MODEL SELECTION METHODS AND THEIR APPLICATIONS IN GENOME-WIDE ASSOCIATION STUDIES

ZHAO JINGYUAN

(Master of Statistics, Northeast Normal University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2008

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor, Associate Professor Chen Zehua for his invaluable advice and guidance, endless patience, kindness and encouragement. I truly appreciate all the time and effort he has spent in helping me to solve the problems I encountered. I have learned many things from him, especially regarding academic research and character building.

I wish to express my sincere gratitude and appreciation to Professor Bai Zhidong for his continuous encouragement and support. I am grateful to Associate Professor Chua Ting Chiu for his timely help. I also appreciate other members and staff of the department for their help in various ways and providing such a pleasant working environment, especially to Ms Yvonne Chow and Mr Zhang Rong for the advice and assistance in computing.

It is a great pleasure to record my thanks to my dear friends: to Ms Wang Keyan, Ms Zhang Rongli, Ms Hao Ying, Ms Wang Xiaoying, Ms Zhao Wanting, Mr Wang Xiping

who have given me much help in my study and life. Sincere thanks to all my friends who helped me in one way or another and for taking caring of me and encouraging me.

Finally, I would like to give my special thanks to my parents for their support and encouragement. I thank my husband for his love and understanding. I also thank my baby for giving me courage and happiness.

Contents

Summary

High dimensional feature selection frequently appears in many areas of contemporary statistics. In this thesis, we propose a high dimensional feature selection method in the context of generalized linear models and apply it in genome-wide association studies. Moreover, the modified SCAD method is developed and the family of extended Bayesian information criteria is discussed in generalized linear models.

In the first part of the thesis, we propose penalizing the original smoothly clipped absoulte deviation (SCAD) penalized likelihood function with the Jeffreys prior for producing finite estimates in case of separation. The SCAD method is a variable selection method with many favorable theoretical properties. However, in case of separation, at least one SCAD estimate tends to infinity and hence the SCAD method cannot work normally. We show that the modification of adding the Jeffreys penalty to the original penalized likelihood function always yields reasonable estimates and maintains the good performance of the SCAD method.

In the second part, we study the family of extended Bayesian information criteria (EBIC) (Chen and Chen, 2008), focusing on its performance of feature selection in the context of generalized linear models with main effects and interactions. There are a variety of model selection criteria such as Akaike information criterion (AIC), Bayesian information criterion (BIC). However, these criteria fail when the dimension of feature space is high. We extend EBIC to generalized linear models with main effects and interactions by deducing different penalties on the number of main effects and the number of interactions.

In the third part, we introduce the generalized tournament screening cum EBIC approach for high dimensional feature selection in the context of generalized linear models. The generalized tournament approach can tackle both main effects and interaction effects, and it is computationally feasible even if the dimension of feature space is ultra high. In addition, one of its characteristics is that the generalized tournament approach jointly evaluates the significance of features, which could improve the selection accuracy.

In the final part, we apply the generalized tournament screening cum EBIC approach to detect genetic variants associated with some common diseases by assessing main effects and interactions. Genome-wide association studies is a hot topic in the genetic study. Empirical evidence suggests that interaction among loci may be responsible for many diseases. Thus, there is a great demand for statistical approaches to identify the

causative genes with interaction structures. The performances of the generalized tournament approach and the multiple testing method (Marchini *et al*., 2005) are compared by some simulation studies. It is shown that the generalized tournament approach not only improve the power for detecting genetic variants but also controls the false discovery rate.

List of Tables

Chapter 1

Introduction

As high dimensional data frequently arise from a variety of areas, feature selection with high dimensional feature space has become a common and imminent problem in contemporary statistics. Genome-wide association studies for identification of multiple loci influenced diseases belong to high dimensional feature selection problem. In this problem, the dimension of the feature space (P) is much larger than the sample size (n), which poses severe challenges to feature selection. Feature selection can be considered as a special case of model selection. However, for such a situation as genome-wide association studies, where the dimension of the feature space is ultra high, it is impossible to implement conventional model selection methods to select causal features. Dimension reduction is an effective strategy to deal with feature selection with high dimensional feature space. On the basis of dimension reduction, some studies are appearing to tackle high dimensional feature selection in the context of linear models.

Besides linear models, other generalized linear models built in high dimensional data are also widely applied in many areas. Thus, it is important to investigate high dimensional feature selection in generalized linear models. In addition, it is common that interaction effects are prominent in explaining the response variable. Hence, it is necessary for high dimensional feature selection methods to consider both main effects and interaction effects.

