이학석사 학위논문

Practical issues for screening and variable selection method in a Genome-Wide Association Analysis

전장유전체 연관분석에서의 변수 선별과 변수 선택 방법의 현실적 사안들

2013 년 2 월

서울대학교 대학원

통계학과

홍 성 연

Practical issues for screening and variable selection method in a Genome-Wide Association Analysis

지도 교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함
2013 년 2 월

서울대학교 대학원
통계학과
홍 성 연

홍성연의 이학석사 학위논문을 인준함
2013 년 2 월

위 원 장 전 종 우 (인)

부위원장 박 태 성 (인)

위 원 Paik Myunghee Cho (인)

Abstract

Practical issues for screening and variable selection method in a Genome-Wide Association Analysis

Sung-Yeon Hong

Department of Statistics

The Graduate School

Seoul National University

Variable selection plays an important role in high dimensional statistical modeling analysis. Computational cost and estimation accuracy are two main concerns for statistical inference of high dimensional data. Recently, many high dimensional data have been generated in biomedical science such as microarray data and single nucleotide polymorphism (SNP) data. Especially, the genome-wide association studies (GWAS) which focus on identifying SNPs associated with a disease of interest, have produced ultra-high dimensional data.

Numerous methods have been proposed to handle GWAS data. Most statistical methods have adopted a two-stage approach: (1) pre-screening for dimensional reduction, (2) variable selection for identification of causal SNPs. The pre-screening step selects SNPs in terms of their p-values or absolute value of regression coefficients in single SNP analysis. Penalized regression such as Ridge, Lasso, adaptive Lasso and Elastic-net are commonly used for the variable selection step. In this paper, we investigate which combination of prescreening method and penalized regression performs best on continuous type response variable via real GWA data containing 327,872 SNPs from 8842 individuals.

주요어 : GWAS, KARE, penalized regression, variable selection

학번 : 2011-20256

Contents

Abstract i

Chapter 1. Introduction 1

1.1 Background 1

1.2 Overview 3

Chapter 2. Methods 5

2.1 Standardization 5

2.2 Pre-screening 6

2.3 Variable Selection 6

2.4 Ordering 10

Chapter 3. Analysis 11

3.1 KARE data 11

3.2 Pre-screening 12

3.3 Variable Selection 15

3.4 Comparison Study 19

Chapter 4. Discussion 21

Bibliography 23

초 록 26

List of Figures

[Figure 1] 14

[Figure 2] 17

[Figure 3] 18

Chapter 1. Introduction

1.1 Background

Recently, many high dimensional data have been generated in biomedical science such as microarray data and single nucleotide polymorphism (SNP) data. Especially, the genome-wide association studies (GWAS), which focus on identifying SNPs associated with a disease of interest, have produced ultra-high dimensional data. We call high dimensionality if $p = O(n^a) for\ some\ a > 0$, and ultra-high dimensionalilty if $\log p = O(n^a)\ for\ some\ a > 0$, for theoretical development. When the dimension p is high, we fatally run into the "curse of dimensionality". The convergence of any estimator to the true value of a smooth function defined on a space of high dimension is very slow. Variable selection plays an important role in high dimensional statistical modeling analysis. Computational cost and estimation accuracy are two main concerns for statistical inference of high dimensional data.

Many efficient approaches have been introduced to overcome these problems. One of them is multi-step strategy.[7,8,9] The first stage reduces the dimensionality p for significant predictor selection in ultra-high dimensional data. This stage is mentioned as a pre-screening. Pre-screening

stage finds the variables which are marginally associated with a response variable. This step reduces the dimension of ultra-high dimensional data and also makes joint analysis possible. Therefore the multi-step approach can solve the ultra-high dimensional problem data stated above. Now several predictor selection tools have been developed to implement the above idea for ultra-high dimensional linear models. Sure independence screening (SIS) is the representative method in prescreening step.[1] SIS ranks the predictor variable according to the marginal correlations. SIS uses absolute value of coefficients criterion. Other pre-screening method is described in Cho et al.[7] They have pre-screening step to select marginally associated response by using P-value criterion. We want to know which method is better on continuous type response variable. Although many pre-screening method are available, we do not know which methods performs better in predicting the quantitative phenotype.

