A COMPARATIVE STUDY ON VARIABLE SELECTION IN MICROARRAY CLASSIFICATION

SIRIKUL LAOSRIVICHIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEMMANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2014

Thesis
entitled

A COMPARATIVE STUDY ON VARIABLE SELECTION IN MICROARRAY CLASSIFICATION

…….....
MissSirikul Laosrivichit
Candidate

.....
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Major advisor

.....
Asst. Prof.Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Co-advisor

.....
Lect. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Co-advisor

…….....
Assoc. Prof. Sombat Thanawan, Ph.D,
Acting Dean
Faculty of Graduate Studies,
Mahidol University

…….
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Program Director
Master of Science Program in Technology of Information System Management
Faculty of Engineering
Mahidol University

Thesis
entitled

A COMPARATIVE STUDY ON VARIABLE SELECTION IN MICROARRAY CLASSIFICATION

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)
on
March 24, 2014

…….....
MissSirikul Laosrivichit
Candidate

.....
Asst. Prof. Adisorn Leelasantitham,
Ph.D. (Electrical and Computer Engineering)
Chair

…….....
Lect. Surapong Pongyupinpanich,
Ph.D. (Electrical and Informatic Engineering)
Member

…….
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Member

…….....
Lect. Waranyu Wongseree,
Ph.D. (Electrical Engineering)
Member

…….
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Member

…….....
Assoc. Prof. Sombat Thanawan, Ph.D,
Acting Dean
Faculty of Graduate Studies,
Mahidol University

…….
Lect. Worawit Israngkul,
M.S. (Technical Management)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

I would like to thanks to Dr. Waranyu Wongseree who always provides valuable advice and encouragementthroughout the research project. I am impressed by his kindness and dedication. He is my inspiration and motivation to work hard to overcome obstacles. I am much appreciated what he has done for me. This thesis was successful because of the support of him.

Moreover, I am grateful to Dr.Sotarat Thammaboosadee and Asst.Prof. Supaporn Kiattisin forproviding useful comments in this thesis.

Finally, I would like to thanks to my family and my friends for all their support and encouragement.

Sirikul Laosrivichit

A COMPARATIVE STUDY ON VARIABLE SELECTION IN MICROARRAY CLASSIFICATION

SIRIKUL LAOSRIVICHIT 5537223 EGTI/M

M.Sc. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THESIS ADVISORY COMMITTEE: SOTARAT THAMMABOOSADEE, Ph.D., SUPAPORNKIATTISIN, Ph.D., WARANYUWONGSEREE, Ph.D.

ABSTRACT

This thesis proposes usingthe shrinkage method for logistic regression. The normal exponential gamma (NEG) distribution is a Bayesian-inspire method. This method uses the normal-exponential-gamma prior for coefficients of model. The typical problems of microarray data have been solved by the LASSO and the elastic net. The comparative study of the two famous methods, NEG distribution and double exponential (DE) distribution were measured by the number of selected variables, predictive accuracy, deviance and computational time. The paired sample t-test was used to analyze the mean difference between two values. The results showed that the NEG distribution was more efficient the LASSO, the elastic net and the DE distribution.

KEY WORDS: NORMAL EXPONENTIAL GAMMA DISTRIBUTION / LASSO / ELASTIC NET /DOUBLE EXPONENTIAL DISTRIBUTION/ DNA MICROARRAY

74pages

การศึกษาเปรียบเทียบการคัดเลือกตัวแปรในการคัดแยกข้อมูลชนิดไมโครอะเรย์
A COMPARATIVE STUDY ON VARIABLE SELECTION IN MICROARRAY CLASSIFICATION

สิริกุล เหล่าศรีวิจิตร 5537223 EGTI/M

วท.ม. (เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: โษฑศ์รัตตธรรมบุษดี, Ph.D., สุภาภรณ์ เกียรติสิน, Ph.D., วรัญญู วงษ์เสรี, Ph.D.

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการลดขนาดสัมประสิทธิ์ของแบบจำลองการถดถอยโลจิสติกโดยอิงแนวคิดจากทฤษฎีของเบย์และใช้การแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติเป็นความน่าจะเป็นก่อนหน้าของสัมประสิทธิ์ของแบบจำลอง ในปัญหาการศึกษาดีเอ็นเอไมโครอะเรย์ การศึกษาเปรียบเทียบประสิทธิภาพวิธีที่นำเสนอกับวิธีที่นิยมใช้กันมากในปัจจุบัน คือวิธีแลซโซ วิธีข่ายยืดหยุ่น และวิธีที่มีการอิงแนวคิดจากทฤษฎีของเบย์และใช้การแจกแจงแบบเลขชี้กำลังคู่ ในการศึกษาประสิทธิภาพจะใช้จำนวนยีนที่ถูกเลือก ค่าความถูกต้องของการทำนาย ค่าความผิดพลาดของการทำนาย และเวลาที่ใช้ในการสร้างแบบจำลองการจำแนก และใช้หลักการทางสถิตินำมาวิเคราะห์หาความแตกต่างของค่าเฉลี่ยสองค่า (paired t-test) จากผลการศึกษาพบว่าการใช้การแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติมีประสิทธิภาพในการทำนายและความสามารถในการคัดเลือกยีนสูงกว่าทั้งวิธีแลซโซ วิธีข่ายยืดหยุ่น และการใช้การแจกแจงแบบเลขชี้กำลังคู่

74หน้า

CONTENTS

LIST OF TABLES

LIST OF TABLES(cont.)

LIST OF TABLES(cont.)

LIST OF FIGURES

CHAPTER I
INTRODUCTION

1.1 Background and Problem Statement

The human genome is a set of chromosomes which contain all genetic information. It is found in cell's nucleus. A number of chromosomes are different in each organism but similar in same species. There are 46 chromosomes in human. They are arranged into 23 pairs by getting one chromosome from mother and the other from father. Characteristics of whosehave inherited come from dominant feature, either father or mother. Each chromosome is the address of DNA. The DNA ,deoxyribonucleic acid, consist of deoxyribose, which is sugar unit, link to phosphate groups and a nitrogenous base. Two strands of DNA are held by hydrogen bond of complementary bases and then form double helix shape of DNA. The sequence of bases determines protein synthesis in organism. A main component of human's body is protein. A smallest unit of protein is amino acid which has more than 20 kinds. Each type of amino acid is built by determination of different bases sequence in DNA. The order of nitrogenous bases in DNA of a human is similar to other human about 99% but differences less than 1 % are enough to tell different characteristic of each human such as color eye, fingerprint. Most genetic diseases are caused by a mutation of gene which changes the base sequence in DNA.

DNA Microarray technology measures gene expression level of thousands gene in one experiment. Gene expression was studied to relation with phenotype. Normal and abnormal cells have different expression level. Gene expression data is usually classified between normal and disease, or categorized into subgroups by building a predictive model from a previous data. The typical problems in analysis are a number of variables much more than a number of samples and highly correlate between variables. The "large p, small n" problem causes overfitting problem, which is a low bias and a large various model. The high correlation of variables problem causes collinearlity problem, which is an unreliable model.

Classification is often done by logistic regression but it is unsatisfied with this data type. There are two reasons for this problem. Firstly, predictive accuracy problem causes from overfitting. Second, model interpretability causes from a large number of variables. A technique for improving is a model selection. There are many kinds of the model selection technique. Subset selection is one of the techniques which can solve the model interpretability problem but this technique is a discrete process. If data have a little change, the model will have many changes. It may decreases performance of the model. Dimension reduction, is another model selection technique, can reduce the number of variables but it cannot define the causal variables. Another technique of model selection is shrinkage technique. It is an attractive method because it selects important variables by shrinking coefficients which are not important toward zero.

Recently, the standard techniques for model selection in genetic studies are LASSO and elastic net which shrink coefficient toward zero and set others to zero by constrain optimization. These techniques had been analyzed with DNA microarray data and other genetic data such as Single nucleotide polymorphism (SNP). Although, these two techniques are popular in nowadays, there is another interested alternative technique. The normal exponential gamma had been compared performance with ridge, LASSO and elastic net and found that overall performance of NEG distribution is the best. The previous study is SNP selection in genome-wide association study. Given the current interest in the NEG distribution, this study will evaluate the performance of the NEG distribution, LASSO and elastic net in DNA microarray classification.

1.2 Objectives

To evaluate the performance of the normal exponential gamma(NEG) distribution, the double exponential (DE)distribution, LASSO and elastic net on microarray classification.

1.3 Scope of Work

This research is performance comparison of variable selection in DNA microarray binary classification between NEG distribution, LASSO, elastic net and DE distribution. The real gene expression level data were used in this research. Three datasets were leukemia microarray, lung cancer microarray and prostate cancer microarray. The performance was measured by predictive accuracy, deviance, the number of selected variable and computational time.

1.4 Result

The normal exponential gamma (NEG) distribution had performance better than LASSO,elastic net and DE distribution in selecting variables which are related to disease.

CHAPTER II
LITERATURE REVIEW

Over the past few decades, medical technology has been rapidly developed. Many people can overcome serious diseases. Scientists do not only find how to treat, but also find how to prevent the disease. Most common diseases, such as heart disease, cancer, diabetes, are multifactorial diseases which have multiple causes. One of the most common causes is a genetic change which might develop the disease. Microarray technology measures genes expression level that different types of cells may show different levels. There are a lot of genes expression levels that were measured in an experiment. It induced problems in data analysis process because there are many variables and data has collinearlity. Many methods for genetic data analysis were developed.

Literature review in this chapter consists of five parts. (1) Genetics (2) Microarray technology (3) Model selection (4) Shrinkage technique

2.1 Genetics

Genetics is the study about of heredity, including genes structure and function to cells. A main functional unit of heredity is gene. Genes consist of DNA which is instruction to determine a molecule synthesis. It is found that the sequence of DNA bases involve to genetic variation of organism. Human genes are estimated about 20,000 to 25,000 genes. One person is inherited a copy of gene from parent. A half of copy gets from father and the other get from mother. Most genes of a person are similar to others. There are about one percent of total genes that are different in each person.

Human body is composed of millions of cells. These cells are basic structure of organs of body. Cells contain various organelles, including the hereditary material. DNA, deoxyribonucleic acid, is the hereditary material which is located in

cell nucleus. DNA is composed of nucleotides which have four different nitrogenous bases. The four bases are adenine (A), guanine (G), cytosine(C) and thymine(T). The hereditary information of DNA is collected as if a code. The different sequence of nitrogenous bases affect to formation of the different genetic code. DNA has double helix structure which is generated from two stands of nucleotides.

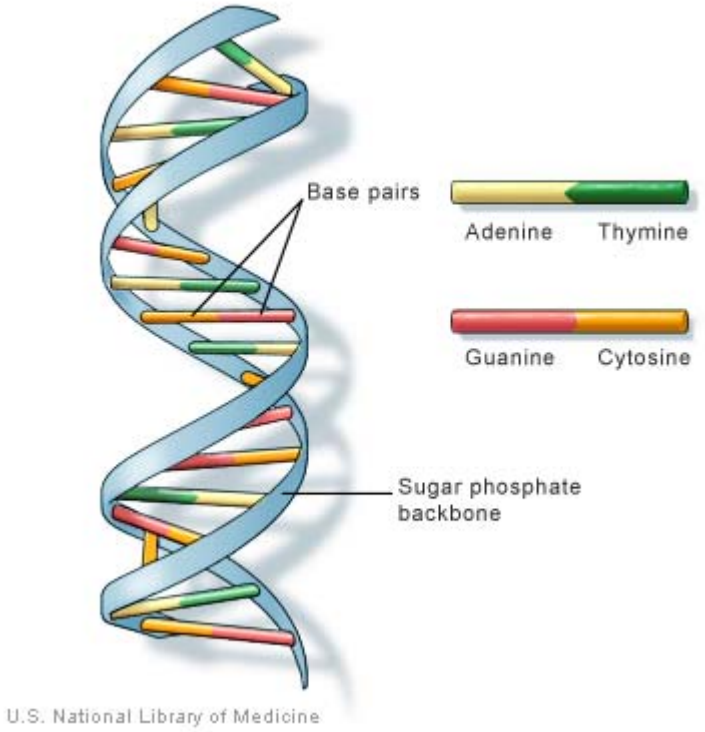


Figure 2.1DNA structure

Chromosome is a packed DNA. The DNA tightly coiled many times and packed into chromosome. There are forty-six chromosomes in human. Forty-four chromosomes called autosomes which are the same both male and female. The other one pair chromosome is sex chromosome.

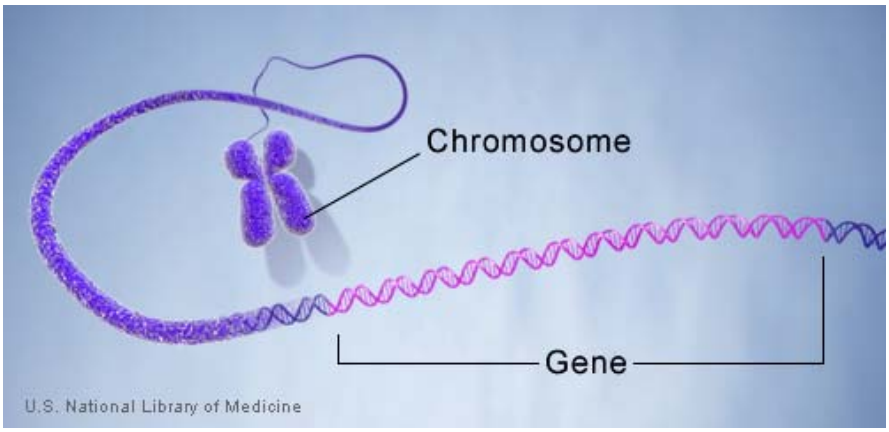


Figure 2.2 Chromosome structure

The main components of human body are proteins which play a critical role for human living. DNA sequences control the phenotypes through protein. The DNA sequence specifies amino acids type, which are the smallest units of protein and involves to the structure and function of protein. Genes regulate expression of cells in the body. If genes become abnormal genes, they will involve cell function.

The most common cause of changing of gene sequences is mutation. The gene mutation is an irreversible change of base sequences within gene. The mutation may be inherited by parent or acquired later. The inherited mutation affects to every cell of person's body and may make person be genetic disorder. The acquired mutation affects to some cells and some person's lifetime. The cause of acquired mutation has not surely known. It is believed that may cause from environmental factor, such as chemical agent, radiation or may cause from mistake in DNA copy process. In normal situation, genes determine protein synthesis so that human body can normally function. If mutation occurs, abnormal genes may instruct malfunction in protein synthesis. The genetic code can be interpreted by measurement of gene expression level. A normal cell has different expression level from abnormal cell.

2.2Microarray technology

Microarray technology is a modern technology which has been attracted from researchers. This technology simultaneously measures thousands gene expression level in an experiment. Microarrays have been applied in many studies.

There are two major objectives of microarray studies. First, microarray data has been created predictive model. The predictive model is fitted by using previous microarray data and classifies new data. Second, this data type has been used to detect pattern. The pattern-detection methods screen for interesting relationships.

The genes-to-protein transformation has two steps. Transcription is the first step to transform DNA to RNA. Translation is the second step to transform RNA to protein. The gene expression means whole steps of genes-to-protein transformation.

The data analysis of DNA microarray is a complex process because an important characteristic of this data is a high dimension matrix which has many real numbers. Furthermore, the DNA microarray has many thousands of genes and always has few samples. It leads to overfitting problem. The overfitting problem was occurred when the amounts of variables more than the amounts of samples. Besides the above problem, collinearlity is an important problem in DNA microarray analysis because of genes sharing biological pathway.

2.3Model selection

Logistic regression is a common method for data classification which is a qualitative response. A classification was done by fitting model which often uses logistic regression. To fitting model must estimate coefficient values. There are many methods for estimating coefficients. One of the popular methods is OLS. The ordinary least squares (OLS) estimates perform by minimizing the residual squared error. The outstanding problems of using OLS are predictive accuracy and model interpretability. The predictive accuracy was not good for OLS estimates because OLS estimates often have low bias and high variance. It tends to overfitting problem which has high accuracy to training data but gives low accuracy to other data. The model interpretability of OLS estimates were not satisfied because there are many variables. It was difficult to identify that which variables affect to response.

Model selection is a method for solving these problems. It was divided into three techniques.

2.3.1 Subset selection

This method identifies a subset of the variable which is believed that associate to response. It can solve a model interpretability problem but it cannot solve the predictive accuracy because this method is discrete process. If the training data has change, the model will have many changes. It leads that model is not stable.

2.3.2 Dimension reduction

This method transforms data to other dimension and uses some part of the transformed data to fitting model. It cannot solve the model interpretability problem because it cannot identify that which variables are important variables.

2.3.3 Shrinkage technique

This method shrinks coefficient values of variables which have not association with response toward zero but variables that are important slightly shrinks coefficient values.