In the following sections, background and literatures related to high dimensional feature selection are reviewed in more details. In Section 1.1, some background of high dimensional feature selection is introduced. In Section 1.2, a topic related to feature selection, model selection is introduced. In Section 1.3, a huge number of literatures about feature selection methods and model selection methods are reviewed. The aim and organization of this thesis are given in Section 1.4.

1.1 Feature selection with high dimensional feature space

With the development of technologies, the collection of high dimensional data becomes feasible commercially. High dimensional data frequently appear in areas such as finance, signal processing, genetics and geology. For example, data from genome-wide association studies contains hundreds of thousands of genetic markers, e.g., single nucleotide polymorphisms (SNPs), which are screened to provide information for identification of causal loci. In these high dimensional data, not all but only a small subset of

features contribute to the response variable, so it is necessary and critical to eliminate irrelevant and redundant features from data. Feature selection with high dimensional feature space has received much attention in contemporary statistics. For high dimensional data, one common characteristic is that the number of candidate features P is much larger than the sample size n, which is the so-called small-n-large-P problem. It is challenging to detect a few causal features from a huge number of candidates to explain the response variable, with a relatively small sample size.

In feature selection with high dimensional feature space, one challenge posed by small-n-large-P problem is that a few casual features mix with a huge number of non-causal features. Another challenge is that the maximal spurious correlation between casual features and non-causal features can be high and usually increase with the dimensionality of feature space, even if all features in population are stochastically independent. If a highly spurious correlation between a casual feature and a non-causal feature exists, this non-causal feature could present a high correlation with the response variable. Thus, it is hard to select truly causal features when the dimension P is large.

Such a problem has become especially prevalent in genome-wide association studies. A genome-wide association study (GWAS) is a promising way to detect genetic variants responsible for some diseases, particularly common complex diseases such as cancer, diabetes, heart disease and mental illnesses. After a new genetic association is identified, it can be employed to develop better strategies to treat and prevent the disease.

In comparison with other approaches for mapping genetic variants, genome-wide association studies need to utilize genotypes of hundreds of thousands of SNPs for human samples. Fortunately, with the advent of high-throughput biotechnologies, a rapid collection of genotypes of densely spaced SNPs throughout the whole genome is becoming the norm, which moves genome-wide associate studies from the futuristic to the realistic. In fact, in these tens or hundreds of thousands of SNPs, there are only a few that contribute to the disease. Thus, the task of genome-wide studies is to detect the genetic variants of common diseases from a huge number of SNPs with a relatively small number of human samples. This is an example of the small-n-large-P problem mentioned above.

In genome-wide association studies, some statistical methods have been developed to detect a few loci involving tens or hundreds of thousands of loci and comparably few observations, see, e.g. Efron adn Tibishirani (2002); Sabatti *et al*. (2003) Storey and Tibshirani (2003); Lin *et al*. (2004); Marchini *et al*. (2005); Lowe, C.E. *et al*. (2004). Most of these methods are based on multiple testing. Besides them, some methods incorporating feature selection into model selection were proposed and applied to tackle genome-wide association studies, so we introduce model selection and its developments in the next section.

1.2 Model selection

A linear regression model is given as follows:

$$Y = \beta_0\mathbf{1} + \beta_1 X_1 + \ldots + \beta_P X_P + \varepsilon = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \tag{1.1}$$

where Y is an $n \times 1$ vector, $\mathbf{X} = (\mathbf{1}, X_1, X_2, \ldots, X_P)$ is an $n \times (P+1)$ matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)^T$ is a $(P+1)$ vector of unknown parameters, and ε follows the distribution with mean 0 and variance matrix $\sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. In the linear model (1.1), the design matrix $\mathbf{X}$ affects the distribution of Y through the linear function $\eta(\mathbf{X}) = \beta_0\mathbf{1} + \beta_1 X_1 + \ldots + \beta_P X_P$, which is equal to the expectation of Y.