We find the variables which jointly associated with response variable, among the remained variables after pre-screening step. When multiple predictor variable exist for a response variable, the joint identification would be a powerful method.[7]

One of the traditional approaches for joint identification is multiple linear/logistic regression method. But when we handle high dimensional data by using a traditional method, we suffer

from several problems. First, high dimension would make multiple linear regression method not clear. Also high dimension causes the computational complexity. Second, multiple linear regression is very sensitive about multicollinearity among SNPs. To overcome this problem, various penalized methods have been addressed. They can find jointly associated variables in high dimensional data. For example, Ridge, bridge, the least absolute shrinkage and selection operator (Lasso), adaptive Lasso, smoothly clipped absolute deviation (SCAD), and Elastic-net[10, 11, 12, 13, 14]. Especially Elastic-net method uses Ridge and Lasso penalty. For that reason, Elastic-net has advantages of both Ridge and Lasso. Elastic-net automatically select significant variables and efficiently resolves the problem caused by multicollinearity. Also iterative adaptive Lasso method is proposed[9] which retains the appealing property of rapid computation, even for ultra-high dimensional problem. This method yields a sparse solution by setting some parameters to 0. Hence, the predictor selection is achieved with the nonzero values.

1.2 Overview

Many methods have been suggested in regarding pre-screening and variable selection. However, we do not know which method perform best for continuous type response

variable. In this paper, we investigate which prescreening method and penalized regression performs best. In order to compare power of prescreening method and penalized regression, we use the Korea Association Resource (KARE) project data. Adjusted R squared is used as a measure of comparison.

Chapter 2. Methods

We formulate a multi-stage strategy for the identification of significant variables among enormous number of explanatory variable. Our strategy consist of three stages. At stage 1, we screen out variables which are weak correlated with the response variable via single-variable association tests. We select variables in terms of their P-values or absolute value of regression coefficients in single variable analysis. At Stage 2, we search for multiple-variable associations by using penalized multiple regression with Elastic-net, Ridge, Lasso and iterative adaptive Lasso method. At stage 3, in case of Elastic-net, we assess the jointly identified variables by using Bootstrap Selection Stability(BSS). BSS have been proposed for the empirical assessment of how consistently a variable is selected from the bootstrap samples.[7] In case of adaptive Lasso method, we assess the jointly identified variables by using effect size.

2.1 Standardization

Supposed that $\mathbf{y}_i$ for $i=1,\cdots,n$ is the response for the ith individual, and $\mathbf{x}_{ij}$ for $j=1,\cdots p$ is the predictors. We assume the predictors are standardized to have mean zero and standard

deviation one in order to keep generality.

$\mathrm{E}(\mathrm{x_{ij}}) = 0$, and $\mathrm{E}(\mathrm{x}_{ij}^2) = 1$, for $i = 1, \ldots, n$, $j = 1, \ldots p$

2.2 Pre-screening

We use the linear regression model in order to eliminate predictors which are weak correlated with the response variable for dimensionality reduction.

$$\mathrm{y_i} = \gamma_0 + \sum_{q=1}^{Q} \gamma_q z_{iq} + \beta_j x_{ij} + \epsilon_i$$

where z_q represent the adjustment variable. All variables are ranked in ascending order of P-values or ranked in descending order of absolute value from single variable analysis. According to the order, the top p variables showing the strongest marginal associations with the response variable are selected.

.

2.3 Variable Selection

Method 1. Penalized regression.

Multiple linear models are fitted for the selected top p variables after adjusting for the covariates.

$$y_i = \gamma_0 + \sum_{q=1}^{Q} \gamma_q z_{iq} + \sum_{j=1}^{P} \beta_j x_{ij} + \epsilon_i$$

We will find optimal solution by using Penalized regression such as Ridge, Lasso and Elastic-net. Penalized regression finds solution as follow

$$\hat{\beta} = \underset{(\beta,\gamma)}{argmin}[-L(\beta,\gamma) + \lambda P_\alpha(\beta)]$$

$$\text{where } P_\alpha(\beta) = \frac{1}{2}(1-\alpha) \parallel \beta \parallel_{l2}^2 + \alpha \parallel \beta \parallel_{l1} = \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right]$$