2.4Shrinkage technique

Shrinkage technique reduces coefficients by constrain optimization. Traditional technique is a ridge regression which uses continuous process. It builds more stable model than subset selection but coefficients of model are not toward zero so the model is still not easy to interpret. Later, the least absolute shrinkage and selection operation (LASSO) are presented. It shrinks some coefficients and set other to zero, while casual variables automatic select into the model. An outstanding advantage of LASSO is to coefficients can be set to zero. The LASSO has still a limitation. If variables highly correlate, it will select only one variable in that group. Afterwards, the elastic netis developed for solving this problem. It also shrinks some associated coefficients and selects them into the model. Other non- casual variables are set to zero. A good feature of elastic net is variable grouping in highly correlate. This technique is hybrid between ridge regression and LASSO. In the most previous studies, the elastic net has performance better than the LASSO. After, some researchers try to develop the LASSO by eliminating weakness of LASSO. The group

LASSO can perform same as the LASSO but it can group variables. It has still limitation of grouping variables. The initial information about grouping structure is limitation of the group LASSO. Then, the OSCAR has been developed to solving the initial data requirement. Octagonal shrinkage and clustering algorithm for regression (OSCAR) simultaneously selects variables while grouping predictive cluster. It is not require the initial information but the previous study found that the performance was not different from the elastic net. Recently, the normal exponential gamma distribution (NEG distribution) has been developed. It is a Bayesian-inspired penalized maximum likelihood approach .It use NEG distribution prior, which has a sharper peak at zero and heavier tails. The sharp peak at zero will have a few numbers of selected variables and the heavy tail will little shrink coefficients of causal variables. It may benefit to variable selection because the ideal variable selection should have a little number of causal variables which select into the model.

2.4.1 Ridge regression

Ridge regression is similar to ordinary least square (OLS) estimates but ridge regression has a regularization term while the OLS has not. The regularization term affects to estimating coefficients of selected variables. The coefficients were estimated by minimizing a slightly quantity.

$$\beta_{ridge} = min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

or

$$\beta_{ridge} = min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad subject\ to \sum_{j=1}^{p} \beta_j^2 \leq t$$

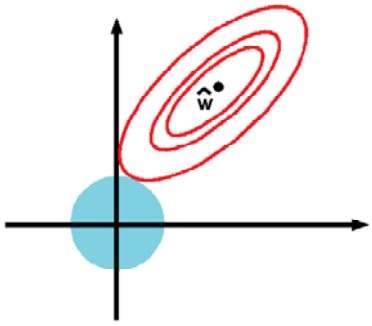


Figure 2.3 Constraint region of ridge regression

From the figure 2.3, the blue area is a constraint region for ridge regression while the red ellipse is the contours of RSS (objective function). The constraint region represents circle.

A tuning parameter is turned on by setting lambda parameter > 0. The regularization term (or called a shrinkage penalty) of ridge regression is $\lambda \sum_{j=1}^{p} \beta_j^2$ and the other part of this equivalence is called RSS ($\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2$). The ridge regression estimates coefficient values by minimizing RSS. If lambda equal zero, the ridge regression will become OLS. When lambda increases, the ridge regression fit will decrease and will decrease variance. The ridge regression can solve the predictive accuracy because it is a continuous process. The model interpretability still is a problem of ridge regression due to it cannot shrink and set coefficient values to exactly zero.

2.4.2 LASSO

The LASSO estimates coefficients by constraint optimization same as the ridge regression but it was developed to overcome the limitation of ridge regression. The LASSO can set coefficients to zero. This advantage leads to solve the model interpretability.

$$\beta_{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right\}$$

or

$$\beta_{lasso} = min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad subject\ to \sum_{j=1}^{p} |\beta_j| \leq t$$

The formulation of LASSO is similar to ridge regression. The difference is regularization term or shrinkage penalty. The formulation of LASSO, β_j^2 was replaced by $|\beta_j|$. The LASSO simultaneously does variable selection and shrinkage.

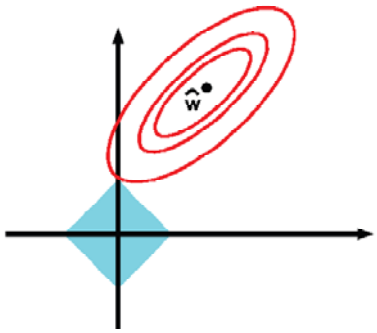


Figure 2.4 Constraint region of LASSO.

From the figure 2.4, the blue area is a constraint region for LASSO while the red ellipse is the contours of RSS (objective function). The constraint region represents diamond shape. It is believed that the contours of RSS will meet the first point of constraint region at corners so the contours of RSS might often meet the diamond-shape constrain region.

Although, the LASSO can solve both the predictive accuracy and model interpretability problem, it still has some limitations as follow:

1. When many more variables than samples, the LASSO has been limited the number of selected variables because of the nature of the convex optimization. It cannot select more the number of variables than the number of samples.

2. In case variables have high correlate together, the LASSO will select only one variable from that group.

3. If data has more the number of samples than the number of variables and variables have high correlate together, it has observed that the ridge regression will have better performance than the LASSO.

2.4.3 Elastic net

The elastic net is hybrid between the ridge regression and the LASSO. It shrinks some coefficients of variables and sets other to zero as the LASSO. It also automatically selects variables.

$$\beta_{elastic\ net} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left((1-\alpha)|\beta_j| + \alpha|\beta_j|^2\right)\right\}$$

or

$$\beta_{lassoelastic\ net} = min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\}\ subject\ to \sum_{j=1}^{p}\left((1-\alpha)|\beta_j| + \alpha|\beta_j|^2\right) \le t$$

The elastic net was developed to solving the limitation of the LASSO. The outstanding advantage of the elastic net is variable grouping. It was said that this method suits for highly correlate data.

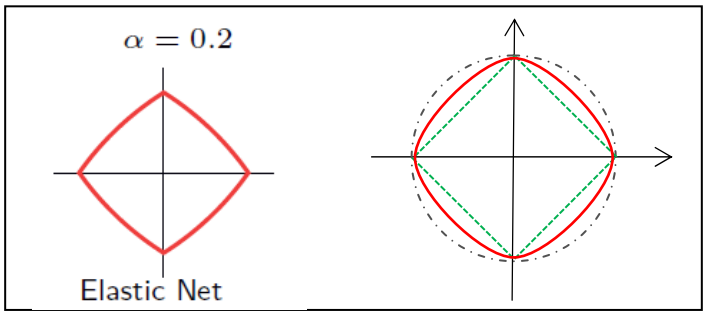


Figure 2.5 Constraint region of elastic net.

From the figure 2.5, the red shape has shown a constraint region for the elastic net. The constraint region represents a shape between the constraint region of LASSO and ridge regression. The shape will change follow the alpha.

2.4.4 Group LASSO

The group LASSO was developed for solving the one of limitation of LASSO. When variables have highly correlated together, the LASSO will select only

one variable from group. The group LASSO can group variables but this method still has a limitation. It requires initial information about grouping structure.

2.4.5 Octagonal shrinkage and clustering algorithm for regression (OSCAR)

The OSCAR does simultaneously select variables while supervised clustering on the important variables. This method does not require initial information same as the group LASSO.

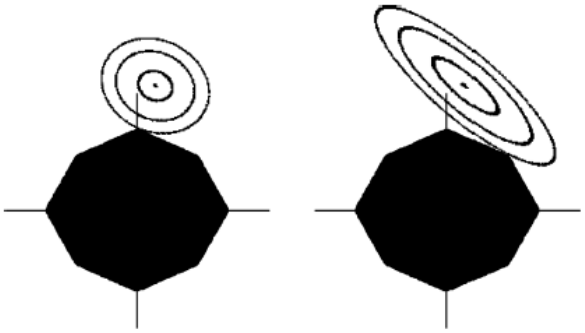


Figure 2.6 Constraint region of OSCAR.

2.4.6 The normal exponential gamma (NEG) distribution

This method is Bayesian-inspired. It is not full Bayesian theory. The NEG distribution seeks only the posterior mode by using maximization algorithm for strongly associated variables. If variables may be not associated, the posterior mode of the coefficient of variables will be zero.

This method uses the normal exponential gamma distribution (NEG) which is a shrinkage prior distribution. The important characteristics of NEG distribution are a sharper peak at zero and a heavier tails. The sharper peak will select few variables to the model. It agrees with the belief that there are few causal variables. The heavier tails will shrink little coefficient values of selected variables.

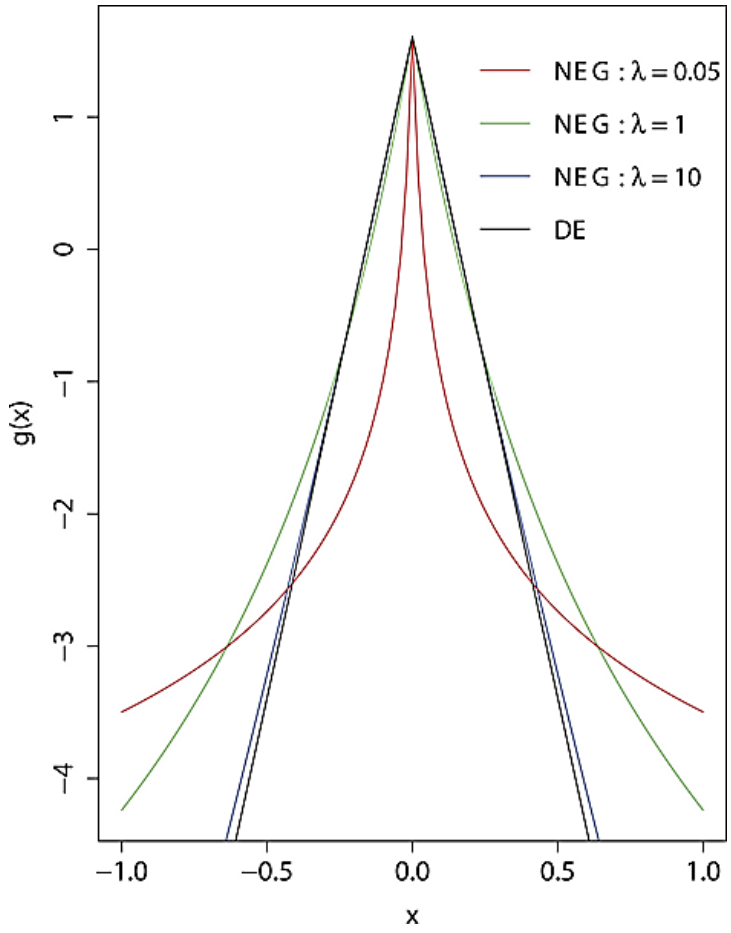


Figure 2.7 Constraint region of NEG distribution

From figure 2.7, the NEG distribution compares with DE distribution. The NEG shows sharper peak and heavier tail than DE distribution.

2.4.7 The double exponential (DE) distribution

This method is Bayesian-inspired same as the NEG distribution but it uses the double exponential distribution (DE) which is a shrinkage prior distribution.

In a Bayesian viewpoint, the coefficients of variables in a model have some prior distribution. When multiplying the prior distribution by using likelihood, it will give the posterior distribution.

It was proved that the LASSO is the posterior mode for coefficients under a double exponential prior.

CHAPTER III
RESEARCH METHODOLOGY

3.1 Material

3.1.1 Leukemia dataset

Gene expression data of leukemia microarray study were classified into two groups: AML and ALL. There were 7,129 variables (genes). The whole data were divided to 38 samples for train set and 34 samples for test set.

3.1.2 Lung cancer dataset

Gene expression data of lung cancer microarray study were classified into two groups: mesothelioma and ADCA. There were 12,533 variables (genes). The whole data were divided to 32 samples for train set and 149 samples for test set.

3.1.3 Prostate cancer dataset

Gene expression data of prostate cancer microarray study were classified into two groups: tumor and normal. There were 12,600 variables (genes). The whole data were divided to 102 samples for train set and 34 samples for test set.

3.2 Method

Case/ control studies are always analyzed by logistic regression that has a form of equation

$$\log \frac{\Pr(Y=1 \mid X)}{\Pr(Y=0 \mid X)} = X^T \beta$$

When X is a matrix of predictive variable which has n×(p+1) size ,Y is a dichotomous phenotype vector and β is a vector of model's coefficient.

Thecoefficientsare typicallydetermined by maximum likelihood estimation that has log-likelihood equation,

$$L(\beta) = \sum_{i=1}^{n} y_i B^T X_i - \log(1 - \exp(B^T X_i))$$

In genetic study, to find associate between disease and variables such as genotype, gene expression often encounter problems due to the number of variables much more than the number of samples. This problem may lead to overfitting problem. Moreover, the data which come from gene studies usually correlate between each variable and lead to collinearlity problem. The standard logistic regression cannot produce a suitable predictive model. Penalized logistic regression methods, which estimate coefficient while select variables, are attractive methods.

3.2.1 Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO is a type of penalized regression method which shrinks some coefficients and automatic selects into the model while sets others to zero. The LASSOlogistic regression estimates parameter by

$$\beta_{lasso} = \min_{\beta} \left\{ \sum_{j=1}^{N} \left(y_i \log h(\beta^T X) + (1 - y_i) \log\left(1 - h(\beta^T X)\right) \right) \right\}$$

and

$$h(z) = \frac{1}{1 + e^{-z}}$$

subject to,

$$\sum_{j=1}^{p} |\beta_j|$$

where,β is a coefficient of model and *N* is the number of samples.

3.2.2 Elastic net

The elastic net is another type of penalized regression method which also shrinks some coefficients, sets other to zero and selects variables but it can group variables. The elastic net logistic regression estimates parameter by

$$\beta_{elastic\ net} = \min_{\beta} \left\{ \sum_{j=1}^{N} \left(y_i \log h(\beta^T X) + (1 - y_i) \log\left(1 - h(\beta^T X)\right) \right) \right\}$$

and

$$h(z) = \frac{1}{1 + e^{-z}}$$

subject to,

$$\sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha) \left| \beta_j \right| \right)$$

where, β is a coefficient of model, N is the number of samples and α is a value for tuning parameter.

3.2.3 Normal Exponential-γ Distribution (NEG Distribution)

The NEG distribution is a Bayesian-inspired method, which finds the posterior mode from defining prior distribution. The NEG distribution has a sharper peak at zero and heavier tails more than double exponential distribution (DE), which is proven that it is similar to LASSO. The NEG distribution can be represented

$$NEG(\beta \mid \lambda, \gamma) = k \exp\left(\frac{\beta^2}{4\gamma^2} \right) D_{-2\lambda-1}\left(\frac{|\beta|}{\gamma} \right)$$

where λ and $\gamma$are shape and scale parameters, k is a integrating constant and D is a parabolic cylinder function. When λ and γ increase, the NEG distribution will be changed to DE distribution. The posterior logarithms in Bayes theorem can be represented

$$\log p(\beta \mid X, y) = L(\beta) - f(\beta) + \text{const},$$

where L is the log-likelihood for a logistic regression model, $f(.)$ is minus the log-prior density. The minus sign of $f(.)$ shows this function penalizes the complex model. The estimator finds to maximize a penalized log-likelihood. The expectation–maximization (EM) algorithm was used in linear regression but this study is designed for logistic regression. Therefore, Newton's method is used instead.

$$\beta_j^{new} = \beta_j - \frac{L'(\beta) - f'(\beta)}{L''(\beta) - f''(\beta)}$$

3.3 Evaluation method

A good predictive model for classification must have a minimal test error to new data so fitting the model should have an error control method. The K-fold cross-validation is a randomly resampling into k size then brings this data to fit a model and remaining data is used to calculate cross-validation error. The K-fold cross-validation is used in the LASSO and elastic net. The type-I error control is used in the NEG distribution.

3.3.1 LASSO

1) Gene expression data was separated into two groups. Frist data set is train set for fitting model. The other is test set for evaluating test error.

2) The train set was used for cross-validation to tuning parameter (lambda).

3) A lambda was selected for tuning parameter that builds the least complicated model. The selected lambda should be a value which gives a smallest deviance (a cross-validation error) or a smallest plus one standard error deviance of the model.

4) The model was fit from a selected lambda.

5) The test set was predicted in the model.

6) The model was evaluated by measuring a deviance (a test error) , a number of selected variables , a predictive accuracy and time.

3.3.2 Elastic net

1) Gene expression data was separated into two groups. Frist data set is train set for fitting model. The other is test set for evaluating test error.

2) The train set was used for cross-validation to tuning parameter.

3) An alpha was selected to fitting a model which has a smallest plus one standard deviation deviance.

4) A lambda was selected for tuning parameter that builds the least complicated model. The selected lambda should be a value which gives a smallest deviance (a cross-validation error) or a smallest plus one standard error deviance of the model.

5) The model was fit from a selected lambda.

6) The test set was predicted in the model.

7)The model was evaluated by measuring a deviance (a test error), a number of selected variables and a predictive accuracy.

3.3.3 NEG distribution

1) Gene expression data was separated into two groups. Frist data set is train set for fitting model. The other is test set for evaluating test error.

2) The train set was built the model by controlling type-I error.

3) A shape and scale parameter were selected to fitting a model

4)The test set was predicted in the model.

5)The model was evaluated by measuring a deviance (a test error), a number of selected variables and a predictive accuracy

3.3.4 DE distribution

1) Gene expression data was separatedinto two groups. Frist data set is train set for fitting model. The other is test set for evaluating test error.

2) The train set was built the model by controlling type-I error.

3) A lambda parameter was selected to fitting a model

4)The test set was predicted in the model.

5)The model wasevaluated by measuring a deviance (a test error), a number of selected variables and a predictive accuracy

CHAPTER IV
RESULTS AND DISCUSSION

The performance of the normal exponential gamma (NEG) distribution, the double exponential (DE) distribution,LASSO, elastic net and DE distribution logistic regression were compared in real data of DNA microarray. There are three datasets in this experiment. First, leukemia microarray data was classified into two groups; ALL and AML. Second, lung cancer microarray data was classified into two groups; mesothelioma and ADCA. Third, prostate cancer microarray data was classified into two groups; tumor and normal.