A generalized linear model is a generalization of the linear regression model given above. Generalized linear models are considered as a way of unifying statistical models, including linear regression model, logistic regression model and Poisson regression model. In a generalized linear model, there are three parts: a random part, a deterministic part and a link function. The random part is the assumption that the response variable Y follows an exponential family distribution. An exponential family is characterized by a probability density function f given by

$$f(y, \theta, \phi) = \exp\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(\phi, y)\} I_A(y),$$

where the set A does not depend on θ (canonical parameter) and ϕ (dispersion parameter). A large class of probability distributions including normal, binomial and Poisson distributions belong to the exponential family. The deterministic part is the assumption

that the covariates affect Y through a linear predictor $\eta(\mathbf{X}) = \beta_0\mathbf{1} + \beta_1 X_1 + ... + \beta_P X_P$. A generalized linear model relates the random part to the deterministic part through a function called the link function: $g(E(Y|X)) = \eta = X\beta$, where $E(Y|X)$ is the conditional expectation of Y given X. The link function provides the relationship between the linear predictor and the mean of the distribution function.

At the beginning of a given modeling problem, a large number of potential covariates are available, but not all of these contribute to the response variable. Some of these may have little or no contribution to the response variable. Model selection, a critical issue in data analysis, is the task of selecting a statistical model from a set of potential models according to some criterion. A model with redundant covariates may result in a better fit with less bias, but suffer high variance and lead to poor prediction performance. Thus, it is necessary to obtain a model which contains as few covariates as possible while still maintains good prediction property. There are a huge number of literatures about model selection methods. Model selection methods can be divided into three classes: classical methods such as forward, backward and stepwise regression, all-subset selection, and the penalized likelihood methodology, see, e.g. Breiman (1995), Tibshirani (1996), Fan and Li (2001) and Efron *et al*. (2004) and Park *et al*. (2006).

Feature selection can be considered as a special case of model selection. The difference is that feature selection only focuses on detecting casual features, whereas the task of model selection focuses on the prediction accuracy of the model. In principle, model

selection procedures mentioned above can be used to detect causal features, but when the dimension P is huge, they will fail for one reason or another. Some studies (Chen and Chen, 2007; Fan and Lv, 2008)have pointed out that dimension reduction is an effective strategy to deal with high dimensionality. When the dimension is reduced to a low level, conventional model selection methods can be implemented to detect causal features. Motivated by this idea, some feature selection procedures have been advocated in the context of linear model with high dimensional data, see Chen and Chen (2007), Fan and Lv (2008). When the purpose is to select a model with good prediction properties, the cross-validation (CV) score, which is an approximation to the prediction error, is an appropriate criterion. CV does not care whether or not the features in the model are causal as long as the model has the best prediction accuracy. However, feature selection focuses on detecting causal features and the accuracy of the selection. Other criteria should be used. Unfortunately, it has been demonstrated in many applications that, when the dimension of the feature space is high, the conventional model selection criteria such as AIC, BIC, etc. fail their functionality. To deal with the difficulty caused by the high dimensionality of the feature space, a family of extended Bayes information criteria (EBIC) has recently been developed by Chen and Chen (2008).

1.3 Literature review

In this section, some feature selection methods are reviewed. We first review some feature selection methods confined to genome-wide association studies in Subsection

1.3.1. Model selection methods and some feature selection methods incorporated into model selection are reviewed in Subsection 1.3.2.

1.3.1 Feature selection methods in genome-wide association studies

In genome-wide association studies, a large number of statistical studies have been developed to detect genetic variants associated with a particular disease. From the point of view of genetics, these approaches can be divided into three categories: single marker analysis, haplotype analysis and gene-gene interaction analysis.

Single marker analysis is based on multiple testing of all possible individual SNPs. In genome-wide association studies, the number of hypothesis tests is equal to the number of SNPs under consideration which can reach hundreds of thousands. An important issue in multiple tests is how to control the overall type I error. Klein *et al*. (2005) used Bonferroni adjustment for the critical value to declare the significance in genome-wide association studies. Instead of Bonferroni correction, the false discovery rate (FDR) was presented by Benjamini and Hochberg (1995), and employed by Efron and Tibshirani (2002) and Storey and Tibshirani (2003). The false discovery rate was expected to be more appropriate than Bonferroni correction, but when too many hypothesis tests are conducted in genome-wide association studies, it is still unsatisfactory. Some other studies on multiple tests were developed in the recent past. Helgadottir *et al*. (2007) suggested to explore SNPs with the lowest p-values. Hoh and Ott (2003) advocated to

utilize the sum-statistics to avoid the multiple testing dilemma.

Many studies (Allen and Satten, 2007) support the idea that the analysis based on haplotype can be more powerful than single marker analysis. Lin *et al*. (2004) employed the multiple testing of haplotype association over all possible windows of segments, using permutation approach as multiple testing adjustment. Besides, another area on the basis of haplotypes focuses on testing untyped variants by coupling typed SNPs with external information from datasets describing linkage disequilibrium (LD) patterns across the genome (Abecasis, 2007; Epstein, Allen and Satten, 2007; Marchini *et al*., 2007; Servin and Stephens, 2007).