The amount of shrinkage is represented by parameter λ. We can find an optimal λ by using ten-fold cross-validation, which accomplishes mean squared error minimization. Ridge estimator(α =0) is shirinkage estimator of least squared estimator.[4] Ridge is biased estimator. But Ridge reduces estimation variance. Therefore Ridge reduce sum of mean square error. In case of high dimensionalilty, Ridge provides shirinkage estimation do can not perform variable selection. Lasso(α=1) have ℓ_1-norm penalty function.[3] Thus, Lasso predicts 0 as insignificant variable. Lasso automatically performs variable selection. The Elastic-net method includes Lasso and Ridge regression. In other words, each of them is a special case where α=1 or α=0. Therefore, Elastic-net has not only advantages of Ridge but also advantages of Lasso regularization method. Variables showing strong joint association with the response are automatically selected via Elastis-net method. The Elastic-net is regularized regression

method which overcomes the limitations of the Lasso and Ridge method. Elastic-net has an ability to perform grouped selection, such as highly correlated variables.

Method 2. Iterative adaptive Lasso

This method has the two-stage procedure. At the 1st stage, an independence learning is implemented in terms of ranking the magnitude of the marginal linear regression estimators. At the 2nd stage, a weighted least-squares type objective function is used to approximate a potential function. This allows us to further define a Penalized Weighted Least square(PWLS) model for moderate-scale selection.

Step1) We need first to predetermine a sparsity parameter size *d*. It is recommended to take *d=n/(logn)*. For each variable, the single-variable association with phenotype is examined using linear regression. The *j*th predictor is $\hat{\beta}_j^M$. The predictors are ranked in descending order of values. From the first predictor to k_1th predictor are considered as the set A_1, where $k_1 = [2d/3]$. k_1 is recommended to guarantee that it will take at least two iterations. Variables of set A_1 fit jointly linear regression. The predictor is $\hat{\beta}_{j\mathcal{M}}$. Then employ the Penalized weighted least square (PWLS) procedure.

$\hat{\beta} = argmin_{(\beta,\gamma)}\,[-L(\beta,\gamma) + \lambda P(\beta)]\,,\ P(\beta) = \sum_{j=1}^{p} w_j|\beta_j|$ to select a subset $\mathcal{M}_1$ of A_1.

Step2) For every $j \in \mathcal{M}_3^C = \{1, \ldots\ldots, p\} \backslash \mathcal{M}_3$, estimate $\hat{\beta}_j^M$

$$y_i = \beta_0 + \sum_{i=1}^{n} \left(X_{i(\mathcal{M}_1)} \beta_{\mathcal{M}_1} + X_{ij} \beta_j^M \right) + \epsilon_i$$

$$\text{where } w_j = \left| \hat{\beta}_{j\mathcal{M}} \right|^{-1}$$

After ordering $\left\{ \left| \hat{\beta}_j^M \right| : j \in \mathcal{M}_3^C \right\}$ pick up a set A_2 of indices of size $k_2 = d - |\mathcal{M}_1|$

Step3) Apply the PWLS procedure at $\{\mathcal{M}_1, \ A_2\}$. The nonzero elements of variable yields a new significant indices $\mathcal{M}_2$.

Step4) Iterate steps 2−3 until $|\mathcal{M}_l| \geq d \ or \ |\mathcal{M}_l| = \ |\mathcal{M}_{l-1}|$.

Step5) We obtain both predictor set M and estimated parameter vector.

The magnitude of the absolute values of the marginal linear regression estimators can preserve the non−sparsity information of the joint regression model. This procedure contains the sure screening property. This large−scale screening method can be regarded as an extension of the SIS procedure.[1] This method contains the sure screening method[2]. It retains the appealing property that it can be rapidly computed, even for ultra−high dimensional problem. PWLS yields a sparse solution by setting some parameters to 0. Therefore, the predictor selection is achieved with the nonzero values. Adaptive Lasso method can reduce bias.

.

