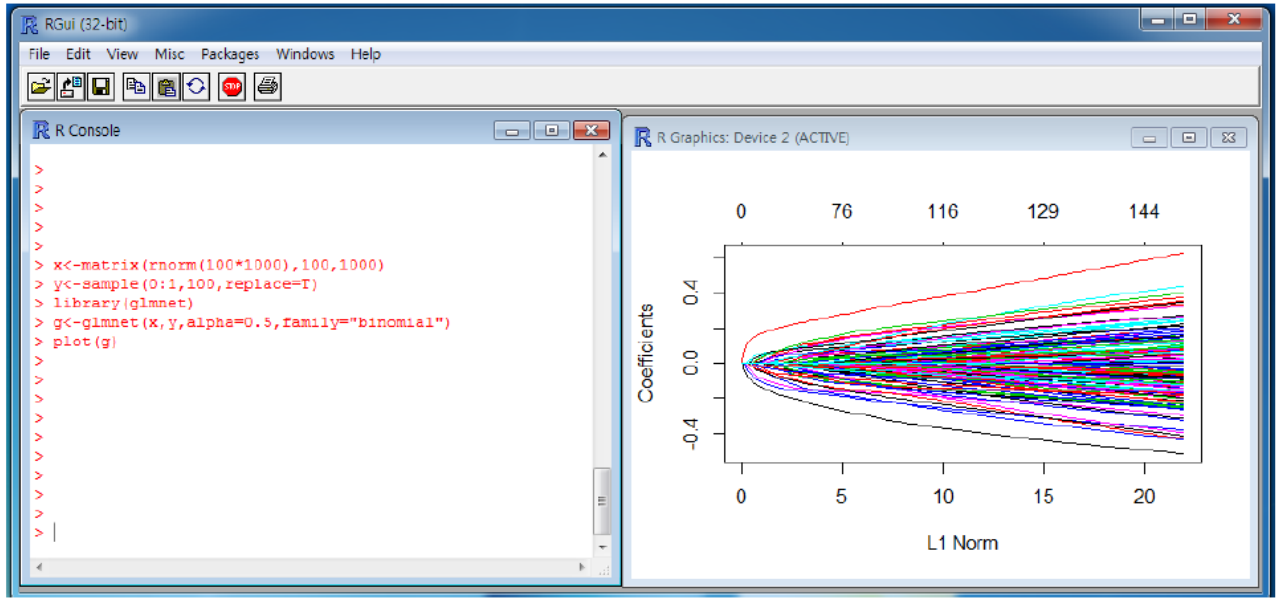
1) Publicly Available Statistical Software

- R packages for regularization procedures
  - 'glmnet' : elastic net solution with a regularization path
  - 'grpreg' : regularization with concave penalty functions
  - 'lars' : lasso solution with a regularization path
  - 'penalized' : elastic net and fused lasso solution
  - 'pclogit' : network-based penalized logistic regression



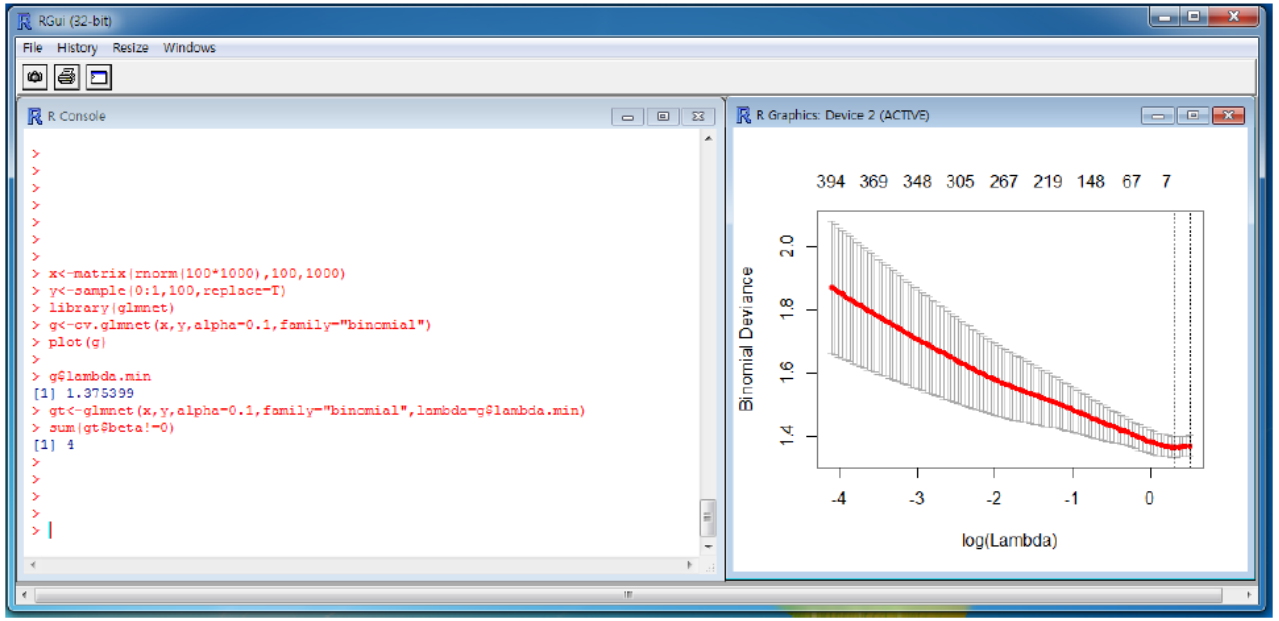
2) R package ‘glmnet’

- High-dimensional data analysis with ‘glmnet’
  - x : $n \times p$ matrix (genomic data)
  - y : n-dimensional vector (phenotype outcome)
  - alpha : controls smoothness between 0 (ridge) and 1 (lasso)
  - family : “binomial” for case-control data



3) Tuning Parameter Selection

- Cross validation
  - Data is separated by training set and test set
  - K-fold cross validation
    - For a fine grid of λ, deviance is compared.
    - Pick the best λ that minimizes device or other criterion.



4) Code

```
# install.packages("glmnet")
library(glmnet)

set.seed(101010)

x <- matrix( rnorm(100*1000), 100, 1000 )
y <- sample(0:1, 100, replace = T)
g <- glmnet(x, y, alpha = 0.5, family = "binomial")
plot( g )

g <- cv.glmnet(x, y, alpha = 0.1, family = "binomial")
plot( g )
g$lambda.min
gt <- glmnet(x, y, alpha = 0.1, family = "binomial", lambda = g$lambda.min )
sum(  gt$beta != 0 )
```