These two kinds of approaches proceed by testing single genetic marker or haplotype individually, but many empirical evidence suggests that interactions among loci may affect many common complex diseases (Zerba, K. E., 2000). Marchini *et al*. (2005) proposed to utilize the multiple testing of all possible pairwise gene-gene interactions to detect genetic variations related to a common complex disease. Log-likelihood ratio tests for each full logistic regression model with case-control data were used. The overall threshold to control overall type I error was suggested to be addressed by Bonferroni correction. One advantage of this method is that it is computationally feasible to undertake in genome-wide association studies given a large computer cluster. Another advantage is that it has greater power for identifying genetic variants in comparison with traditional single marker analyses. However, since Bonferroni correction is so

conservative that an extremely small $p-$value is needed to declare the genome-wide significance, the power to identify genetic variants would be still low. Moreover, some non-causal variations may be wrongly detected since the multiple testing may declare some interactions between non-causal and causal variants to be significance.

The interest of these feature selection methods was confined to genome-wide association studies. Moreover, methods based on multiple testing have a lot of limitations such as ignoring multi-feature joint effects. Recently, some studies focused on incorporating feature selection into model selection. The next subsection will review conventional model selection methods, as well as feature selection methods in high dimensional space.

1.3.2 Model selection methods

As model selection is an important issue in modern data analysis, a large number of model selection methods were proposed. They can be classified into three categories: classical methods such as forward, backward and stepwise selection; all-subset selection methods, AIC and BIC; the penalized likelihood methods including non-negative garrote, least absolute shrinkage and selection operator (LASSO) and SCAD.

Forward, backward selection methods select variables by adding or deleting one at a time based on reducing the sum-square-error. Stepwise selection by Efroymson (1960)

is a combination of forward and backward selection. The backward selection is not suitable to the situation where the number of covariates is much larger than the sample size. Moreover, both forward and stepwise selections suffer a serious drawback from their greedy property.

All-subset selection examines all possible sub-models and picks the best model by optimizing some selection criteria. Although all-subset selection methods are easy to use in practice, they have several drawbacks. One main drawback is that all-subset selection methods are the most unstable procedure (Breiman, 1996). Moreover, all-subsets procedure is impracticable in terms of computational cost when the number of independent covariates is large.

In recent years, researchers have proposed a new class of model selection methods. They include the non-negative garrote by Breiman (1995), the LASSO by Tibshirani (1996), the least angle regression (LARS) by Efron et al. (2004), Elastic Net by Zou and Hastie (2005), the adaptive Lasso by Zou (2006) and the SCAD by Fan and Li(2001). Generally speaking, these methods estimate the unknown parameters by minimizing a penalized sum of squares of residuals in linear model. They can perform the parameter estimation and variable selection simultaneously. In the following, we review penalized likelihood methods in the context of linear model.

Breiman introduced the non-negative garrote method in 1995. The garrote starts with

the ordinary least squares estimates of the full model and then shrinks them by non-negative factors whose sum is constrained. The garrote estimates can be obtained by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{P} c_j \hat{\beta}_j \, x_{ij})^2 \text{ subject to } c_j \geq 0, \ \sum_{j=1}^{P} c_j \leq t, \tag{1.2}$$

where $\hat{\beta}_j, \ j = 1, \ldots, P$ are the ordinary least squares estimates. The non-negative garrote method enjoys consistently lower prediction error than all-subset selection and is competitive with ridge regression except when the true model contains many small non-zero coefficients. However, the garrote estimates depend on both the sign and the magnitude of the ordinary least squares estimates. Moreover, when there are highly correlated covariates, the ordinary least squares estimates behave poorly, which may affect the garrote estimates.