2.4 Ordering

After we select significant predictor variables, we have to range them in order of importance. In case of penalized regression, we use Bootstrap Selection Stability(BSS). Joint selection of SNPs via Elastic-net is performed for the bootstrap samples. Bootstrapping is technique of resampling. Bootstrapping sample is random sample with replacement from the original dataset. Bootstrap sample size is equal to the original data set size. B bootstrap samples are generated. Bootstrap Selection Stability(BSS) is defined for ith variables as follow.

$$\mathrm{BSS_i} = \frac{1}{B}\sum_{b=1}^{B} I_i^b \text{ , where } I_i^b = \begin{cases} 1 & \textit{if replicated in } b^{th} \textit{ bootstrap sample} \\ 0 & \textit{otherwise} \end{cases}$$

BSS signify the frequency how many times each selected predictor variable is replicated in B bootstrap datasets. SNPs are ranked in descending order of BSS.

In case of adaptive Lasso method, selected significant predictor variables are ranked in descending order of effect size.

Chapter 3. Analysis

These methods are applied to a Korea Association Resource (KARE) dataset with 327,872 SNPs from 8842 individuals to investigate which combination of prescreening method and penalized regression performs best. First of all we perform quality control process; eliminate SNPs having missing value. We use adjusted R square of fitted linear regression model in order to compare between combinations of prescreening method and penalized regression performs.

3.1 KARE data

The Korea Association Resource (KARE) project started on 2007. Participants in this project were recruited from two community-based cohorts; rural Ansung cohort and urban Ansan cohorts, in Gyeonggi Province of South Korea. The number of people of the Ansung and Ansan cohort 5,018 and 5,020. The age range is from 40 to 69 years. More than 260 phenotype have been widely surveyed through physical examinations, epidemiological surveys and laboratory tests. Now we focus on height phenotype, because height is most highly descendible polygenic characteristic.[5]

Originally KARE data contain 500,568 SNPs. Before analysis, quality control processes are performed as formerly drawn in Cho et al.[6] and missing genotypes were imputed by using PLINK software and the JPT/CHB reference panel in HAPMAP.[7] After these process, finally we get data set with 327,872 SNPs from 8842 individuals.

3.2 Pre-screening

At this step, we use the linear regression model in order to perform single SNP analysis for 327,872 SNPs. This linear regression model includes adjustment variables such as gender, age and recruitment area (rural Ansung and urban Ansan).

$$\text{height}_i = \gamma_0 + \gamma_1 SEX_i + \gamma_2 AGE_i + \gamma_3 AREA_i + \beta_j SNP_{ij} + \epsilon_i$$

where i=1, 2,…, 8842 and 8842 is the number of individuals; j=1, 2, … , 327,872 and 327,872 is the number of SNPs. All SNPs are ranked in ascending order of P-values or ranked in descending order of absolute value of coefficients. from single variable analysis. We will use high 1000 ranked SNPs each criterion which are P-value and absolute value of coefficients. We compare the minor allele frequency and the number of missing value of the selected SNPs on each criterion. In case of P-value criterion, the number of rare variants, which has value of MAF less than 0.05, is 87. In case of absolute value of coefficients criterion, the number of rare variants is

991.(Fig. 1) We observe that the common variants tend to have large number of missing value.

Although we have imputation process, missing values still exist. We have to eliminate individuals who have at least one missing value. When data at least has one individual who has at least one missing value, penalized regression such as elastic-net and adaptive Lasso can not be performed. After eliminating process performed, the remaining number of individual is 4183 on P-value criterion, 7496 on absolute value of coefficients criterion. The number of overlaped individual each criterion is 3740. We have to use these overlaped individuals to compare each combination method. But the loss of individual is too many. This is improper data to compare each combination method. In order to reduce the loss of data, we eliminated the SNPs which have the number of missing value above 30. Then the number of remained SNPs is 944 on P-value criterion, and the number of remained SNPs is 984 on absolute value of coefficients criterion. After eliminating process performed, the remaining number of individual is 7481 for P-value criterion, 8164 for absolute value of coefficients criterion. The number of overlaped individual each criterion is 7061. From now on, we use these individuals.