The performances of all methods were evaluated by measuring the number of selected variables, predictive accuracy, deviance and computational time. The experiment was repeated 30 times in all methods and all datasets. Then values were compared by using paired sample t-test. All methods were paired to compare performances in each datasets. Paired t-test compares mean between two methods. If the p-value more than 0.05, it shows that two mean values is not difference.

4.1 Tuning parameter

A shrinkage method was tuned parameters and selected a suitable parameter for fitting model.

Table 4.1 summary of tuning parameter between DE, LASSO, NEG and elastic net

Dataset	parameter	DE	LASSO	NEG	Elastic net
Leukemia	lambda	0.1	0.123	0.1	0.1617
	alpha	-	-	-	0.3
	shape	-	-	0.1	-
Lung cancer	lambda	0.1	0.00078	0.1	0.0013
	alpha	-	-	-	0.9
	shape	-	-	0.1	-
Prostate cancer	lambda	0.1	0.0348	0.1	0.0386
	alpha	-	-	-	0.9
	shape	-	-	0.1	-

4.2Performance

When a predictive model was created from training set data, the model was evaluated performance by measuring the number of selected variables, predictive accuracy, deviance and computational time.

4.2.1 Leukemia microarray dataset

4.2.1.1 DE distribution

Table 4.2 Performance of DE distribution in leukemia

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	25	0.052	0.4396	88.2353	18.5416
2	25	0.052	0.4412	91.1765	17.4094
3	25	0.052	0.4398	88.2353	16.269
4	24	0.052	0.4399	91.1765	15.4716
5	24	0.052	0.4403	91.1765	16.2095
6	24	0.052	0.4406	88.2353	18.3986
7	23	0.052	0.4394	91.1765	17.2146
8	24	0.052	0.4409	88.2353	16.8856
9	21	0.052	0.4404	88.2353	17.3951
10	26	0.052	0.4393	91.1765	16.5052
11	25	0.052	0.4404	88.2353	17.4992
12	27	0.052	0.438	94.1176	15.9804
13	23	0.052	0.4393	91.1765	16.6021
14	24	0.052	0.4417	88.2353	17.7906
15	24	0.052	0.4405	91.1765	16.1582
16	26	0.052	0.4398	91.1765	16.9872
17	24	0.052	0.4391	91.1765	17.5861
18	21	0.052	0.4399	91.1765	16.838
19	23	0.052	0.4395	88.2353	17.6303
20	21	0.052	0.4393	88.2353	17.4648
21	22	0.052	0.4395	88.2353	19.2666
22	24	0.052	0.4391	91.1765	16.6654
23	25	0.052	0.4401	88.2353	17.3669
24	26	0.052	0.4375	91.1765	16.0465
25	26	0.052	0.4404	88.2353	18.9069
26	27	0.052	0.4405	91.1765	16.7942
27	22	0.052	0.4405	88.2353	17.4851
28	26	0.052	0.4403	91.1765	15.9638
29	25	0.052	0.4404	91.1765	17.18
30	24	0.052	0.4401	91.1765	15.0019

4.2.1.2 LASSO

Table 4.3 Performance of LASSO in leukemia

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	10	4.07	24.1975	73.5294	30.9337
2	5	3.97	32.0101	70.5882	35.7578
3	9	4.80	25.2513	73.5294	33.1021
4	14	4.34	14.078	88.2353	20.8129
5	14	4.23	15.5777	88.2353	20.6396
6	10	4.65	23.0276	73.5294	30.9337
7	7	4.34	25.6379	70.5882	34.5688
8	10	4.45	23.5401	73.5294	30.9337
9	17	4.23	8.2535	88.2353	19.5402
10	9	4.75	23.9318	73.5294	33.1021
11	9	4.45	23.9369	73.5294	33.1021
12	6	4.23	31.6507	70.5882	34.8379
13	11	3.98	20.1419	79.4118	27.6216
14	12	3.89	21.3398	79.4118	26.1635
15	9	4.45	23.7264	73.5294	33.1021
16	11	4.23	22.3442	79.4118	29.1509
17	9	4.12	24.2324	73.5294	33.1021
18	7	4.23	27.2873	70.5882	34.5688
19	9	4.23	23.6341	73.5294	33.1021
20	14	4.12	13.5251	88.2353	20.2209
21	11	4.56	22.4842	79.4118	29.1509
22	7	4.67	28.8765	70.5882	34.5688
23	9	4.78	23.3966	73.5294	33.1021
24	10	4.87	22.8963	73.5294	30.9337
25	14	4.66	11.3732	88.2353	20.1714
26	11	4.98	22.1688	79.4118	29.1509
27	6	4.89	25.9674	70.5882	34.8379
28	9	4.34	26.3177	73.5294	33.1021
29	9	4.56	25.3941	73.5294	33.1021
30	9	4.12	27.4609	73.5294	33.1021

4.2.1.3 NEG distribution

Table 4.4 Performance of NEG distribution in leukemia

Iteration	Nonzero	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	9	0.151333	0.3867	94.1176	14.6676
2	13	0.151333	0.3798	94.1176	15.2617
3	15	0.151333	0.3792	94.1176	15.1537
4	14	0.151333	0.3799	97.0588	14.5962
5	15	0.151333	0.3787	94.1176	15.1432
6	8	0.151333	0.3868	94.1176	14.7273
7	14	0.151333	0.3787	94.1176	15.2861
8	13	0.151333	0.3784	94.1176	15.067
9	13	0.151333	0.379	94.1176	15.3194
10	16	0.151333	0.3746	97.0588	12.9487
11	12	0.151333	0.3754	97.0588	12.933
12	14	0.151333	0.3781	94.1176	15.3426
13	15	0.151333	0.3783	94.1176	15.2286
14	13	0.151333	0.382	94.1176	14.868
15	11	0.151333	0.3824	94.1176	14.8943
16	15	0.151333	0.3785	94.1176	15.2169
17	13	0.151333	0.3796	94.1176	15.2853
18	14	0.151333	0.379	94.1176	15.2655
19	8	0.151333	0.3844	94.1176	15.0698
20	14	0.151333	0.3787	94.1176	15.1611
21	12	0.151333	0.3791	94.1176	15.1974
22	14	0.151333	0.3806	94.1176	14.85
23	7	0.151333	0.3871	94.1176	14.6254
24	14	0.151333	0.379	94.1176	15.1111
25	14	0.151333	0.379	94.1176	15.1655
26	18	0.151333	0.3782	97.0588	14.776
27	13	0.151333	0.3788	94.1176	15.1333
28	14	0.151333	0.3817	94.1176	14.7369
29	13	0.151333	0.3788	97.0588	14.8161
30	11	0.151333	0.3821	94.1176	14.9553

4.2.1.4 elastic net

Table 4.5 Tuning alpha parameter of elastic net in leukemia

Alpha	No. of variables	Train Accuracy	Train Deviance (min+1SE)
0.2	115	100	19.790
0.3	86	100	18.433
0.4	56	100	19.120
0.5	33	100	20.581
0.6	20	100	23.398
0.7	18	100	24.196
0.8	13	100	26.119
0.9	11	100	25.798

Table 4.6 Performance of elastic net in leukemia

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	86	12.60	16.9129	82.3529	19.9814
2	70	8.57	20.3859	79.4118	22.2707
3	90	8.83	16.0595	85.2941	18.721
4	77	8.75	16.8179	82.3529	21.2601
5	88	8.89	15.08	85.2941	19.3095
6	95	9.01	13.4443	85.2941	16.8982
7	70	8.45	18.5479	79.4118	22.2707
8	90	8.23	15.0195	85.2941	18.721
9	88	9.12	15.9839	85.2941	19.3095
10	91	9.03	14.7782	85.2941	18.2067
11	95	8.45	13.7622	85.2941	16.8982
12	93	8.78	16.5678	85.2941	17.2971
13	79	8.98	17.2652	82.3529	20.7692
14	74	9.34	19.7044	82.3529	21.7572
15	70	9.78	18.2253	79.4118	22.2707
16	90	8.90	15.9575	85.2941	18.721
17	74	9.45	18.3529	82.3529	21.7572
18	77	9.03	18.1996	82.3529	21.2601
19	77	9.06	18.0603	82.3529	21.2601
20	62	8.23	21.5353	76.4706	23.4517
21	79	9.50	17.1045	82.3529	20.7692
22	91	8.23	16.3941	85.2941	18.2067
23	86	8.67	15.9278	82.3529	19.9814
24	86	9.23	16.4098	82.3529	19.9814
25	62	8.32	20.0688	76.4706	23.4517

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
26	67	9.14	19.471	79.4118	22.8101
27	79	8.45	17.1721	82.3529	20.7692
28	90	9.12	16.9844	85.2941	18.721
29	102	8.45	12.962	85.2941	15.5662
30	79	9.54	19.7445	82.3529	20.7692

4.2.2 Lung cancer microarray dataset

4.2.2.1 DE distribution

Table 4.7 Performance of DE distribution in lung cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	14	0.115333	0.2614	96.6443	24.3095
2	16	0.115333	0.2605	95.302	24.5153
3	15	0.115333	0.261	95.302	23.3561
4	15	0.115333	0.2607	95.302	24.2399
5	14	0.115333	0.2612	95.302	24.2682
6	15	0.115333	0.263	96.6443	26.3998
7	15	0.115333	0.2597	95.302	23.0089
8	16	0.115333	0.2603	96.6443	23.6914
9	14	0.115333	0.26	95.302	23.8844
10	13	0.115333	0.2616	95.302	23.9381
11	15	0.115333	0.259	95.302	23.673
12	13	0.115333	0.2591	95.302	23.985
13	11	0.115333	0.2625	94.6309	25.009
14	14	0.115333	0.2622	95.302	24.3892
15	15	0.115333	0.2616	96.6443	25.0668
16	15	0.115333	0.2604	96.6443	23.6981
17	14	0.115333	0.2603	95.302	23.3999
18	14	0.115333	0.2613	96.6443	26.1627
19	15	0.115333	0.2622	95.9732	27.164
20	14	0.115333	0.2618	96.6443	24.9919
21	14	0.115333	0.2613	96.6443	25.8785
22	14	0.115333	0.2628	96.6443	22.6839
23	16	0.115333	0.2617	96.6443	26.9024
24	14	0.115333	0.2631	95.302	25.8649
25	14	0.115333	0.2629	95.302	23.7587
26	15	0.115333	0.2602	95.302	24.2306
27	16	0.115333	0.2596	96.6443	25.622

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
28	15	0.115333	0.2615	95.302	25.0188
29	18	0.115333	0.2604	96.6443	26.3877
30	14	0.115333	0.2619	96.6443	27.6271

4.2.2.2 LASSO

Table 4.8 Performance of LASSO in lung cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	16	6.55	0.3584	95.9732	23.9242
2	16	5.27	1.6306	95.302	23.6456
3	17	6.08	0.8481	95.9732	24.2121
4	17	5.34	0.3529	95.9732	24.1099
5	15	6.23	0.6627	95.302	23.6619
6	16	6.34	0.4249	95.9732	23.9242
7	15	5.56	1.4938	95.302	23.4924
8	16	5.34	0.9433	95.302	23.4504
9	16	6.23	0.9348	95.302	23.7223
10	15	6.45	1.8724	95.302	24.1973
11	16	5.54	0.6242	95.302	23.8488
12	17	6.76	0.2479	95.9732	24.3205
13	17	5.34	0.2383	95.9732	24.1099
14	15	5.42	4.3332	95.302	28.4616
15	15	5.12	0.7819	95.302	23.4924
16	16	6.45	0.9386	95.302	23.7821
17	16	6.12	0.3588	95.9732	23.9242
18	16	6.55	0.9214	95.302	23.8488
19	15	7.02	1.3309	95.302	23.5575
20	15	5.45	1.5192	95.302	23.4924
21	16	6.23	1.1639	95.302	23.4504
22	16	6.45	1.6336	95.302	23.4504
23	15	4.56	1.5595	95.302	23.6052
24	16	5.34	0.6307	95.302	23.7821
25	16	5.45	0.5607	95.9732	23.9242
26	16	6.12	0.5754	95.302	23.7223
27	15	6.34	3.7921	95.302	29.1329
28	17	6.73	0.1992	95.9732	24.3205
29	16	5.32	0.7386	95.302	23.8488
30	16	5.23	1.7476	95.302	23.6456

4.2.2.3 NEG distribution

Table 4.9 Performance of NEG distribution in lung cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	10	0.169667	0.2509	94.6309	30.141
2	11	0.169667	0.2495	94.6309	29.7462
3	10	0.169667	0.2499	94.6309	31.7739
4	11	0.169667	0.25	95.302	29.34
5	9	0.169667	0.2511	94.6309	32.7393
6	11	0.169667	0.2501	94.6309	30.5238
7	11	0.169667	0.2508	94.6309	31.3511
8	10	0.169667	0.2505	94.6309	31.904
9	12	0.169667	0.2515	95.9732	28.5263
10	12	0.169667	0.2513	96.6443	29.0163
11	11	0.169667	0.2517	96.6443	29.602
12	13	0.169667	0.2524	95.9732	26.5953
13	8	0.169667	0.2492	94.6309	32.5387
14	10	0.169667	0.2515	96.6443	29.8868
15	12	0.169667	0.2509	94.6309	30.4136
16	14	0.169667	0.2518	95.9732	28.6773
17	12	0.169667	0.2518	95.302	37.2937
18	11	0.169667	0.2507	94.6309	30.189
19	13	0.169667	0.2523	94.6309	35.4907
20	15	0.169667	0.2518	97.9866	19.534
21	9	0.169667	0.2507	94.6309	32.3839
22	7	0.169667	0.2497	94.6309	32.7537
23	12	0.169667	0.2519	97.3154	24.3721
24	16	0.169667	0.2526	97.3154	19.9449
25	14	0.169667	0.2508	97.9866	18.945
26	8	0.169667	0.2495	94.6309	33.7037
27	13	0.169667	0.2523	95.302	37.5491
28	10	0.169667	0.2503	94.6309	32.9719
29	9	0.169667	0.2511	94.6309	32.5756
30	13	0.169667	0.2523	95.9732	23.906

4.2.2.4 elastic net

Table 4.10 Tuning alpha parameter of elastic net in lung cancer

Alpha	No. of variables	Train Accuracy	Train Deviance (min+1SE)
0.1	566	100	0.9258
0.2	320	100	1.099
0.3	215	100	1.1831
0.4	147	100	1.1292
0.5	117	100	1.065
0.6	90	100	1.0241
0.7	66	100	0.8691
0.8	40	100	0.6307
0.9	31	100	0.4059

Table 4.11 Performance of elastic net in lung cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	31	10.03	0.3293	97.9866	15.37
2	31	9.45	0.3979	97.9866	15.3398
3	31	9.47	0.6829	97.9866	15.4063
4	31	8.58	0.4324	97.9866	15.4491
5	31	9.57	0.3197	97.9866	15.3398
6	32	9.35	0.4888	97.9866	15.6196
7	31	9.02	0.4652	97.9866	15.37
8	31	9.43	0.4291	97.9866	15.4989
9	31	8.92	0.6273	97.9866	15.4989
10	25	9.98	1.7292	97.3154	18.2063
11	32	8.47	0.747	97.9866	15.6196
12	31	8.68	0.3187	97.9866	15.3398
13	31	10.20	0.2712	97.9866	15.37
14	24	10.03	2.7698	97.3154	19.1897
15	30	10.08	1.1707	97.9866	16.6125
16	32	10.08	0.9917	97.9866	15.7645
17	30	9.27	0.9181	97.9866	16.4541
18	32	9.52	0.8374	97.9866	15.8501
19	29	9.62	1.8348	97.3154	17.1876
20	31	10.05	0.414	97.9866	15.37
21	31	12.70	1.2358	82.3529	16.0561

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
22	32	9.53	0.8662	97.9866	15.6881
23	27	9.85	2.1291	97.9866	17.4151
24	32	10.33	0.7802	97.9866	15.7645
25	31	9.57	0.484	97.9866	15.4491
26	31	9.32	0.4034	97.9866	15.4989
27	21	8.38	3.2907	97.3154	22.5707
28	31	9.97	0.2805	97.3154	15.4063
29	32	9.17	0.8058	97.9866	15.6196
30	30	8.67	1.7687	97.9866	16.7861

4.2.3Prostate cancer microarray dataset

4.2.3.1 DE distribution

Table 4.12 Performance of DE distribution in prostate cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	42	0.382	0.591	94.1176	86.0048
2	44	0.382	0.5924	94.1176	91.692
3	41	0.382	0.5924	94.1176	99.8087
4	40	0.382	0.5918	91.1765	122.1934
5	43	0.382	0.5913	91.1765	120.4543
6	41	0.382	0.5919	94.1176	98.9922
7	40	0.382	0.5921	88.2353	132.3951
8	45	0.382	0.593	91.1765	125.2544
9	41	0.382	0.5929	91.1765	123.9737
10	41	0.382	0.5914	94.1176	118.8265
11	41	0.382	0.5927	94.1176	87.0423
12	43	0.382	0.5926	94.1176	108.0147
13	42	0.382	0.591	94.1176	118.6211
14	41	0.382	0.5926	94.1176	110.6828
15	41	0.382	0.5922	94.1176	112.331
16	42	0.382	0.593	94.1176	116.0481
17	45	0.382	0.5916	91.1765	119.563
18	41	0.382	0.5927	91.1765	116.376
19	40	0.382	0.5924	94.1176	104.4922
20	44	0.382	0.5921	94.1176	94.1389
21	41	0.382	0.5935	94.1176	94.9467
22	44	0.382	0.5921	91.1765	118.8375
23	41	0.382	0.593	94.1176	115.4402
24	42	0.382	0.5918	94.1176	92.5644