Motivated by the idea of non-negative garrote method, Tibshirani (1996) proposed a new method via the L_1 penalty, called the Lasso, for "least absolute shrinkage and selection operator". In Lasso, the parameter estimates are obtained by minimizing the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. The Lasso penalized estimators are obtained by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{P} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{P} |\beta_j| \tag{1.3}$$

In (1.3), $\lambda \geq 0$ is a tuning parameter controlling the amount of shrinkage to ordinary least squares estimates. For a small value of λ, the solution of (1.3) approaches the ordinary least squares estimates; while for sufficiently large λ, some of the parameter

estimates will be exactly zero. Efron *et al*. (2004) proposed a sequential variable selection algorithm also via the L_1 penalty, called Least Angle Regression (LARS) which is useful and less greedy than forward selection method. The procedure in the LARS algorithm is helpful to understand the mechanism of the Lasso. In the penalized likelihood method, the tuning parameter λ controls the number of nonzero coefficient estimates, with larger λ yielding sparser nonzero estimates. As the tuning parameter λ decreased from ∞ to 0, a series of solutions is called the solution path. The algorithm of LARS is much simpler and uses less computational time to track the entire solution path, although the LARS method yield nearly the same solution path as the Lasso.

Although the Lasso/LARS algorithm has many advantages, it also has some limitations. First, the L_1 penalty shifts the ordinary least squares estimates, which leads to unnecessary bias even when the true parameters are large. Second, the L_1 penalized likelihood estimators cannot work as well as if the correct submodel were known in advance. Another drawback is the number of variables selected by the L_1 penalty is bounded by the sample size n.

There are some LARS extensions described in the literature. Zou and Hastie (2005) proposed the Elastic Net method, whose penalty function is a combination of the L_1 penalty and the L_2 penalty. The number of variables selected by Elastic Net is not bounded by the sample size. Furthermore, the Elastic Net considers the group effect, so highly correlated variables can be selected or removed together. Zou (2006) advocated

the adaptive Lasso, a new version of the Lasso. Unlike Lasso which applies the same penalty to all coefficients, the adaptive Lasso utilizes adaptive weights for penalizing different coefficients in the L_1 penalty. The adaptive Lasso enjoys the oracle properties, whereas the Lasso does not. Park *et al*. (2006) introduced the GLM path algorithm, a path-following algorithm to fit generalized linear models with the L_1 penalty. The GLM path uses the predictor-corrector method of convex-optimization to compute solutions along the entire regularization path.

Fan and Li (2001) pointed out that a good penalty function should result in estimators with three theoretical properties:

- *Unbiasedness*: The estimator is unbiased when the true unknown parameter is large.
- *Sparsity*: The estimator has a threshold structure, which automatically sets small estimated coefficients to zero.
- *Continuity*: The estimator is continuous in data.

These properties can make the model selection avoid unnecessary bias, redundant variables and instability. The L_q penalty function $p_\lambda(|\theta|) = \lambda|\theta|^q$ does not simultaneously satisfy these three properties. Fan and Li (2001) proposed a penalty function possessing all these properties, called the smoothly clipped absolute deviation (SCAD) function. It is based on the L_1 penalty function and the clipped penalty function. Its derivative is

expressed by

$$p'_n(\theta) = \lambda_n I(|\theta| \leq \lambda_n) + \frac{(a\lambda_n - |\theta|)^+}{(a-1)\lambda_n} I(|\theta| > \lambda_n),\ a > 2. \tag{1.4}$$

Assume that the columns of $\mathbf{X}$ is orthonormal, the SCAD penalized likelihood estimators are given by

$$\tilde{\theta} = \begin{cases} sgn(\hat{\theta})(|\hat{\theta}| - \lambda_n)^+, & |\hat{\theta}| \leq 2\lambda_n, \\ \{(a-1)\hat{\theta} - sgn(\hat{\theta})a\lambda_n\}/(a-2), & 2\lambda_n < |\hat{\theta}| < a\lambda_n, \\ \hat{\theta}, & |\hat{\theta}| > a\lambda_n, \end{cases} \tag{1.5}$$

where λ_n and a are two tuning parameters, and $\hat{\theta}$ is the ordinary least squares estimate. From (1.5), it is seen that when the ordinary least square estimate of the unknown parameter is sufficiently large, the SCAD penalty function does not penalize it. Furthermore, the SCAD estimate $\tilde{\theta}$ is a continuous function of the ordinary least squares estimate $\hat{\theta}$. Under some general regularity conditions, the SCAD estimates have oracle property when the smoothing parameter λ_n is appropriately chosen. The oracle property is that the SCAD penalized likelihood estimates perform as well as if the true underlying model is given in advance. Nevertheless, when the separation phenomenon exists in a logistic model, the SCAD method is infeasible. The problem of separation is non-negligible and usually observed in a logistic model with a small sample size and a huge number of possible factors. In case of separation, the log-likelihood function is monotone on at least one unknown parameter. This, combined with the fact that the SCAD penalty function is bounded, results in at least one infinity SCAD penalized estimate.