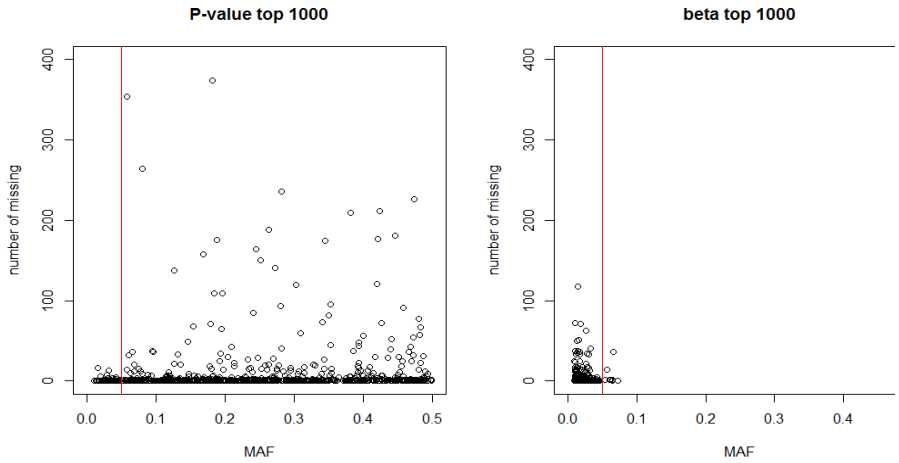

Figure 1. These plots represent the distribution of MAF and the number of missing value on each pre-screening method.

3.3 Variable Selection

We fit multiple linear regression model to jointly select associated SNPs for top 944, 984 SNPs, each criterion.

$$\text{height}_i = \gamma_0 + \gamma_1 SEX_i + \gamma_2 AGE_i + \gamma_3 AREA_i + \sum_{j=1}^{P} \beta_j SNP_{ij} + \epsilon_i$$

In case of Elastic-net regularization method, we set a tuning parameter α =1/2. We choose the value of tuning parameter λ=0.03602 on P-value criterion. We can choose the value of tuning parameter λ =0.1045 on absolute value of coefficients criterion. In case of Lasso regularization method, we set a tuning parameter α=1. We choose the value of tuning parameter λ=0.03792 on P-value criterion. We can choose the value of tuning parameter λ =0.1102 on absolute value of coefficients criterion. In case of Ridge regularization method, we set a tuning parameter α=0. We choose the value of tuning parameter λ=0.04376 on P-value criterion. We can choose the value of tuning parameter λ =0.1296 on absolute value of coefficients criterion. All tuning parameter λ were determined by ten-fold cross-validation, which minimized the mean squared error.

We can make eight combination, (P-value+Elastic-net), (P-value+Lasso), (P-value+Ridge), (P-value+iterative adaptive Lasso), (absolute value of coefficients+Elastic-net), (absolute value of coefficients+Lasso), (absolute value of

coefficients+Ridge), (absolute value of coefficients+iterative adaptive Lasso). After each combination method perform, each 524, 504, 944, 471, 549, 548, 984, 530 SNPs were jointly identified as putative height-related genetic variants.

Then, we generate 1000 bootstrapped sets which size is 7481, with replace by using elastic-net regularization method. The same fixed value of λ is used to generate bootstrapped data sets. Then we can find BSS value of each SNPs. SNPs are ranked in descending order of BSS. Ridge can not perform variable selection. Ridge selects all SNPs. So BSS is meaningless for Ridge. Therefore, in case of Ridge and adaptive Lasso method, SNPs are ranked in descending order of effect size.