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
25	40	0.382	0.5921	94.1176	91.3048
26	44	0.382	0.5926	94.1176	86.8961
27	42	0.382	0.5932	94.1176	111.9765
28	40	0.382	0.5923	94.1176	102.28
29	42	0.382	0.5933	94.1176	104.1894
30	41	0.382	0.593	94.1176	97.517

4.2.3.2 LASSO

Table 4.13 Performance of LASSO in prostate cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	22	10.13	50.9001	82.3529	47.8602
2	22	10.22	56.9901	85.2941	46.9909
3	23	10.21	43.3699	82.3529	40.834
4	22	10.34	53.4536	82.3529	47.8602
5	30	9.89	43.5719	94.1176	23.0734
6	24	9.23	41.6667	82.3529	38.8993
7	22	10.54	57.6007	85.2941	46.9909
8	23	10.15	46.7609	82.3529	44.3118
9	24	9.34	47.855	82.3529	38.8993
10	23	10.45	45.186	82.3529	44.3118
11	24	9.67	46.0281	82.3529	38.8993
12	22	9.23	50.8122	85.2941	47.2843
13	23	10.45	45.13	82.3529	40.834
14	26	9.23	38.0093	88.2353	29.9176
15	22	9.78	49.1116	82.3529	47.8602
16	26	10.56	43.8376	91.1765	25.2951
17	23	10.65	37.5676	82.3529	37.6269
18	23	10.12	43.7642	82.3529	37.6269
19	23	9.28	44.7687	82.3529	37.6269
20	23	9.97	48.8419	82.3529	40.834
21	23	10.56	41.1411	85.2941	35.8614
22	23	10.12	46.8474	82.3529	44.3118
23	23	10.65	42.4065	85.2941	36.5893
24	24	10.34	45.9204	82.3529	38.8993
25	23	10.11	51.6318	82.3529	44.3118
26	26	9.54	41.3783	88.2353	29.9176
27	22	9.45	57.8443	85.2941	46.9909

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
28	23	9.12	39.1883	85.2941	36.5893
29	26	9.85	37.1811	88.2353	29.9176
30	23	10.54	52.3313	82.3529	44.3118

4.2.3.3 NEG distribution

Table 4.14 Performance of NEG distribution in prostate cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	36	0.309333	100	94.1176	76.8981
2	36	0.309333	100	94.1176	86.2466
3	38	0.309333	100	94.1176	72.2725
4	34	0.309333	100	94.1176	72.2725
5	35	0.309333	100	94.1176	80.6565
6	36	0.309333	100	94.1176	88.0919
7	33	0.309333	100	91.1765	79.1058
8	34	0.309333	100	94.1176	129.292
9	35	0.309333	100	94.1176	118.1338
10	35	0.309333	100	94.1176	114.4855
11	32	0.309333	100	94.1176	84.3877
12	38	0.309333	100	94.1176	92.037
13	35	0.309333	100	91.1765	96.5618
14	34	0.309333	100	94.1176	127.5577
15	37	0.309333	100	94.1176	84.8608
16	38	0.309333	100	94.1176	93.6265
17	38	0.309333	100	94.1176	95.9948
18	36	0.309333	100	94.1176	79.4716
19	36	0.309333	100	94.1176	75.826
20	36	0.309333	100	94.1176	91.9031
21	33	0.309333	100	94.1176	108.0091
22	34	0.309333	100	94.1176	87.6592
23	36	0.309333	100	91.1765	73.6174
24	33	0.309333	100	94.1176	117.1927
25	35	0.309333	100	91.1765	75.2635
26	35	0.309333	100	94.1176	72.5832
27	36	0.309333	100	94.1176	97.754
28	37	0.309333	100	94.1176	94.8584
29	38	0.309333	100	94.1176	88.2276
30	32	0.309333	100	94.1176	96.0139

4.2.3.4 elastic net

Table 4.15 Tuning alpha parameter of elastic net in prostate cancer

Alpha	No. of variables	Train Accuracy	Train Deviance (min+1SE)
0.1	243	97.3684	53.0278
0.2	145	97.3684	50.5521
0.3	88	97.3684	48.8003
0.4	60	94.7368	49.2405
0.5	45	97.3684	49.4871
0.6	42	97.3684	46.9873
0.7	36	97.3684	46.3547
0.8	31	100	45.0036
0.9	28	100	44.078

Table 4.16 Performance of elastic net in prostate cancer

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
1	26	11.12	51.0063	82.3529	26.3759
2	26	8.67	57.1474	82.3529	24.8216
3	28	8.63	41.6331	88.2353	24.8549
4	26	12.42	53.0525	82.3529	26.3759
5	30	11.87	45.6977	91.1765	21.2001
6	28	12.78	43.6482	82.3529	24.5512
7	26	12.50	57.5312	82.3529	24.8216
8	27	12.63	50.6317	82.3529	26.2886
9	28	12.25	46.5524	88.2353	24.5203
10	27	12.87	46.1489	82.3529	24.8193
11	28	13.00	44.0758	88.2353	24.5203
12	27	13.25	51.2724	82.3529	26.2886
13	27	13.17	47.0509	82.3529	24.8193
14	30	12.60	40.9156	88.2353	23.8516
15	29	12.72	47.8679	88.2353	23.8674
16	29	13.43	47.8679	88.2353	23.8674
17	28	13.40	38.5145	88.2353	24.5203
18	28	13.18	46.1806	82.3529	24.5512
19	29	14.27	44.1572	88.2353	23.8674
20	27	12.88	50.8546	82.3529	24.8193
21	30	13.32	41.3727	88.2353	23.8516

Iteration	No. of variables	Time (min.)	Train Deviance	Test Accuracy	Test Deviance
22	27	12.45	46.9832	82.3529	24.8193
23	28	11.80	45.3877	88.2353	24.8549
24	28	13.32	44.9215	88.2353	24.5203
25	28	13.03	49.1179	88.2353	24.8549
26	30	13.17	42.9415	91.1765	21.2001
27	26	13.35	57.3918	82.3529	24.8216
28	28	12.58	42.6957	88.2353	24.8549
29	31	11.82	37.8211	94.1176	20.7503
30	26	12.80	53.908	82.3529	26.3759

4.3Summary result

Table 4.17 Comparison summary of performance between DE, LASSO, NEG and elastic net

Dataset	performance	DE	LASSO	NEG	Elastic net
Leukemia	No.of variables	24.2	9.9	12.9667	81.9
	Accuracy	90	76.37	94.60	82.7451
	deviance	17.05051	30.21729	14.8934	21.1139
	Time (min.)	0.052	4.4063	0.15	9.0043
Lung cancer	No.of variables	14.5667	15.8667	11.2333	30.1667
	Accuracy	95.8837	95.5034	95.481	97.3536
	deviance				
	Time (min.)	0.1153	5.8977	0.1697	9.5763
Prostate	No.of variables	41.8333	23.5333	35.3667	27.8667
cancer	Accuracy	93.2353	84.3137	93.7255	85.8823
	deviance	107.4286	39.717	91.695	24.483
	Time (min.)	0.382	9.9907	0.3093	12.5093

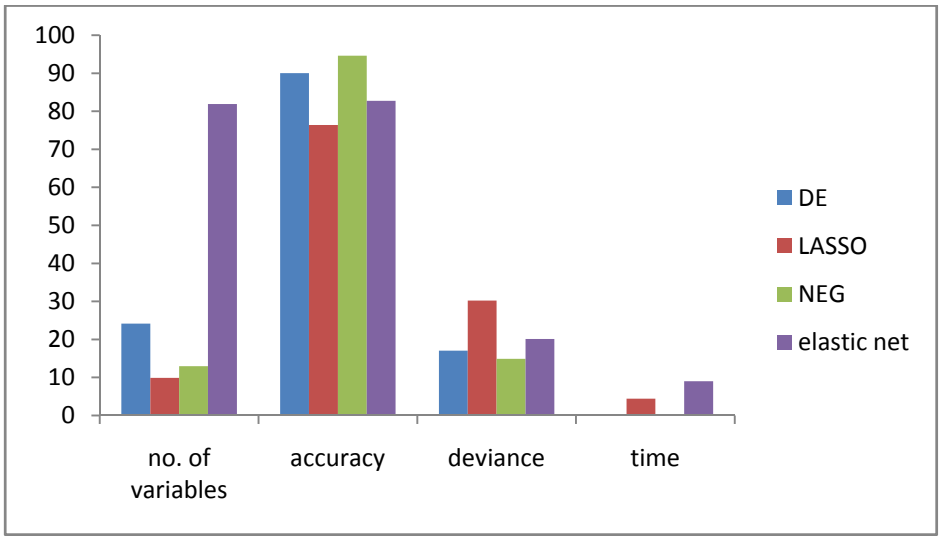


Figure 4.1 Comparison performance chart between no.of variables accuracy deviance and time in leukemia

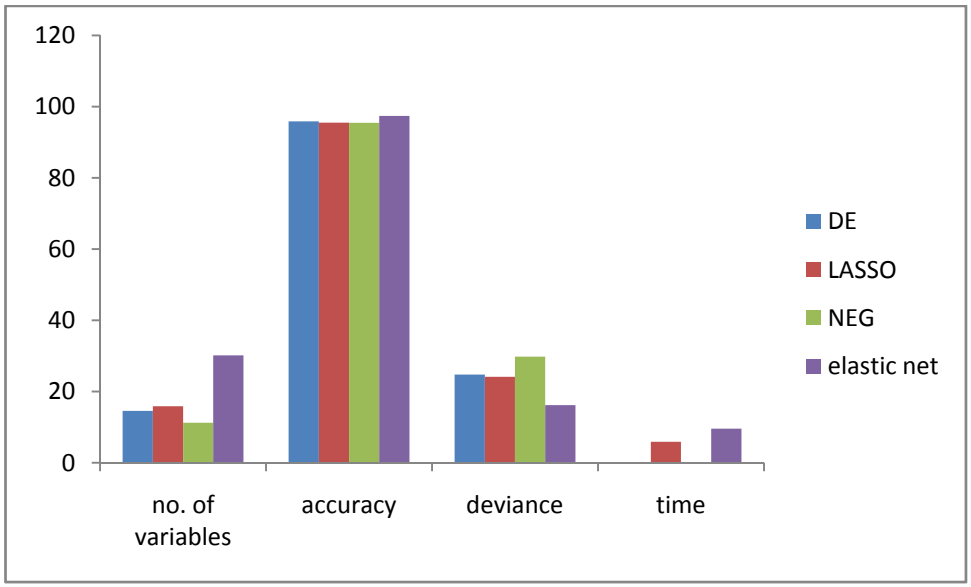


Figure 4.2 Comparison performance chart between no.of variables accuracy deviance and time in lung cancer

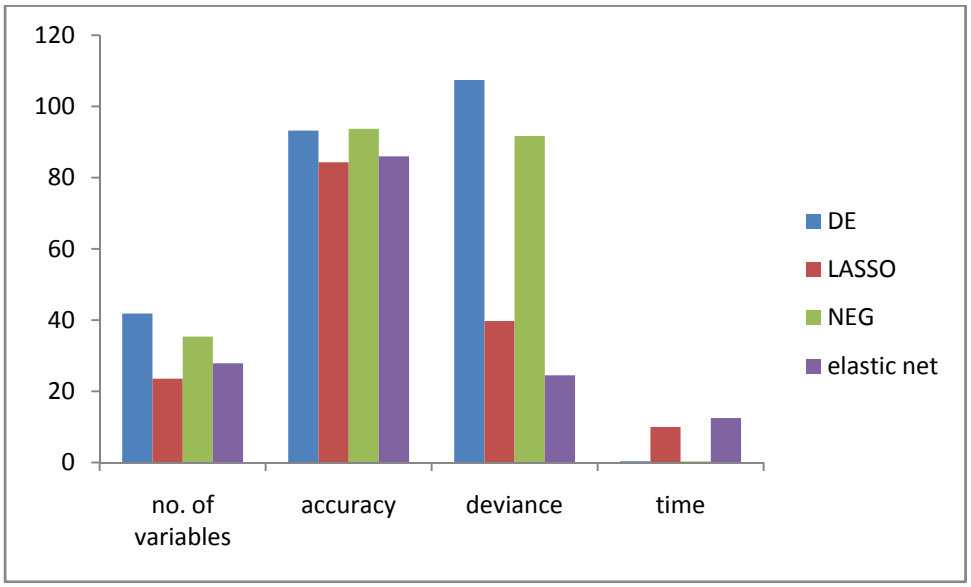


Figure 4.3 Comparison performance chart between no.of variables accuracy deviance and time in prostate cancer

4.4Paired sample t-test

4.4.1 Leukemia microarray dataset

4.4.1.1 Compare mean between the DE distribution and LASSO

Table 4.18 Paired t-test of number of selected variables, predictive accuracy,deviance between DE and LASSO in microarray

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	DE	24.2	1.66919	0.30475	.000
	LASSO	9.9	2.72093	0.49677	
Predictive accuracy	DE	90.0	1.6566	0.30245	.000
	LASSO	76.37	6.05643	1.10575	
Deviance	DE	17.05	0.97286	0.17762	.000
	LASSO	30.217	5.0415	0.92045	

Table 4.19 Comparison of time between DE and LASSO in leukemia

		Mean
Time	DE	0.052
	LASSO	4.4063

Table 4.18 showed result of paired t-test of predictive accuracy between DE and LASSO. It was believed that the LASSO is same as the DE. When compare between the DE and LASSO, the DE distribution which was controlled by type I error had higher accuracy than the LASSO which used 10 fold cross validation. When

compare deviances, result of deviance agreed with accuracy. When consider table 4.19, time of DE running less than the LASSO but variables selection of DE selected the number of variables more than the LASSO. To compare performance of the LASSO and the DE distribution in leukemia dataset found that performance of the DE distribution overcome the LASSO in accuracy, deviance and time.

4.4.1.2 Compare mean between the LASSO and elastic net

Table 4.20 Paired t-test of number of selected variables, predictive accuracy, deviance between the LASSO and elastic net in leukemia

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	LASSO	9.9	2.720	0.496	.000
	Elastic net	81.9	10.32	1.885	
Predictive accuracy	LASSO	76.3725	6.056	1.105	.000
	Elastic net	82.7451	2.645	0.483	
Deviance	LASSO	30.2173	5.04	0.920	.000
	Elastic net	20.1139	2.03628	0.37177	

Table 4.21 Comparison of time between the LASSO and elastic net in leukemia

		Mean
Time	LASSO	4.4063
	Elastic net	9.0043

Table 4.20 showed result of paired t-test of predictive accuracy between the LASSO and elastic net. It was proved that the elastic net suits for highly correlated data while the LASSO has limitations. When compare between the elastic net and LASSO, the elastic net which can variables grouping had higher accuracy than the LASSO which selects one variable from correlated group. The result agreed with the number of selected variable. The elastic net selected many variables into the model.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.21, time of the elastic net running more than the LASSO because of limitation of LASSO which selects variables at most n before it saturates. To compare performance of the LASSO and the elastic net in leukemia dataset found that performance of the elastic net overcome the LASSO in accuracy and deviance.

4.4.1.3 Compare mean between the elastic net and NEG distribution

Table 4.22 Paired t-test of number of selected variables, predictive accuracy, deviance between the elastic net and NEG distribution in leukemia

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	Elastic net	81.9	10.326	1.885	.000
	NEG	12.9667	2.428	0.44	
Predictive accuracy	Elastic net	82.7451	2.645	0.483	.000
	NEG	94.6078	1.114	0.203	
deviance	Elastic net	20.1139	2.036	0.371	.000
	NEG	14.8934	0.576	0.576	

Table 4.23 Comparisonof time between the elastic net and NEG distribution

		Mean
time	Elastic net	9.0043
	NEG	0.15

Table 4.22 showed result of paired t-test of predictive accuracy between the elastic net and NEG distribution. When compare between the elastic net and NEG distribution, the NEG distribution which was controlled type I error had higher accuracy than the elastic net. The result of the number of selected variable showed the elastic net selected many variables into the model than the NEG distribution which have a shape peak at zero.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.23, time of the elastic net running more than the NEG distribution. To compare performance of the NEG distribution and the elastic net in leukemia dataset found that performance of the NEG distribution overcomes the elastic net in accuracy, deviance, the number of selected variables and time.

4.4.1.4 Compare mean between the DE and NEG distribution

Table 4.24 Paired t-test of number of selected variables, predictive accuracy, deviance between the DE and NEG distribution in leukemia

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	DE	24.2	1.669	0.304	.000
	NEG	12.96	2.428	0.44	
Predictive accuracy	DE	90	1.65	0.302	.000
	NEG	94.60	1.114	0.203	
deviance	DE	17.05	0.972	0.177	.000
	NEG	14.89	0.576	0.105	

Table 4.25 Comparisonof time between the DE and NEG distribution in leukemia

		Mean
time	DE	0.052
	NEG	0.151

Table 4.24 showed result of paired t-test of predictive accuracy between the DE and NEG distribution. When compare between the DE and NEG distribution, the NEG distribution which has a sharper peak and a heavier tail had higher accuracy than the DE. The result of the number of selected variable showed the DE selected many variables into the model than the NEG distribution which have a sharper peak than DE distribution.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.25, time of the NEG running more than the DE distribution. To compare performance of the NEG distribution and the DE in leukemia dataset found that performance of the NEG distribution overcomes the DE in accuracy, deviance and the number of selected variables.