An appropriate model selection criterion is needed to identify the optimal model from

all candidate models. Many model selection criteria have been developed, including cross-validation (CV) by Stone (1974), generalized cross-validation (GCV) by Craven and Wahba (1979), Akaike information criterion (AIC) by Akaike (1973), Bayesian information criterion (BIC) by Schwarz (1978). However, it was observed that all conventional selection criteria tend to select too many spurious variables by Broman and Speed (2002), Chen and Chen (2007). The extended Bayesian information criterion (EBIC) proposed by Chen and Chen (2007) provides an appropriate model selection criterion for high dimensional feature selection since it can effectively control the number of spurious variables. However, the extended Bayesian information criterion was only discussed in the linear regression model with main effects.

When the dimensionality P is huge, both traditional model selection methods and the penalized likelihood methodology are infeasible mainly because of the small-n-large-P problem. Fortunately, a new series of approaches have been proposed to tackle feature selection with high dimensional feature space. In general, this kind of approaches first reduce a high dimensional feature space to a low dimensional one. Then, model selection method is utilized to find causal features from the reduced feature space. In the following, two high dimensional feature selection methods are reviewed.

Fan and Lv (2008) proposed the sure independent screening (SIS) procedure to reduce the dimensionality of feature space from high to a relatively small scale (d) below the sample size (n) in the context of linear model. SIS procedure applies the componen-

twise regression to select the features with the largest d componentwise magnitudes. After the dimension of the original feature space is reduced, the penalized likelihood methods such as SCAD, LASSO are suggested for estimating unknown parameters or selecting causal features. The procedure of SIS is identical to selecting features by comparing correlations between features and the response variable. This feature makes SIS procedure to be promising because the computation is very simple even if the dimension of feature space is ultra high.

Chen and Chen (2007) developed another procedure called the tournament screening (TS) to reduce the dimension of high dimensional feature space in linear model. In TS procedure, the dimension of feature space is reduced gradually until it reaches a desirable level. At each stage, the features which survived in the previous stage are divided into some non-overlapping groups randomly. Then, a specified number of features are selected by some model selection methods in each group and pooled together as candidates in the next stage. This process is repeated until the dimension of the feature space is reduced to an expected number. After pre-screening, all the features entered the final stage are jointly assessed by the penalized likelihood methodology and grouped into a sequence of nested subsets. For each subset, an un-penalized likelihood model is fitted and then evaluated by some model selection criterion. The tournament screening would be efficient and feasible for feature selection with high dimensional feature space.

1.4 Aim and organization of the thesis

Combining model selection with dimension reduction is an effective strategy to deal with feature selection with high dimensional feature space. Besides linear regression models, other generalized linear regression models built by high dimensional data also play an important role in many areas. For instance, logistic regression model is used to describe the relationship between the phenotype and genotypes in genome-wide association studies. Hence, it is an important and urgent task to investigate high dimensional feature selection in the context of generalized linear models. In this thesis, we provide the generalized tournament screening cum EBIC approach to achieve this purpose and apply it in genome-wide association studies for the identification of genetic variations.

The SCAD method proposed by Fan and Li (2001) is an effective variable selection method with many favorable theoretical properties. Unfortunately, the SCAD method encounters a problem that at least one parameter estimate diverges to infinity in case of the separation phenomenon. Furthermore, the separation phenomenon is non-negligible and primarily occurs in the data with a small sample size and a huge number of possible factors. We introduce the modified SCAD method, which is applicable in case of the separation phenomenon.

The Extended Bayesian information criterion (EBIC; Chen and Chen, 2007) is extremely useful in moderate or high dimensional feature selection, since it can effectively

control the false discovery rate whereas conventional model selection criteria cannot. As the idea of incorporating feature selection into model selection is becoming popular, the EBIC would become more attractive. Its performance was only demonstrated in linear regression models with main effects. In this thesis, we extend EBIC to the generalized linear models with both main effects and interaction effects. Meanwhile, EBIC is a necessary element in the generalized tournament approach.

The thesis is organized as follows:

In Chapter 2, we focus on the problem raised by the separation phenomenon in the original SCAD method. We propose a modified SCAD method by adding the logarithm of the Jeffreys penalty to the SCAD penalized log-likelihood function. The properties and performance of the modified SCAD method are shown by some justifications and simulation studies.