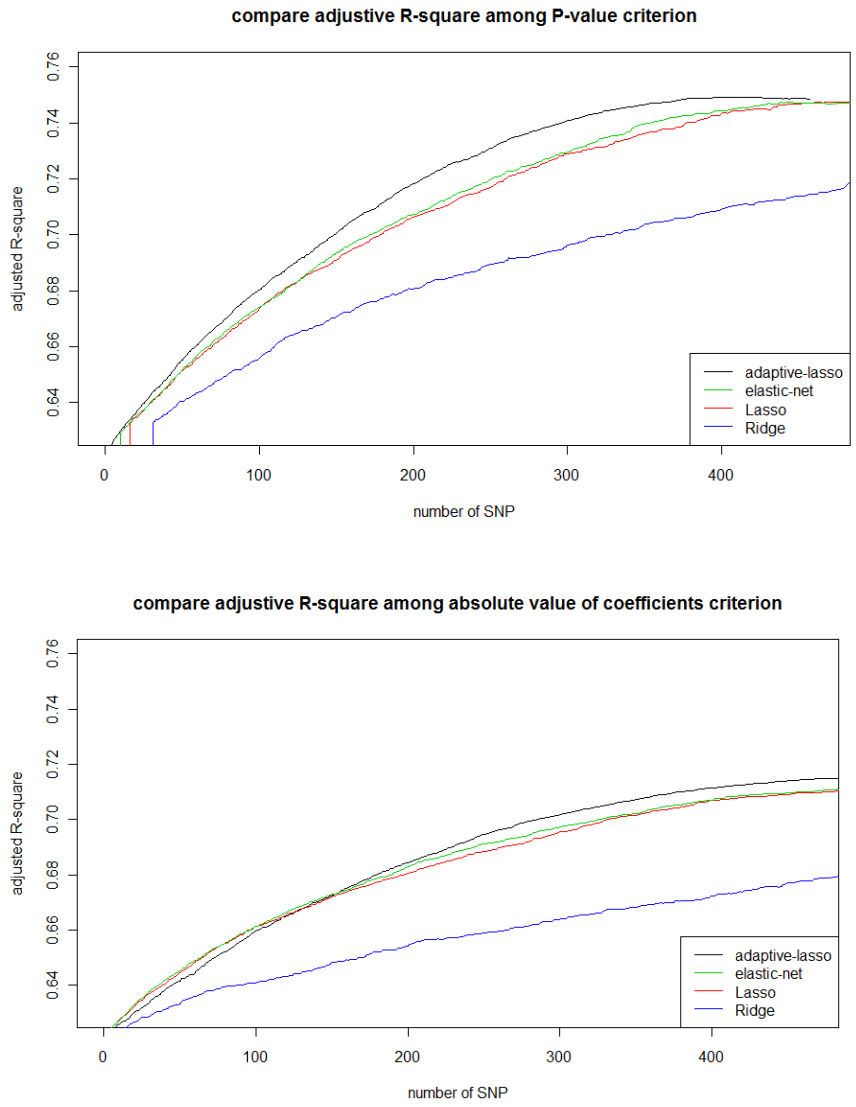

Figure 2. Adjusted R-square are plotted along the y-axis and the number of SNPs is plotted along x-axis on each prescreening criterion. The SNPs are ranked using BSS in case of Elastic net and Lasso. And the SNPs are ranked using effect size in case of iterative adaptive Lasso and Ridge. Explanatory power comparison among variable selection methods via adjusted R square.

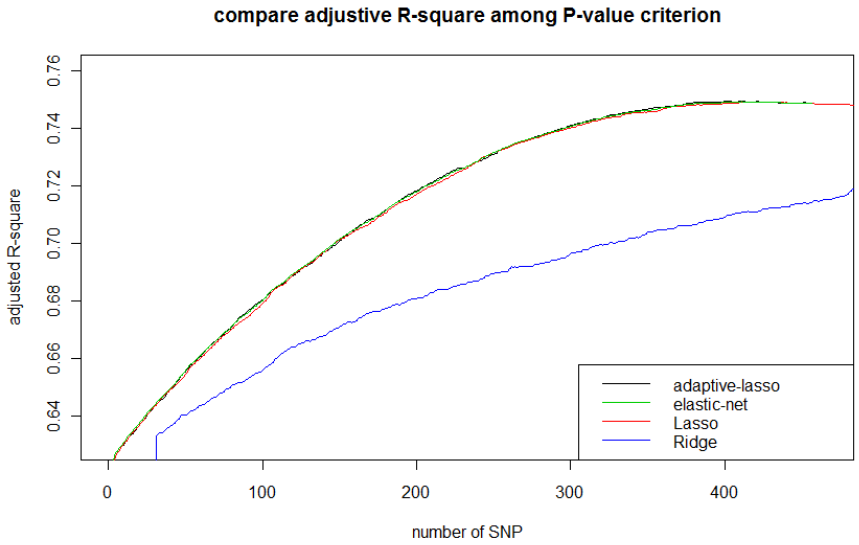

Figure 3. Adjusted R−square are plotted along the y−axis and the number of SNPs is plotted along x−axis on P−value criterion. The SNPs are ranked using only effect size. Explanatory power comparison among variable selection methods via adjusted R square.