4.4.1.5Compare mean between the LASSO and NEG distribution

Table 4.26 Paired t-test of number of selected variables, predictive accuracy, deviance between the LASSO and NEG distribution in leukemia

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	LASSO	9.9	2.72093	0.49677	.000
	NEG	12.9667	2.42804	0.4433	
Predictive accuracy	LASSO	76.3725	6.05643	1.10575	.000
	NEG	94.6078	1.11486	1.11486	
deviance	LASSO	30.2173	5.0415	0.92045	.000
	NEG	14.8934	0.57628	0.10521	

Table 4.27 Comparisonof time between the LASSO and NEG distribution in leukemia

		Mean
time	LASSO	4.4063
	NEG	0.1513

Table 4.26 showed result of paired t-test of predictive accuracy between the LASSO and NEG distribution. When compare between the LASSO and NEG distribution, the NEG distribution which has a sharp peak and a heavy tail and was controlled by type I error had higher accuracy than the LASSO. The result of the number of selected variable showed the NEG selected many variables into the model than the LASSO because of limitation of LASSO.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.27, time of the NEG running less than the LASSO. To compare performance of the NEG distribution and the LASSO in leukemia dataset found that performance of the NEG distribution overcomes the LASSO in accuracy, deviance the number of selected variables and time.

4.4.2 Lung cancer microarray dataset

4.4.2.1 Compare mean between the DE distribution and LASSO

Table 4.28 Paired t-test of number of selected variables, predictive accuracy, deviance between DE and LASSO in lung cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	DE	14.5667	1.22287	0.22326	.000
	LASSO	15.8667	0.68145	0.12441	
Predictive accuracy	DE	95.8837	0.6911	0.12764	.022
	LASSO	95.5034	0.31284	0.05712	
deviance	DE	24.7709	1.29015	0.23555	.064
	LASSO	24.1354	1.29659	0.23672	

Table 4.29 Comparisonof time between DE and LASSO in lung cancer

		Mean
time	DE	0.1153
	LASSO	5.8977

Table 4.28 showed result of paired t-test of predictive accuracy between DE and LASSO. It was believed that the LASSO is same as the DE. When compare between the DE and LASSO, the DE distribution which was controlled by type I error had slightly higher accuracy than the LASSO which used 10 fold cross validation. When compare deviances, result of deviance was not different in two methods. When consider table 4.29, time of DE running less than the LASSO but variables selection of DE selected the number of variables more than the LASSO. To compare performance of the LASSO and the DE distribution in lung dataset found that performance of the DE distribution overcome the LASSO in accuracy and time.

4.4.2.2 Compare mean between the LASSO and elastic net

Table 4.30 Paired t-test of number of selected variables, predictive accuracy, deviance between the LASSO and elastic net in lung cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	LASSO	15.8667	0.68145	0.12441	.000
	Elastic net	30.1667	2.58755	0.47242	
Predictive accuracy	LASSO	95.5034	0.31284	0.05712	.001
	Elastic net	97.3536	2.8445	0.51933	
deviance	LASSO	24.1354	1.29659	0.23672	.000
	Elastic net	16.2037	1.52238	0.27795	

Table 4.31 Comparison of time between the LASSO and elastic net in lung cancer

		Mean
time	LASSO	5.8977
	Elastic net	9.5763

Table 4.30 showed result of paired t-test of predictive accuracy between the LASSO and elastic net. It was proved that the elastic net suits for highly correlated data while the LASSO has limitations. When compare between the elastic net and LASSO, the elastic net which can variables grouping had higher accuracy than the LASSO which selects one variable from correlated group. The result agreed with the number of selected variable. The elastic net selected many variables into the model.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.31, time of the elastic net running more than the LASSO because of limitation of LASSO which selects variables at most n before it saturates. To compare performance of the LASSO and the elastic net in lung cancer dataset found that performance of the elastic net overcome the LASSO in accuracy and deviance.

4.4.2.3 Compare mean between the elastic net and NEG distribution

Table 4.32 Paired t-test of number of selected variables, predictive accuracy, deviance between the elastic net and NEG distribution in lung cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	Elastic net	30.1667	2.58755	0.47242	.000
	NEG	11.2333	2.11209	0.38561	
Predictive accuracy	Elastic net	97.3536	2.8445	0.51933	.001
	NEG	95.481	1.11377	0.20335	
deviance	Elastic net	16.2037	1.52238	0.27795	.000
	NEG	29.813	4.63718	0.84663	

Table 4.33 Comparison of time between the elastic net and NEG distribution in lung cancer

		Mean
time	Elastic net	9.5763
	NEG	0.1697

Table 4.32 showed result of paired t-test of predictive accuracy between the elastic net and NEG distribution. When compare between the elastic net and NEG distribution, the elastic net had higher accuracy than the NEG. The result of the number of selected variable showed the elastic net selected many variables into the model than the NEG distribution which have a shape peak at zero.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.33, time of the elastic net running more than the NEG distribution. To compare performance of the NEG distribution and the elastic net in lung cancer dataset found that performance of the elastic net overcomes the NEG in accuracy and deviance. The performance of the NEG overcomes the elastic net in the number of selected variables and time.

4.4.2.4 Compare mean between the DE and NEG distribution

Table 4.34 Paired t-test of number of selected variables, predictive accuracy, deviance between the DE and NEG distribution in lung cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	DE	14.5667	1.22287	0.22326	.000
	NEG	11.2333	2.11209	0.38561	
Predictive accuracy	DE	95.8837	0.69911	0.12764	.000
	NEG	95.481	1.11377	0.20335	
deviance	DE	24.7709	1.29015	0.23555	.000
	NEG	29.813	4.63718	4.84663	

Table 4.35 Comparison of time between the DE and NEG distribution in lung cancer

		Mean
time	DE	0.1153
	NEG	0.1697

Table 4.34 showed result of paired t-test of predictive accuracy between the DE and NEG distribution. When compare between the DE and NEG distribution, the DE distribution had higher accuracy than the NEG. The result of the number of selected variable showed the DE selected many variables into the model than the NEG distribution which have a sharper peak than DE distribution.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.35, time of the NEG running more than the DE distribution. To

compare performance of the NEG distribution and the DE in lung cancer dataset found that performance of the DE distribution overcomes the NEG in accuracy and deviance.

4.4.2.5 Compare mean between the LASSO and NEG distribution

Table 4.36 Paired t-test of number of selected variables, predictive accuracy, deviance between the LASSO and NEG distribution in lung cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	LASSO	15.8667	0.68145	0.12441	.000
	NEG	11.2333	2.11209	0.38561	
Predictive accuracy	LASSO	95.5034	0.31284	0.05712	.919
	NEG	95.481	1.11377	0.20335	
deviance	LASSO	24.1354	1.29659	0.23672	.000
	NEG	29.813	4.63718	0.84663	

Table 4.37 Comparison of time between the LASSO and NEG distribution

		Mean
time	LASSO	5.8977
	NEG	0.1697

Table 4.36 showed result of paired t-test of predictive accuracy between the LASSO and NEG distribution. When compare between the LASSO and NEG distribution, It was not different in accuracy. The result of the number of selected variable showed the NEG selected many variables into the model than the LASSO because of limitation of LASSO.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.37, time of the NEG running less than the LASSO. To compare performance of the NEG distribution and the LASSO in lung cancer dataset found that performance of the LASSO overcomes theNEG distribution in accuracy, deviance and the number of selected variables.

4.4.3 Prostate cancer microarray dataset

4.4.3.1 Compare mean between the DE distribution and LASSO

Table 4.38 Paired t-test of number of selected variables, predictive accuracy, deviance between DE and LASSO in prostate cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	DE	41.8333	1.5105	0.27578	.000
	LASSO	23.5333	1.73669	0.31707	
Predictive accuracy	DE	93.2353	1.57345	0.28727	.000
	LASSO	84.3137	3.02453	0.5522	
deviance	DE	107.4286	13.18242	2.40677	.000
	LASSO	39.7179	6.81663	1.24454	

Table 4.39 Comparison of time between DE and LASSO in prostate cancer

		Mean
time	DE	0.382
	LASSO	9.9907

Table 4.38 showed result of paired t-test of predictive accuracy between DE and LASSO. It was believed that the LASSO is same as the DE. When compare between the DE and LASSO, the DE distribution which was controlled by type I error had higher accuracy than the LASSO which used 10 fold cross validation. When compare deviances, result of deviance of the DE was not agree with accuracy. When consider table 4.39, time of DE running less than the LASSO but variables selection of DE selected the number of variables more than the LASSO. To compare performance of the LASSO and the DE distribution in prostate cancer dataset found that performance of the DE distribution overcome the LASSO in accuracy and time.

4.4.3.2 Compare mean between the LASSO and elastic net

Table 4.40 Paired t-test of number of selected variables, predictive accuracy, deviance between the LASSO and elastic net in prostate cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	LASSO	23.5333	1.73669	0.31707	.000
	Elastic net	27.8667	1.4077	0.25701	
Predictive accuracy	LASSO	84.3137	3.02453	0.5522	.016
	Elastic net	85.8823	3.57311	0.65236	
deviance	LASSO	39.7179	6.81663	1.24454	.000
	Elastic net	24.4835	1.39031	0.25384	

Table 4.41 Comparison of time between the LASSO and elastic net

		Mean
time	LASSO	9.9907
	Elastic net	12.5093

Table 4.40 showed result of paired t-test of predictive accuracy between the LASSO and elastic net. It was proved that the elastic net suits for highly correlated data while the LASSO has limitations. When compare between the elastic net and LASSO, the elastic net which can variables grouping had higher accuracy than the LASSO which selects one variable from correlated group. The result agreed with the number of selected variables. The elastic net selected many variables into the model.

When compare deviances, result of deviance agreed with accuracy. When consider table 4.41, time of the elastic net running more than the LASSO because of limitation of LASSO which selects variables at most n before it saturates. To compare performance of the LASSO and the elastic net in lung cancer dataset found that performance of the elastic net overcome the LASSO in accuracy and deviance.

4.4.3.3 Compare mean between the elastic net and NEG distribution

Table 4.42 Paired t-test of number of selected variables, predictive accuracy, deviance between the elastic net and NEG distribution in prostate cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	Elastic net	27.8667	1.4077	0.25701	.000
	NEG	35.3667	1.7711	0.32336	
Predictive accuracy	Elastic net	85.8823	3.57311	0.65236	.000
	NEG	93.7255	1.01687	0.18565	
deviance	Elastic net	24.4835	1.39031	0.25384	.000
	NEG	91.6954	16.33056	2.98154	

Table 4.43 Comparison of time between the elastic net and NEG distribution in prostate cancer

		Mean
time	Elastic net	12.5093
	NEG	0.3093

Table 4.42 showed result of paired t-test of predictive accuracy between the elastic net and NEG distribution. When compare between the elastic net and NEG distribution, the NEG distribution which was controlled type I error had higher

accuracy than the elastic net. The result of the number of selected variable showed the elastic net selected less variables into the model than the NEG distribution.

When compare deviances , result of deviance was not agree with accuracy. When consider table 4.43, time of the elastic net running more than the NEG distribution. To compare performance of the NEG distribution and the elastic net in prostate cancer dataset found that performance of the NEG distribution overcomes the elastic net in accuracy and time.

4.4.3.4 Compare mean between the DE and NEG distribution

Table 4.44 Paired t-test of number of selected variables, predictive accuracy, deviance between the DE and NEG distribution in prostate cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	DE	41.8333	1.5105	0.27578	.000
	NEG	35.3667	1.7711	0.32336	
Predictive accuracy	DE	93.2353	1.57345	0.28727	.134
	NEG	93.7255	1.01687	0.18565	
deviance	DE	107.42	13.18242	2.40677	.000
	NEG	91.6954	16.33056	2.98154	

Table 4.45 Comparison of time between the DE and NEG distribution

		Mean
time	DE	0.382
	NEG	0.3093

Table 4.44 showed result of paired t-test of predictive accuracy between the DE and NEG distribution. When compare between the DE and NEG distribution, the result was not different. The result of the number of selected variable showed the DE selected many variables into the model than the NEG distribution which have a sharper peak than DE distribution.

When compare deviances, the DE had higher than NEG. When consider table 4.45, time of the DE running more than the NEG distribution. To compare performance of the NEG distribution and the DE in prostate cancer dataset found that performance of the NEG distribution overcomes the DE in deviance and the number of selected variables.

4.4.3.5 Compare mean between the LASSO and NEG distribution

Table 4.46 Paired t-test of number of selected variables, predictive accuracy, deviance between the LASSO and NEG distribution in prostate cancer

		Mean	Standard Deviation	Standard Error Mean	Sig. (2-tailed)
Number of selected variables	LASSO	23.5333	1.73669	0.31707	.000
	NEG	35.3667	1.7711	0.32336	
Predictive accuracy	LASSO	84.3137	3.02453	0.5522	.000
	NEG	93.7255	1.01687	0.18565	
deviance	LASSO	39.7179	6.81663	1.24454	.000
	NEG	91.6954	16.33056	2.98154	

Table 4.47 Comparison of time between the LASSO and NEG distribution

		Mean
time	LASSO	9.9907
	NEG	0.3093

Table 4.46 showed result of paired t-test of predictive accuracy between the LASSO and NEG distribution. When compare between the LASSO and NEG distribution, the NEG had higher than the LASSO. The result of the number of selected variable showed the NEG selected many variables into the model than the LASSO because of limitation of LASSO.

When compare deviances, result of deviance was not agree with accuracy. When consider table 4.47, time of the NEG running less than the LASSO. To compare performance of the NEG distribution and the LASSO in prostate cancer dataset found that performance of the NEG overcomes theLASSOin accuracy and computational time.

4.5Limitation

Lasso is a constraint optimization method which was developed for solving model interpretation problem but still has some limitation about grouping variables. Elastic net was developed for solving the LASSO's limitation.The two methods were famous methods for genetic association study. NEG distribution is a Bayesian-inspire method which shrinks coefficients by setting normal exponential gamma prior.

4.5.1 LASSO

4.5.1.1 The number of selected variables of LASSA were limited by the number of samples. LASSO cannot select more variables than the number of samples because of the nature of convex optimization

4.5.1.2 LASSO cannot group variables which are correlated each other. This method was select one variable from group of correlated variables.

4.5.2 Elastic net

4.5.2.1 Elastic net can group variables which are correlated and selects the whole of correlated group into a model. This method selected the number of selected variable more than other methods. It is difficult to interpret model.

4.5.2.2 Elastic net uses cross validation. It uses computation time more than other methods.

4.5.3 NEG distribution

4.5.3.1 NEG distribution is a non-convex optimization. It would find a local optimum.

CHAPTER V
CONCLUSION

This study proposed the normal exponential gamma (NEG) distribution to classify microarray data. The typical problems of the kind of microarray data are overfitting and colinearlity. The standard logistic regression cannot make satisfied result. The LASSO and the elastic net are a constraint optimization which are famous in genetic association study. The candidate method, NEG distribution, is Bayesian-inspire method which determines the normal exponential gamma prior. The outstanding characteristics of NEG distribution are sharp peak at zero and heavy tail. It was interesting to do a comparative study between the famous methods and this method. The advantages of the NEG distribution are variable selection which had less the number of selected variables , shrinkage coefficients which little shrink in selected variables and using less time than the famous methods.

Although, the NEG distribution has many advantages, it still has some limitations. In some real data, the performance of the NEG distribution were not appreciate in overall. It was difficult to identify that what are the cause of this problem because using real data. In the future work, other microarray data may be used to testing performance.

REFERENCES

1 H. M. Kingston (2004).*ABC of Clinical Genetics*, Third edition, the BMJ Publishing Group, London, 2002. Bus SA, Ulbrecht JS, Cavanagh PR,*Pressure relief and load redistribution by custom-made insoles in diabetic patients with neuropathy and foot deformity*. Clinical Biomechanics, vol 19, 629-38.

2 R. Schleif.(1993).*Genetics and Molecular Biology*. Second Edition, the United States of America.

3 D. P. Berrar, W. Dubitzky, M. Granzow.(2003)*A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers, New York.

4 H. Hu, J. Li, A. Plank, H. Wang and G. Daggard.(2006).A Comparative Study of Classification Methods for Microarray Data Analysis. *Proceedings of the Fifth Australasian Data Mining Conference*. 33–37.

5 R. Tibshirani. Regression Shrinkage and Selection via the LASSO. (1996).*Journal of the Royal Statistical Society. Series B-Statistical Methodology*, vol. 58, no.1, 267–288.

6 G. James, D. Witten, T. Hastie and R. Tibshirani. (2013).*An Introduction to Statistical Learning with Applications in R*, New York. Springer.

7 S. Ma and J. Huang.(2008).Penalized Feature Selection and Classification in Bioinformatics. *Briefing in Bioinformatics*, vol. 9, no. 5, 392–403.

8 H. Zou and T. Hastie.(2005).Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B-Statistical Methodology*, vol. 67, part 2, 301–320.

9 K. L. Ayers and H. J.(2010). Cordell. SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression. *Genetic Epidemiology*, vol. 34, no. 8, 879–891.

10 K. P. Murphy. (2012).*Machine Learning: a Probabilistic Perspective*, USA. Cambridge, MA,MIT Press.

11 H. D. Bondel and B. J. Reich. (2008).Simultaneous Regression Shrinkage Variable Selection and Supervised Clustering of Predictors with OSCAR.*Biometrics,* vol. 64, 115–123.