In Chapter 3, we focus on the extended Bayesian information criterion (EBIC) in the context of generalized linear models. EBIC can be used in the model with both main effects and interaction effects. Simulation studies are conducted to demonstrate the performance of EBIC in the medium or high dimensional generalized linear models in comparison with the Bayesian information criterion.

In Chapter 4, we focus on the generalized tournament screening cum EBIC in gen-

eralized linear models. We introduce its whole procedure including the pre-screening step and the final selection step. In addition, some strategies for two steps are proposed.

In Chapter 5, the generalized tournament screening cum EBIC is applied in genome-wide association studies. The penalized logistic model with main effects and interaction effects is introduced. Some numerical studies are conducted to compare the performances of the generalized tournament approach and the multiple testing for gene-gene interactions (Marchini *et al*., 2005).

In Chapter 6, we give the conclusions on the thesis and discuss some future works including choosing an appropriate parameter value for the extended Bayesian information criterion, combining the group selection methods with the generalized tournament approach and constraining the order of selecting main effects and interaction effects.

Chapter 2

The Modified SCAD Method for Logistic Models

The SCAD method (Fan and Li, 2001) is a variable selection method with some favorable theoretical properties. It is suitable to several models including generalized linear regression models. However, the separation phenomenon in logistic regression model pose a challenge to the SCAD method. Separation frequently occurs when the binary outcome variable can be perfectly separated by a single covarite or by a linear combination of the covariates (Albert and Anderson, 1984). In case of separation, the log-likelihood is monotone on at least one unknown parameter. This, combined with the fact that the SCAD penalty is bounded by a constant, causes at least one infinite parameter estimate. It has been shown that the separation phenomenon is non-negligible and primarily occurs in datasets with a small sample size relative to the number of pos-

sible risk factors. To solve the problem raised by separation, we propose the modified SCAD method in this chapter. The modified SCAD method adds the algorithm of the Jeffreys invariant prior (Jeffreys, 1946) to the original SCAD penalized log-likelihood function. This modification ensures finite parameter estimate even in case of separation. We apply the Newton-Raphson algorithm to maximize the modified SCAD penalized likelihood function. In case of no separation, simulation studies are conducted to compare the modified SCAD method with the original SCAD method. It is shown that when the sample size is large enough, the performance of modified SCAD method is the same as that of the original SCAD method with regards to variable selection. Therefore, the modified SCAD method not only provides a solution to the problem of separation but also maintains the performance of the SCAD method.

In the following sections, the modified SCAD method is described in more details. In Section 2.1, we describe the separation phenomenon and review the solution to the problem of separation in the maximum likelihood method. The modified SCAD method is explored and discussed in Section 2.2. In Section 2.3, the performance of the modified SCAD method is illustrated with simulated datasets.

2.1 Introduction to the separation phenomenon

Logistic regression model is used extensively in many areas such as genome-wide association studies and medical studies. Examples of a binary response variable (0/1)

include disease or free of disease, the success of some medicine in treating patients (yes/no). Let Y denote a binary response variable:

$$Y = \begin{cases} 1 \quad , & \text{if the subject falls into a certain category,} \\ 0 \quad , & \text{otherwise.} \end{cases}$$

The logistic regression model has the form

$$\log \frac{p(Y=1|X)}{p(Y=0|X)} = X^T\boldsymbol{\beta}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)$, β_0 denotes the intercept item and $X = (1, X_1, \ldots, X_P)$. The likelihood function of $\boldsymbol{\beta}$ with n observations $\{(y_i, \mathbf{x}_i),\ i = 1, .., n\}$ is given by

$$L(\boldsymbol{\beta}) = \Pi_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{(1-y_i)}, \tag{2.2}$$

where

$$\pi_i = \frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}} \qquad \text{or} \qquad \log\frac{\pi_i}{1-\pi_i} = \mathbf{x}_i^T\boldsymbol{\beta}.$$

Therefore, the log-likelihood function is expressed by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i\ (\mathbf{x}_i^T\boldsymbol{\beta}) - \log[1+\exp\ (\mathbf{x}_i^T\boldsymbol{\beta})]\}. \tag{2.3}$$

Either in medical or in genome-wide association studies, datasets are commonly small or sparse, which tends to cause the separation phenomenon. Separation frequently occurs when the binary outcome variable can be perfectly separated by a single covariate or by a linear combination of the covariates (Albert and Anderson, 1984). For example, 'Age' is one covariate in the logistic model. Consider a situation where every value of the response variable is 0 if the age is less than 40 and every value is 1 if the is age is grater than or equal to 40. The value of response can be perfectly separated by the

covariate 'Age'. It has been shown that the separation phenomenon is a non-negligible problem and primarily occurs in the datasets with a small sample size and some highly predictive risk factors (Heinze and Schemper, 2002). The simplest case of separation is in the analysis of a 2×2 table with one zero cell count. The separation phenomenon renders some methods relevant to estimation of unknown parameters unable to work normally. In the remainder of this section, we describe the problem caused by separation in the maximum likelihood method and review a solution to this problem.