3.4 Comparison Study

We calculated the adjusted R square for the selected SNPs to investigate which combination of prescreening method and penalized regression performs best in continuous type response variable. The SNPs are ranked using BSS in case of Elastic net and Lasso. The SNPs are ranked using effect size in case of iterative adaptive Lasso and Ridge. (Fig.2) There is a tendency that adjusted R-square increases as the number of SNPs increases. Fig.2 represents that the adjusted R square of P-value criterion method converges to 0.75, except Ridge. The adjusted R square of absolute value of coefficients criterion converge to 0.71, except Ridge. Therefore, we can conclude that P-value criterion selects more significant casual SNPs than absolute value of coefficients criterion method in case of continuous type response variable.

Ridge more slowly increases than iterative adaptive-Lasso, Lasso and Elastic-net on each pre-screening method. We conclude that Ridge can not perform variable selection, and Ridge is not proper method for high dimensional data in case of continuous type response variable.

Fig.2 looks like that Iterative adaptive-Lasso is rising faster than Elastic-net and Lasso. But Iterative adaptive-Lasso uses effect size order and Lasso, Elastic-net use BSS order.

They have different ordering. Fig.3 represents that the result of Elastic-net, Lasso, Ridge and iterative adaptive-Lasso use same ordering. Elastic-net, Lasso and iterative adaptive-Lasso are increasing equally.

Chapter 4. Discussion

Recently, many high dimensional data have been generated in biomedical science such as microarray data and single nucleotide polymorphism (SNP) data. Multi-step strategy have been introduced to analyze these data. The first stage is pre-screening. Pre-screening step selects marginally associated response variable by using various criterion. The second stage is variable selection. Various penalized methods have been addressed to analyze high dimensional data. For example, Ridge, bridge, the least absolute shrinkage and selection operator (Lasso), adaptive Lasso, smoothly clipped absolute deviation (SCAD), and Elastic-net. However, we do not know which method perform best for continuous type response variable. Adjusted R squared is used as a measure of comparison. In this study, we can conclude that P-value criterion selects more significant predictor variables than absolute value of coefficients criterion method in case of continuous type response variable. Also we can conclude that Elastic-net, Lasso and iterative adaptive-Lasso have same performance ability of variable selection on continuous type response variable, except Ridge.

In this study, we only use continuous type response

variable(height). We can use binary type response variable such as diabetes, high blood pressure.

Because of missing value, we unavoidably eliminate SNPs and individuals which have at least one missing value. This loss of information reduces the accuracy of study. We need to increase the accuracy of study by applying the proper imputation method by using simulated datasets.

.

Bibliography

[1] Fan, J.Q. and Lv,J.C. (2008) “Sure independence screening for ultrahigh dimensional feature space” Journal of the Royal Statistical Society, Series B 70 849−911

[2] Fan, J.Q. and Song, R.(2010) “Sure Independence Screening in Generalized Linear Models with NP−dimensionality”. Annals of Statistics 38 3567−3604

[3] Robert Tibshirani, “Regression Shrinkage and Selection via the Lasso” J. R. Statist. Soc., vol. 58, 1996, pp 267−288, doi:10.2307/2346178

[4] S. Le Cessi and J. C. Van Houwelingen, “Ridge Estimators in Logistic Regression” Appl. Statist., vol. 41, 1992. Pp 191−201, doi:10.2307/2347628

[5] Li, M., Liu, P., Li, Y., Qin, Y., Liu, Y. & Deng, H. (2004) A major gene model of adult height is suggested in Chinese. J Hum Genet 49, 148−153

[6] Cho, Y., Go, M., Kim, Y., Heo, J., Oh, J., Ban, H., Yoon, D., Lee, M., Kim, D. & Park, M. (2009) A large−scale genome−wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat Genet 41, 527−534

[7] Seoae Cho, Kyunga Kim, Young Jin Kim, Jong−Keuk Lee, Yoon Shin Cho, Jong−Young Lee, Bok−Ghee Han, Heebal Kim,

Jurg Ott and Taesung Park, "Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis," Ann. of Hum Genet., vol. 74, Sep. 2010, pp. 416-428, doi: 10.1111/j.1469-1809.2010.00597.x.