12 C. J. Hoggart, J. C. Whittaker, M. Delorio and D. J. Balding.(2008).Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies.*PLoSGenetics,* vol. 4, no. 7.

APPENDICES

APPENDIX A

Publication

การศึกษาเปรียบเทียบการคัดเลือกตัวแปรในการจำแนกดีเอ็นเอไมโครอะเรย์

A Comparative Study on Variable Selection in DNA Microarray Classification

สิริกุล เหล่าศรีวิจิตร[1]โษฑศ์รัตต ธรรมบุษดี[1]สุภาภรณ์ เกียรติสิน[1] และวรัญญู วงษ์เสรี[2]

[1]สาขาวิชาเทคโนโลยีการจัดการระบบสารสนเทศ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยมหิดล
[2]ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
Email: cooly_kul@hotmail.com, sotarat.tha@mahidol.ac.th, tom_kiattsin@hotmail.com
waranyu.wongseree@gmail.com

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการลดขนาดสัมประสิทธิ์ของแบบจำลองการถดถอยโลจิสติกโดยอิงแนวคิดจากทฤษฎีของเบย์และใช้การแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติเป็นความน่าจะเป็นก่อนหน้าของสัมประสิทธิ์ของแบบจำลองในปัญหาการศึกษาดีเอ็นเอไมโครอะเรย์ ผลการเปรียบเทียบประสิทธิภาพในการทำนายและความสามารถในการเลือกยีนกับแลซโซและข่ายยืดหยุ่นในการจำแนกชนิดย่อยของมะเร็งเม็ดเลือดขาวพบว่าการใช้การแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติมีประสิทธิภาพในการทำนายและความสามารถในการคัดเลือกยีนสูงกว่าทั้งแลซโซและข่ายยืดหยุ่น

คำสำคัญการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติแลซโซข่ายยืดหยุ่นดีเอ็นเอไมโครอะเรย์

1. บทนำ

ดีเอ็นเอไมโครอะเรย์ (DNA Microarray) เป็นเทคโนโลยีสำหรับวัดระดับการแสดงออกของยีน (Gene Expression) จำนวนหลายพันยีนพร้อมกันในหนึ่งการทดลอง เซลล์ปกติและเซลล์ที่มีความผิดปกติ เช่น เซลล์มะเร็ง จะมีระดับการแสดงออกของยีนแต่ละยีนแตกต่างกัน [1] ข้อมูลชนิดนี้จึงนิยมนำไปใช้จำแนกการเกิดโรคออกจากภาวะปกติ จำแนกระยะของการเกิดโรค และจำแนกกลุ่มย่อยที่มีความสำคัญต่อการพยากรณ์และการรักษาโรค โดยนำข้อมูลการแสดงออกของยีนมาสร้างแบบจำลองการจำแนกและใช้แบบจำลองจำแนกข้อมูลในปัจจุบัน

วิธีที่นิยมใช้สร้างแบบจำลองการจำแนกคือการถดถอยโลจิสติกแต่การวิเคราะห์ดีเอ็นเอไมโครอะเรย์ที่มีจำนวนตัวแปรหรือยีนมากกว่าจำนวนตัวอย่างมีผลลัพธ์ไม่ดี เนื่องมาจากเหตุผลหลัก 2 ข้อ [2,3] คือ 1) ค่าความถูกต้องของการทำนาย เนื่องจากการถดถอยโลจิสติกจะสร้างแบบจำลองที่มีความเอนเอียงต่ำแต่ความแปรปรวนสูง นั่นก็คือแบบจำลองที่ฟิตเกิน (Overfitting) และ 2) การตีความแบบจำลอง (Model Interpretability) ชุดข้อมูลที่มีตัวแปรเป็นจำนวนมากจะต้องเลือกเซตย่อยที่มีผลกระทบสูงต่อการสร้างแบบจำลองออกมาเท่านั้น การเลือกแบบจำลอง (Model Selection)

เป็นวิธีสำหรับแก้ปัญหานี้โดยการลดความแปรปรวนของแบบจำลองที่นำไปสู่การเพิ่มความแม่นยำในการทำนาย[4]

การเลือกแบบจำลองสามารถแบ่งออกเป็น 3 วิธี [3]คือ 1) การเลือกเซตย่อย (Subset Selection) เป็นวิธีค้นหาเซตตัวแปรทำนาย 2) การลดขนาด (Shrinkage) เป็นวิธีปรับลดค่าสัมประสิทธิ์ของแบบจำลองและ 3) การลดมิติ (Dimension Reduction) เป็นการแปลงตัวแปรไปอยู่ในอีกมิติแล้วใช้ตัวแปรที่ถูกแปลงเพียงบางส่วนในการสร้างแบบจำลองการเลือกเซตย่อยสามารถแก้ปัญหาการตีความแบบจำลองได้ แต่เนื่องจากการค้นหาเซตย่อยเป็นกระบวนการแบบไม่ต่อเนื่อง (Discrete Process) ถ้ามีการเปลี่ยนแปลงข้อมูลเพียงเล็กน้อยจะส่งผลให้แบบจำลองที่ได้เปลี่ยนแปลงอย่างมากและทำให้ประสิทธิภาพของแบบจำลองลดลง [3]ส่วนการลดมิติถึงแม้จะช่วยแก้ปัญหาได้ แต่การแปลงตัวแปรทำให้ไม่สามารถระบุได้ว่าตัวแปรใดที่มีความสำคัญ [3]และการลดขนาดเป็นวิธีการเลือกตัวแปรโดยการลดขนาดของสัมประสิทธิ์ของตัวแปรที่ไม่มีความสำคัญให้มีค่าเข้าใกล้ศูนย์และลดขนาดของสัมประสิทธ์ของตัวแปรที่สำคัญเพียงเล็กน้อยหรือไม่ลดเลย จึงเป็นวิธีที่นิยมใช้ในการศึกษาหาความสัมพันธ์ของข้อมูลทางพันธุกรรมกับการเกิดโรค [5]

วิธีการลดขนาดสัมประสิทธิ์ของแบบจำลองสามารถทำได้โดยวิธีการหาค่าเหมาะสมแบบมีเงื่อนไข (Constrained Optimization) วิธีการดั้งเดิมของการลดขนาดสัมประสิทธิ์คือการถดถอยแบบริดจ์(Ridge Regression) [3]ที่เป็นกระบวนการแบบต่อเนื่อง (Continuous Process) ทำให้แบบจำลองมีเสถียรภาพกว่าการเลือกเซตย่อยแต่ค่าสัมประสิทธิ์ของแบบจำลองที่ได้จากการถดถอยแบบริดจ์ยังไม่เป็นค่าที่เข้าใกล้ศูนย์ จึงไม่สามารถแก้ปัญหาการตีความแบบจำลองได้ ต่อมามีการนำเสนอการหดค่าสัมบูรณ์น้อยสุดและตัวดำเนินการเลือก (Least Absolute Shrinkage and Selection Operator) หรือแลซโซ (LASSO) [2]ที่ลดขนาดสัมประสิทธิ์ของแบบจำลองด้วยกระบวนการแบบต่อเนื่องและเลือกตัวแปรไปพร้อมกัน โดยการรวมข้อดีของการเลือกเซตย่อยและการถดถอยแบบริดจ์มารวมไว้ด้วยกัน คุณสมบัติเด่นของแลซโซคือค่าสัมประสิทธิ์ของแบบจำลองมีโอกาสเป็นศูนย์ได้ [4] แต่แลซโซยังมีข้อจำกัดบางประการคือถ้าตัวแปรมีสหสัมพันธ์ระหว่างกันสูง แลซโซจะเลือกตัวแปรในกลุ่มนั้นมาเพียงตัวเดียว [6] โดยปกติข้อมูลดีเอ็นเอไมโครอะเรย์จะเป็นข้อมูลที่ยืนมีสหสัมพันธ์กันเอง จึงมีการพัฒนาวิธีที่เรียกว่าข่ายยืดหยุ่น (Elastic Net) [6] เป็นวิธีซึ่งผสมระหว่างแลซโซและริดจ์ที่มีความสามารถในการลดขนาดสัมประสิทธิ์ให้เข้าใกล้ศูนย์และจัดกลุ่มตัวแปรที่มีสหสัมพันธ์ระหว่างกันได้ พบว่าข่ายยืดหยุ่นมีประสิทธิภาพสูงกว่าแลซโซในปัญหาที่มีจำนวนตัวแปรมากกว่าจำนวนตัวอย่างมากและกลุ่มตัวแปรมีสหสัมพันธ์กันเอง นอกจากนี้ยังมีการพัฒนาแลซโซให้สามารถแก้ปัญหานี้ได้คือแลซโซเชิงกลุ่ม (GroupLASSO) [7]แต่ปัญหาคือจะต้องทราบโครงสร้างของตัวแปรกลุ่มนั้นก่อนถึงจะสามารถวิเคราะห์ข้อมูลได้เพื่อแก้ปัญหาดังกล่าวจึงมีการพัฒนาออสกา (Octagonal Shrinkage and Clustering Algorithm for Regression: OSCAR) [8]ที่สามารถเลือกตัวแปรและจัดกลุ่มตัวแปรโดยการแบ่งกลุ่มแบบมีผู้ฝึกสอน (Supervised Clustering) ตัวแปรที่มีความสำคัญโดยไม่จำเป็นต้องมีข้อมูลโครงสร้างของกลุ่มตัวแปร แต่พบว่าออสกาให้ผลลัพธ์ไม่แตกต่างกับข่ายยืดหยุ่น

นอกจากการลดขนาดด้วยวิธีการหาค่าเหมาะสมแบบมีเงื่อนไขแล้วยังสามารถทำได้โดยอิงแนวคิดจากทฤษฎีของเบย์โดยการกำหนดความน่าจะเป็นก่อนหน้าสำหรับสัมประสิทธิ์ของแบบจำลอง การแจกแจงความน่าจะเป็นที่ดีควรมียอดแคบและหางยาว การมียอดแคบจะเลือกตัวแปรออกมาน้อย และการมีหางยาวจะทำให้การลดขนาดสัมประสิทธ์ของตัวแปรที่ถูกเลือกอยู่ในแบบจำลองเกิดขึ้นเพียงเล็กน้อย [9]แลซโซได้รับการพิสูจน์แล้วว่าเป็นวิธีที่เหมือนกับการใช้การแจกแจงแบบเลขชี้กำลังคู่ (Double Exponential

Distribution: DE distribution) เป็นความน่าจะเป็นก่อนหน้าของสัมประสิทธิ์ของแบบจำลอง [2]มีการนำเสนอการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติ (Normal Exponential Gamma Distribution: NEG Distribution) [9] เป็นความน่าจะเป็นก่อนหน้าของสัมประสิทธิ์ของแบบจำลอง โดยการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติจะมียอดแคบและหางยาวกว่าการแจกแจงแบบเลขชี้กำลังคู่ จากการทดลองพบว่าการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติมีประสิทธิภาพในการเลือกตัวแปรที่สำคัญสูงกว่าการแจกแจงแบบเลขชี้กำลังคู่

ไม่นานมานี้มีการศึกษาเปรียบเทียบประสิทธิภาพการคัดเลือกสนิป (Single Nucleotide Polymorphism: SNP) หรือตัวแปรของการศึกษาความสัมพันธ์ทั้งจีโนม (Genome-Wide Association Study) ระหว่างริดจ์ แลซโซ ข่ายยืดหยุ่นและการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติ พบว่าการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติเป็นวิธีที่มีประสิทธิภาพในการเลือกตัวแปรดีที่สุด [5] แต่ยังไม่เคยมีการศึกษาเปรียบเทียบประสิทธิภาพกับข้อมูลดีเอ็นเอไมโครอะเรย์จึงเป็นที่มาของการศึกษานี้ โดยนำมาเปรียบเทียบกับวิธีที่นิยมใช้ในปัจจุบันคือแลซโซและข่ายยืดหยุ่น

2. ขั้นตอนวิธีที่นำเสนอ

แบบจำลองการถดถอยโลจิสติกมีสมการในรูปของ

$$\log \frac{\Pr(Y=1 \mid X)}{\Pr(Y=0 \mid X)} = X^T \beta$$

เมื่อ X คือ เมทริกซ์ของตัวแปรทำนายขนาด $n \times (p+1)$ (n และ p คือจำนวนตัวอย่างและจำนวนตัวแปรทำนายตามลำดับ) β คือเวกเตอร์ของสัมประสิทธิ์ของแบบจำลอง โดยปรกติใช้วิธีประมาณแบบความควรจะเป็นสูงสุด(Maximum Likelihood) สำหรับหาสัมประสิทธิ์ของแบบจำลองและมีสมการล็อกความควรจะเป็น (Log-Likelihood)

$$L(\beta) = \sum_{i=1}^{n} y_i B^T X_i - \log(1 - \exp(B^T X_i))$$

ปัญหาของการที่จำนวนตัวแปรมีมากกว่าจำนวนตัวอย่างและตัวแปรมีสหสัมพันธ์กันเองสูงคือการฟิตเกินและภาวะร่วมเส้นตรง(Collinearity)ส่งผลให้การประมาณค่าสัมประสิทธิ์ไม่มีเสถียรภาพ วิธีการหนึ่งในการแก้ปัญหานี้คือการประมาณแบบความน่าจะเป็นภายหลังสูงสุดที่ใช้ความน่าจะเป็นก่อนหน้าในการลดขนาดสัมประสิทธิ์ของแบบจำลอง ความน่าจะเป็นก่อนหน้าที่ใช้ในงานวิจัยนี้มีการแจกแจงแบบแกมมาเลขชี้กำลังปรกติดังสมการ

$$NEG(\beta \mid \lambda, \gamma) = k \exp\left(\frac{\beta^2}{4\gamma^2}\right) D_{-2\lambda-1}\left(\frac{|\beta|}{\gamma}\right)$$

เมื่อ λ และ γ คือ พารามิเตอร์บ่งรูปร่างและบ่งขนาดตามลำดับ k คือ ค่าคงที่ของการอินทิเกรทและ D คือ ฟังก์ชันทรงกระบอกเชิงพาราโบลา (Parabolic CylinderFunction) เมื่อ λ และ γ มีค่าเพิ่มขึ้นการแจกแจงจะลู่เข้าสู่การแจกแจงแบบเลขชี้กำลังคู่แต่การแจกแจงแบบแกมมาเลขชี้กำลังปรกติจากทฤษฎีของเบย์สามารถเขียนสมการล็อกความน่าจะเป็นภายหลังดังสมการ

$$\log p(\beta \mid X, y) = L(\beta) - f(\beta) + \text{const}$$

เมื่อ $f(\cdot)$ คือค่าติดลบของล็อกความน่าจะเป็นก่อนหน้า เครื่องหมายติดลบแสดงว่าฟังก์ชันนี้เป็นการทำโทษแบบจำลองที่มีความซับซ้อน และวิธีการประมาณค่านี้จะหาค่าสูงสุดของล็อกความควรจะเป็นที่ถูกลงโทษ (Penalized Log-Likelihood)เนื่องจากไม่มีผลเฉลยสำหรับการถดถอยโลจิสติกจึงประยุกต์ใช้วิธีนิวตันในการหาค่าพารามิเตอร์ของแบบจำลองทีละตัวดังสมการ

$$\beta_j^{new} = \beta_j - \frac{L'(\beta) - f'(\beta)}{L''(\beta) - f''(\beta)}$$

3. วิธีการดำเนินการวิจัย

ชุดข้อมูลในการศึกษานี้เป็นข้อมูลการแสดงออกของยีนจากการศึกษาไมโครอะเรย์ของผู้ป่วยมะเร็งเม็ดเลือดขาว(Leukemia) [10] ในการจำแนกยีนที่มีการแสดงออกแตกต่างกันใน 2 ชนิดย่อยของผู้ป่วยมะเร็งเม็ดเลือดขาวชนิดเฉียบพลันโดยชุดข้อมูลนี้จำแนกชนิดย่อยของมะเร็งเม็ดเลือดขาวออกเป็น 2 ชนิดคือ มะเร็งเม็ดเลือดขาวชนิดเฉียบพลันแบบลิมฟอยด์ (Acute Lymphoblastic Leukemia: ALL) และมะเร็งเม็ดเลือดขาวชนิดเฉียบพลันแบบไมอีลอยด์ (Acute Myeloid Leukemia: AML) ระดับการแสดงออกของยีนถูกวัดทั้งหมด 7,129 ยีน โดยแบ่งเป็นชุดข้อมูลฝึกสอนจำนวน 38 ตัวอย่างประกอบด้วยลิมฟอยด์ (ALL)จำนวน 27 ตัวอย่างและไมอีลอยด์ (AML)จำนวน 11 ตัวอย่าง และชุดข้อมูลทดสอบจำนวน 34 ตัวอย่างประกอบด้วยลิมฟอยด์ (ALL)จำนวน 20 ตัวอย่างและไมอีลอยด์ (AML)จำนวน 14 ตัวอย่าง

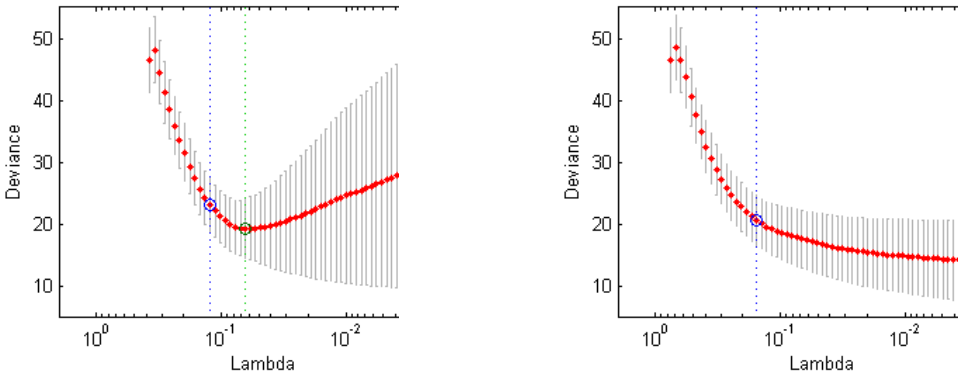
การทดลองจะเปรียบเทียบประสิทธิภาพการทำนายและความสามารถในการคัดเลือกยีนระหว่างวิธีการใช้การแจกแจงแบบแกมมาเลขชี้กำลังปรกติแลซโซและข่ายยืดหยุ่น แลซโซและข่ายยืดหยุ่นใช้วิธีการตรวจสอบความสมเหตุสมผลแบบไขว้ (Cross-Validation) กับข้อมูลเรียนรู้สำหรับหาค่าพารามิเตอร์ที่เหมาะสม โดยเลือกพารามิเตอร์ที่ให้แบบจำลองที่มีความซับซ้อนน้อยที่สุดที่มีค่าดิเวียนซ์ (deviance) หรือลบสองเท่าของค่าล็อกความควรจะเป็นแตกต่างจากแบบจำลองที่ดีที่สุดไม่เกิน 1 ค่าคลาดเคลื่อนมาตรฐาน (Standard Error) จากนั้นใช้พารามิเตอร์นี้สร้างแบบจำลองกับข้อมูลเรียนรู้ทั้งหมด และจำนวนตัวแปรที่ได้จากแบบจำลองคือจำนวนยีนที่แต่ละวิธีเลือก สุดท้ายนำแบบจำลองไปทดสอบกับข้อมูลทดสอบเพื่อวัดประสิทธิภาพการทำนาย ส่วนวิธีการใช้การแจกแจงแบบแกมมาเลขชี้กำลังปรกติจะเลือกพารามิเตอร์โดยการควบคุมค่าผิดพลาดแบบที่หนึ่ง (Type-I Error) ต่อตัวแปร โดยกำหนดให้มีค่าแอลฟาเท่ากับ 10^{-5}และแบบจำลองที่ดีที่สุดจะเลือกจากฐานนิยมของความน่าจะเป็นภายหลัง [9]

4. ผลการทดลอง

รายละเอียดการเลือกพารามิเตอร์ของแลซโซและข่ายยืดหยุ่นแสดงดังรูปที่ 1 เป็นกราฟแสดงความสัมพันธ์ระหว่างค่าดิเวียนซ์และพารามิเตอร์แลมป์ดาที่ได้จากวิธีการตรวจสอบความสมเหตุสมผลแบบไขว้ จุดตรงกลางคือค่าเฉลี่ยขอบเขตบนและล่างแสดงค่าคลาดเคลื่อนมาตรฐานจากค่าเฉลี่ย และเลือกพารามิเตอร์ที่ให้แบบจำลองที่มีความซับซ้อนน้อยที่สุด (เส้นประยาวทางด้านซ้าย) ที่มีค่าดิเวียนซ์แตกต่างจากแบบจำลองที่ดีที่สุด (เส้นประยาวทางด้านขวา) ไม่เกิน 1 ค่าคลาดเคลื่อนมาตรฐาน

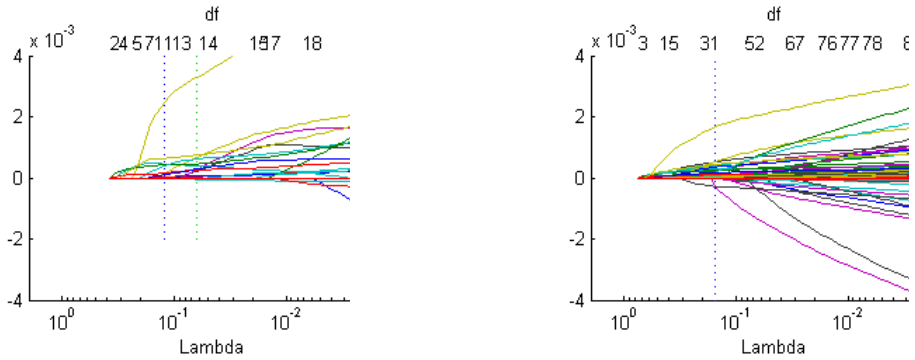
สัมประสิทธิ์ของแบบจำลองจากการสร้างแบบจำลองกับข้อมูลเรียนรู้ด้วยพารามิเตอร์ที่เลือกจากวิธีการตรวจสอบความสมเหตุสมผลแบบไขว้แสดงดังรูปที่ 2 แสดงความสัมพันธ์ระหว่างค่าสัมประสิทธิ์ของแบบจำลองและค่าพารามิเตอร์แลมป์ดา ตัวเลขด้านบนแสดงองศาอิสระ (Degree of Freedom) หรือจำนวนตัวแปรของแบบจำลอง เมื่อเปรียบเทียบการเลือกตัวแปรระหว่างสองวิธีพบว่าแลซโซเลือกตัวแปรออกมาน้อยกว่าข่ายยืดหยุ่นแต่สัมประสิทธิ์ของตัวแปรที่เลือกจะมีขนาดสูงกว่า

ผลการเปรียบเทียบประสิทธิภาพการทำนายและความสามารถในการคัดเลือกยีนระหว่างวิธีการใช้การแจกแจงแบบแกมมาเลขชี้กำลังปรกติแลซโซและข่ายยืดหยุ่นแสดงดังตารางที่ 1 พบว่าวิธีการใช้การแจกแจงแบบแกมมาเลขชี้กำลังปรกติมีประสิทธิภาพในการทำนาย 91.18% สูงกว่าทั้งแลซโซและข่ายยืดหยุ่นที่มีความถูกต้อง 82.35% และ 79.41% ตามลำดับ นอกจากนี้ยังมีความสามารถในการคัดเลือกยีนที่จำเป็นในการสร้างแบบจำลองสูงกว่าทั้งสองวิธีด้วย คือสามารถคัดเลือกยีนที่ใช้ในการจำแนกชนิดย่อยของมะเร็งเม็ดเลือดขาวออกมาเพียง 10 ยีน ในขณะที่แลซโซและข่ายยืดหยุ่นเลือกออกมา 11 และ 33 ยีนตามลำดับ



(ก) (ข)

รูปที่ 1ความสัมพันธ์ระหว่างค่าดิเวียนซ์และค่าแลมป์ดาของ(ก) แลซโซและ(ข) ข่ายยืดหยุ่น



(ก) (ข)

รูปที่ 2 ความสัมพันธ์ระหว่างสัมประสิทธิ์ของตัวแปรและค่าแลมป์ดาของ(ก) แลซโซและ(ข) ข่ายยืดหยุ่น

ตารางที่ 1จำนวนยีนที่ถูกเลือกและความถูกต้องของการทำนายระหว่างแลซโซ ข่ายยืดหยุ่นและวิธีการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติ (NEG)

วิธี	จำนวนยีน	ความถูกต้อง (%)
แลซโซ	11	79.41
ข่ายยืดหยุ่น	33	82.35
NEG	10	91.18

5. สรุปผลการทดลองและวิจารณ์ผล

เนื่องจากข้อมูลดีเอ็นเอไมโครอะเรย์เป็นข้อมูลที่วัดระดับการแสดงออกของยีน ซึ่งยีนส่วนใหญ่มาจากวิถี (Pathway) เดียวกันทำให้ตัวแปรมีสหสัมพันธ์ระหว่างกันสูง [6]แลซโซจึงไม่มีประสิทธิภาพในการแก้ปัญหา ขณะที่ข่ายยืดหยุ่นมีความสามารถในการแก้ปัญหานี้ได้ดีกว่าแลซโซ แต่เมื่อเปรียบเทียบกับวิธีการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติพบว่ามีประสิทธิภาพต่ำกว่าทั้งความถูกต้องในการทำนายและความสามารถในการคัดเลือกยีน เนื่องจากทั้งแลซโซและข่ายยืดหยุ่นเป็นการแก้ปัญหาค่าเหมาะสมแบบคอนเวกซ์ (Convex Optimization Problem) ถ้าพิจารณาในรูปของความ

น่าจะเป็นก่อนหน้าจะมียอดที่กว้างและหางสั้นกว่า วิธีการแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติ เป็นผลให้จำนวนตัวแปรที่ถูกเลือกจะมีสูงกว่าและยังมีค่าสัมประสิทธิ์ในแต่ละตัวแปรต่ำกว่าด้วย ซึ่งสอดคล้องกับการศึกษาก่อนหน้านี้ [5,7]

เอกสารอ้างอิง

[1] H. Hu, J. Li, A. Plank, H. Wang and G. Daggard,"A Comparative Study of Classification Methods for Microarray Data Analysis,"*Proceedings of the Fifth Australasian Data Mining Conference*, pp.33–37,Dec. 2006.

[2] R. Tibshirani,"Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society. Series B-Statistical Methodology*, vol. 58, no.1, pp. 267–288, 1996.

[3] G. James, D. Witten, T. Hastie and R. Tibshirani,*An Introduction to Statistical Learning with Applications in R*, New York. Springer,2013.

[4] S. Ma and J. Huang,"Penalized Feature Selection and Classification in Bioinformatics,"*Briefing in Bioinformatics*, vol. 9, no. 5, pp. 392–403, Sep. 2008.

[5] K. L. Ayers and H. J. Cordell,"SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression,"*Genetic Epidemiology,* vol. 34, no. 8, pp. 879–891, Dec. 2010.

[6] H. Zou and T. Hastie,"Regularization and Variable Selection via the Elastic Net,"*Journal of the Royal Statistical Society. Series B-Statistical Methodology*, vol. 67, part 2, pp.301–320, 2005.

[7] K. P. Murphy,*Machine Learning: a Probabilistic Perspective*,USA. Cambridge, MA,MIT Press, 2012.

[8] H. D. Bondel and B. J. Reich,"Simultaneous Regression Shrinkage Variable Selection and Supervised Clustering of Predictors with OSCAR,"*Biometrics,* vol. 64, pp. 115–123, Mar. 2008.

[9] C. J. Hoggart, J. C. Whittaker, M. Delorio and D. J. Balding,"Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies,"*PLoSGenetics*, vol. 4, no. 7, Jul. 2008.

[10] T. R. Golub ,D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek,J. P. Mesirov, et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,"*Science,* vol. 286, no. 5439, pp. 531–537, Oct. 1999.