[8] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel and Kenneth Lange, "Genome-wide association analysis by Lasso penalized logistic regression," Bioinformatics, vol. 25, Mar. 2009, pp 714-721, doi:10.1093/bioinformatics/btp041.

[9] Pei-Rong Xu, Wen-Jiang Fu, Mo-Miao Xiong and Li-Xing Zhu, "GEE based screening and weighted least squares predictor selection for ultra-high dimensional generalized linear models in longitudinal data.",

[10] Hui Zou and and Trevor Hastie, "Regularization and variable selection via the elastic net," J. R. Statist. Soc., vol. 67, Apr. 2005, pp. 301-320, doi: .1111/j.1467868.2005.00503.x.

[11] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso," J. R. Statist. Soc., vol. 58, 1996, pp 267-288, doi:10.2307/2346178.

[12] S. Le Cessi and J. C. Van Houwelingen, "Ridge Estimators in Logistic Regression," Appl. Statist., vol. 41, 1992, pp 191-201, doi:10.2307/2347628.

[13] Jianqing Fan and Runze Li, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,"

JASA, vol. 96, Dec. 2001, pp 1348-1360, doi:10.1198/016214501753382273.

[14] Hui Zou, "The adaptive Lasso and its orachle properties," JASA, vol.

101, 2006, pp 1418-1429, doi:10.1198/016214506000000735.

[15] Seo-Jin Bang, Yong-Gang Kim, Taesung Park, "Joint selection of SNPs for improving prediction in Genome-wide Association Studies",

초 록

변수선택 방법은 고차원 자료의 통계적 분석에서 중요한 역할을 수행해 왔다. 고차원 자료의 통계적 추론 과정에서는 계산에 소요되는 비용과 추정 정확도가 가장 중요한 관심사이다. 최근에 분자생물학 실험(microarray) 자료와 단일염기다형성 (single nucleotide polymorphism: SNP) 자료와 같은 생물의학 자료에서 많은 양의 고차원 자료가 생성되고 있다. 특히 복합한 질환을 갖고 있는 사람과 정상인과의 차이를 나타내는 SNP을 발굴하는 전장유전체 연관분석 (genome-wide association study: GWAS)에서 초고차원의 데이터가 생성되고 있다.

이와 같은 전장유전체 연관분석 자료를 효과적으로 다루기 위해서 다양한 방법들이 제안되어 왔다. 그 중에서 다단계 접근 방법이 가장 대표적이다. 첫 번째 단계는 고차원 SNP 자료의 차원축소를 위해서 형질과 개별적으로 연관이 있는 SNP만을 선별하는 변수 선별 단계이다. 두 번째 단계는 전 단계에서 선별된 SNP들을 서로간에 공동적으로 고려하여 변수를 선택하는 변수선택 단계이다. 변수 선별 단계에서는 두 가지 기준으로 변수를 선별한다. 하나는 단일 SNP과 형질과의 연관성을 나타내는 회귀 계수의 유의확률이고 다른 하나는 단일 SNP과 형질과의 연관성을 나타내는 회귀계수의 절대값 크기이다. 다음으로 변수선택 단계에서는 계수추정 및 변수선택을 동시에 수행하는 Ridge, Lasso, adaptive Lasso 그리고 Elastic-net과 같은 벌점화를 통한 축소추정법이 사용된다.

본 논문에서는 연속형의 형질과 SNP과의 관계에서 다양한 변수선별 기준과, 변수선택의 다양한 벌점화를 통한 축소추정법에서 가장 효과적인 조합의 발굴에 대하여 고찰하였다.

연구를 위하여 The Korea Association Resource (KARE) project 자료에서 얻어진 8,842명의 표본과 327,872개의 SNP을 사용하였다.

주요어 : 전장유전체 연관분석; The Korea Association Resource (KARE) project; 변수선택

학 번 : 2011 – 20256