APPENDIX B

Figure of results

1. Leukemia data

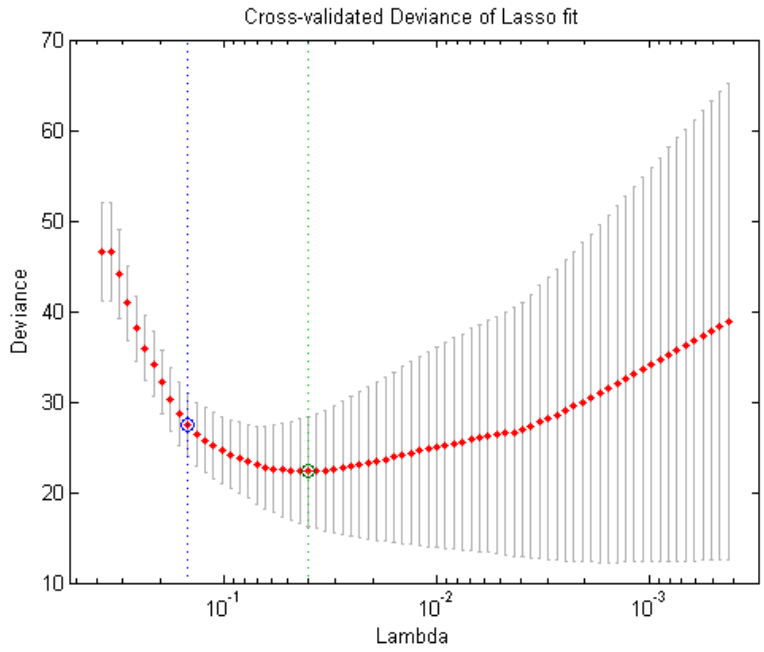


Figure 7.1 Cross –validated deviance of lasso fit in leukemia

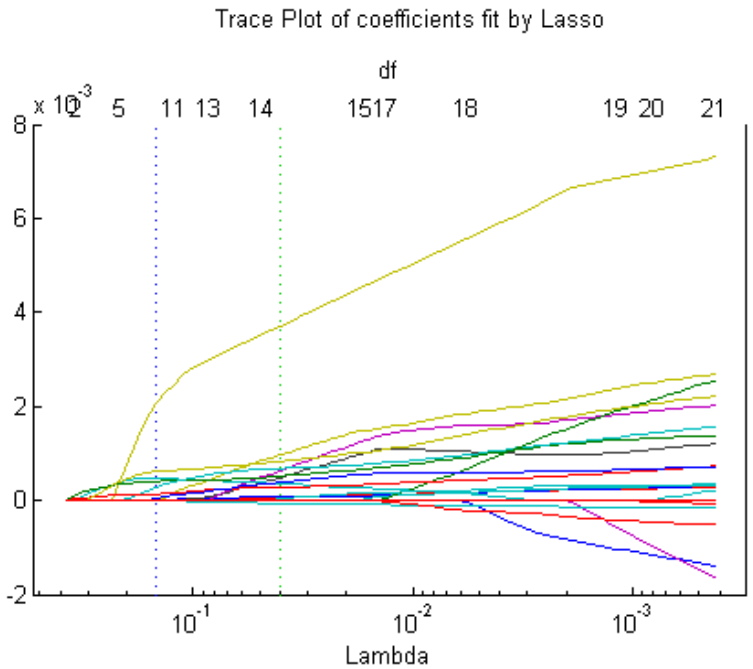


Figure 7.2 Trace plot of coefficients fit by LASSO in leukemia

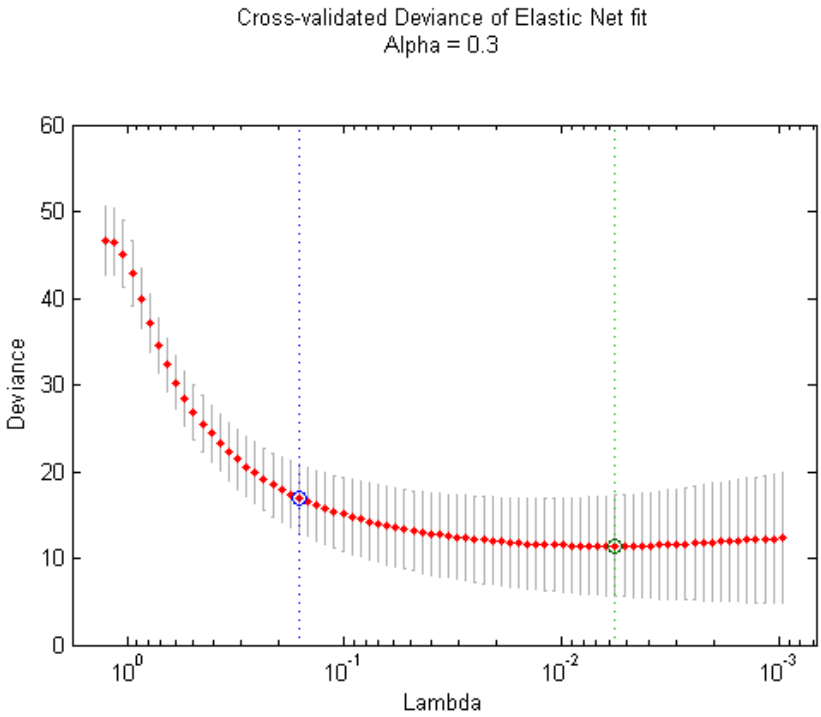


Figure 7.3 Cross –validated deviance of elastic net fit in leukemia

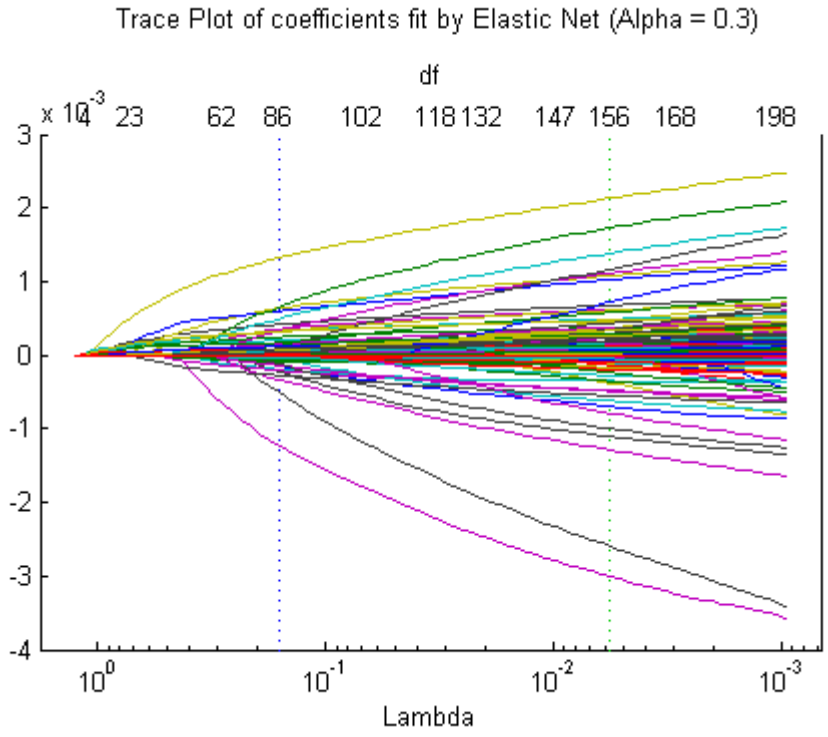


Figure 7.4 Trace plot of coefficients fit by elastic net in leukemia

2. Lung cancer

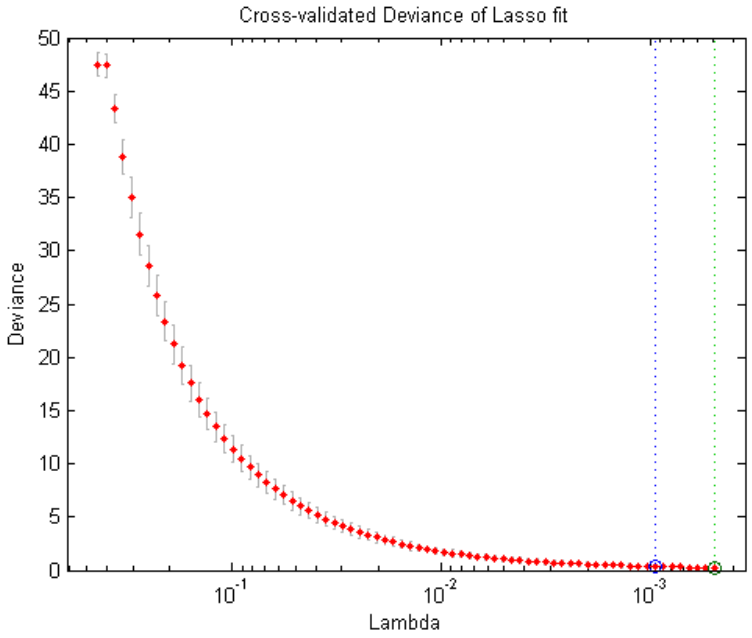


Figure 7.5 Cross –validated deviance of lasso fit in lung cancer

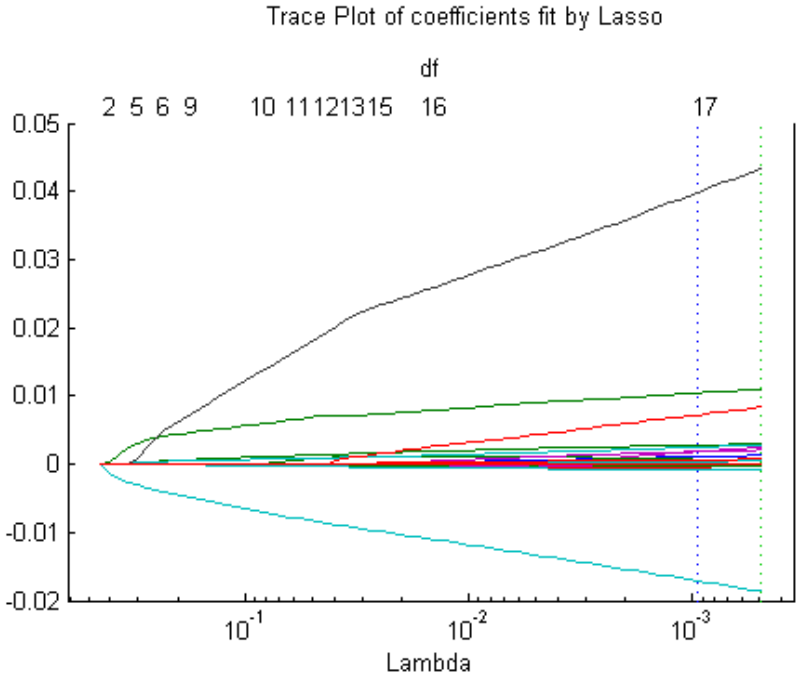


Figure 7.6 Trace plot of coefficients fit by LASSO in lung cancer

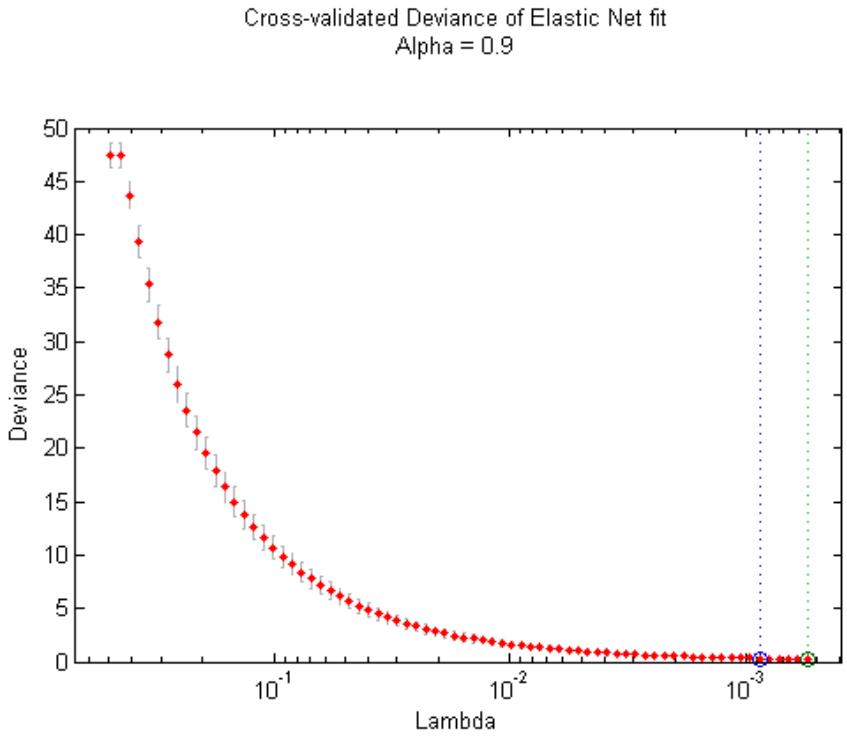


Figure 7.7 Cross –validated deviance of elastic net fit in lung cancer

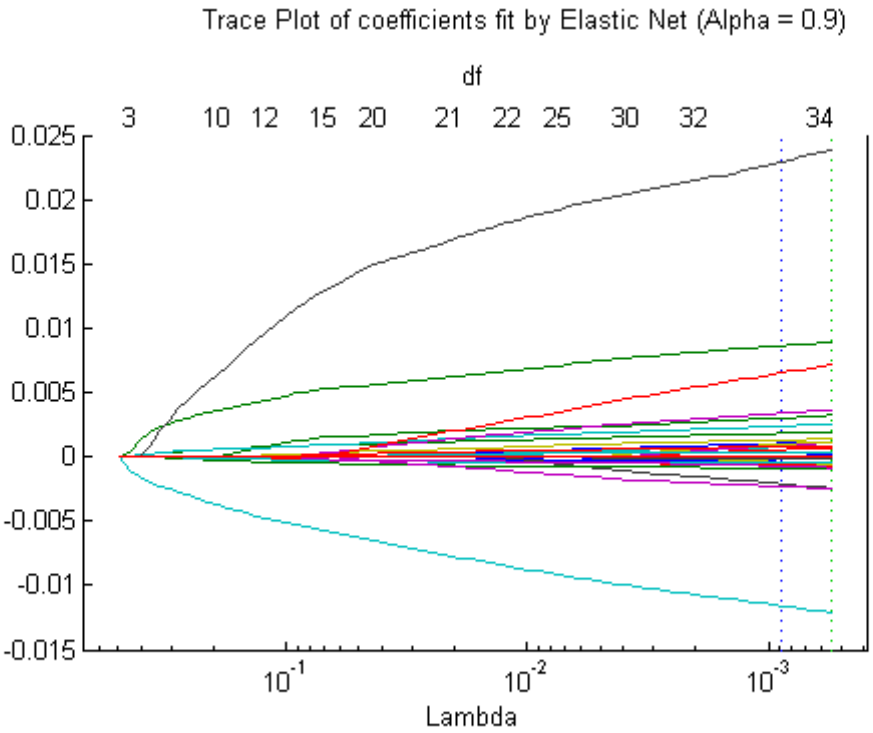


Figure 7.8 Trace plot of coefficients fit by elastic net in lung cancer

3. Prostate cancer

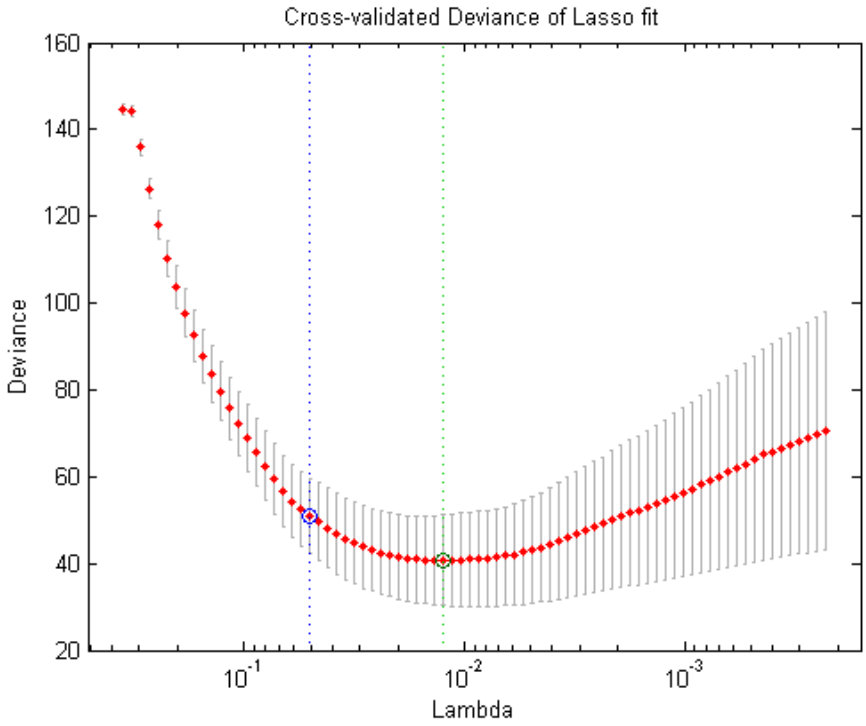


Figure 7.9 Cross –validated deviance of lasso fit in prostate cancer

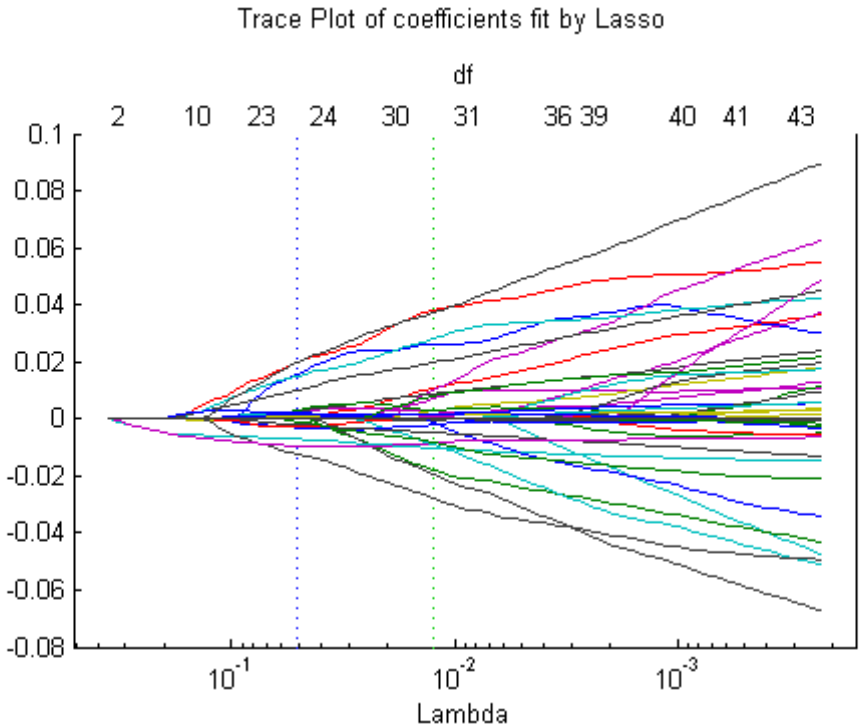


Figure 7.10 Trace plot of coefficients fit by LASSO in prostate cancer

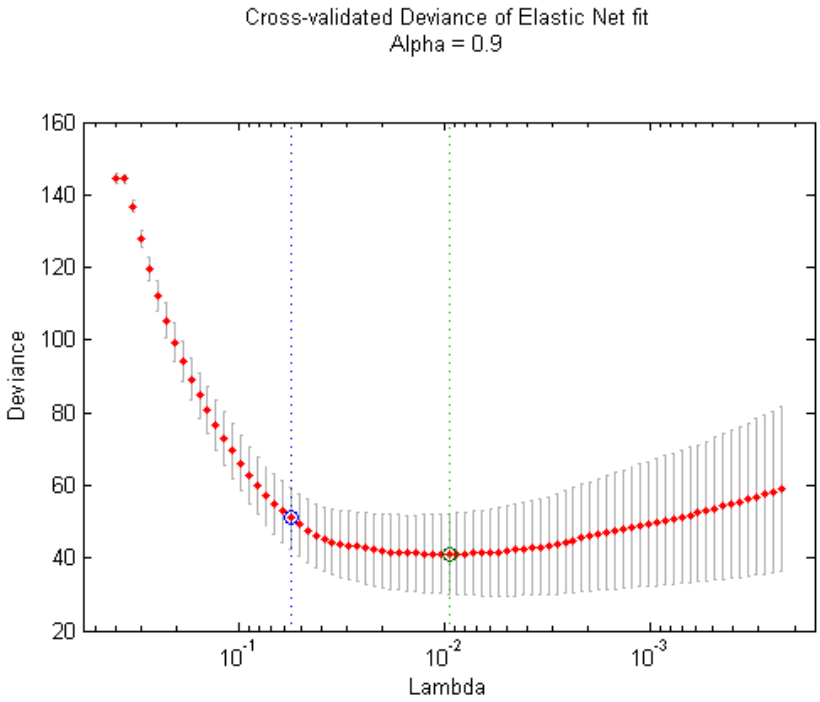


Figure 7.11 Cross –validated deviance of elastic net fit in prostate cancer

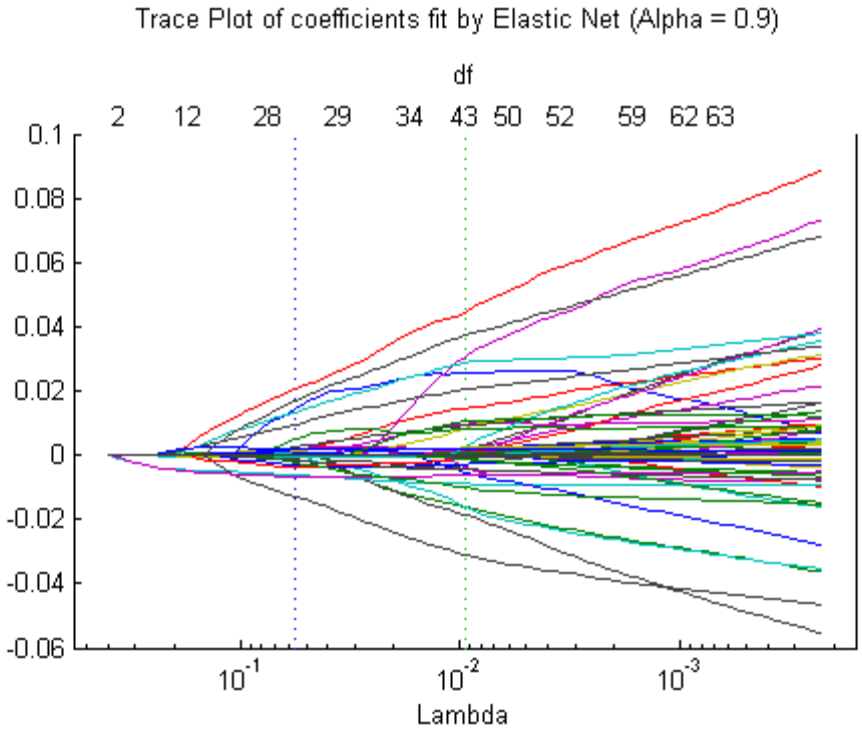


Figure 7.12 Trace plot of coefficients fit by elastic net in prostate cancer

BIOGRAPHY

NAME	MissSirikul Laosrivichit
DATE OF BIRTH	4April 1984
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Mahidol University, 2002-2006 Bachelor of Science (Medical Technology) Mahidol University, 2012-2014 Master of Science (Technology of Information System Management**)**
HOME ADDRESS	341/41Sukhumvit 101/1 Bangjak Phakanong Bangkok10260 Tel 081-889-0710 Email: cooly_kul@hotmail.com
PUBLICATION / PRESENTATION	Sirikul Laosrivichit, Sotarat Thammaboosadee, Supaporn Kiattisin, Waranyu Wongseree, The 6th National Conference on Information Technology :NCIT) Thailand, 27-28February2014