In logistic regression, the maximum likelihood estimate (MLE) of unknown parameters is obtained by an iteratively weighted least-squares algorithm. In the fitting process, it is likely that although the likelihood function converges to a finite value, at least one parameter estimate diverges to infinity. As a result, the corresponding estimated odds ratio is zero or infinite. It has been recognized that this problem is caused by the separation phenomenon. In practice, infinite parameter or zero (infinite) odds ratio is usually considered unrealistic. Therefore, it once seemed that the separation phenomenon posed a challenge to the maximum likelihood method. However, it was found that in exponential family, the penalized likelihood function with a penalty function $|I(\theta)|^{\frac{1}{2}}$ provides a solution to this problem. This penalty is the Jeffreys invariant prior (Jeffreys, 1946).

The asymptotic bias of the maximum likelihood estimate $\hat{\theta}$ can be expressed by $b(\theta) = b_1(\theta)/n + b_2(\theta)/n^2 + \ldots$, where n is the sample size. In a logistic regression model, the

$O(n^{-1})$ bias can be written by

$$b_1(\theta)/n = (X^T WX)^{-1} X^T W\xi, \tag{2.4}$$

where $W = \text{diag}\{\pi_i(1-\pi_i)\}$, $W\xi$ has i-th element $h_i(\pi_i - 1/2)$ and h_i is the diagonal element of the matrix $H = W^{1/2}X(X^T WX)^{-1}X^T W^{1/2}$. Firth (1993) proposed a modified score procedure to remove $O(n^{-1})$ bias for MLE. In exponential family, its effect is to penalize the likelihood function by the Jeffreys invariant prior. Firth illustrated with one example that this modification produces finite estimate instead of infinite MLE in case of separation. Heinze and Schemper (2002) pointed out that Firth's modified score procedure can solve the problem of separation in the maximum likelihood method. Furthermore, Heinze and Ploner (2003) developed a statistical software package in R, a comprehensive tool to facilitate the application of Firth's modified score procedure in logistic regression.

Let $\{(y_i,\ \mathbf{x}_i),\ i = 1, ..., n\}$ denote a sample of n observations with the response variable Y and the covariate vector X of dimension P. In general, the maximum likelihood estimate of the unknown parameter $\boldsymbol{\beta}$ is the solution of the score equation $U(\boldsymbol{\beta}) = \partial \log L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = 0$, where $L(\boldsymbol{\beta})$ is the likelihood function. However, the maximum likelihood estimate may be seriously biased when the sample size is small. In order to reduce the bias, Firth suggested to use Firth's modified score equations instead of the original ones $U(\beta_r) = 0$. In exponential family, the modified score equations is given by

$$U(\beta_r)^* = U(\beta_r) + \frac{1}{2}\text{trace}[I(\boldsymbol{\beta})^{-1}\{\frac{\partial I(\boldsymbol{\beta})}{\partial \beta_r}\}] = 0,\ r = 1, ..., P, \tag{2.5}$$

where $I(\boldsymbol{\beta})$ is the Fisher information matrix, i.e. the negative of the expected second derivative of the log-likelihood function. It was shown that the modified score equation (2.5) can remove the $O(n^{-1})$ bias of the maximum likelihood estimate. Moreover, in exponential family with canonical parameterization, Firth's modified score procedure is corresponding to the penalized log-likelihood function $\log L(\boldsymbol{\beta})^* = \log L(\boldsymbol{\beta}) + \log |I(\boldsymbol{\beta})|^{1/2}$, where the penalty $|I(\boldsymbol{\beta})|^{1/2}$ is named as Jeffreys invariant prior (Jeffreys, 1946).

Since the original purpose of Firth's modified score procedure is to reduce the bias of the maximum likelihood estimate, its function relevant to the separation problem was not fully recognized. Thus, Heinze and Schemper (2002) reviewed Firth's modified score procedure and suggested to use it to produce finite estimate in case of separation. Firth's modified score function for logistic regression model is

$$U(\beta_r)^* = \sum_{i=1}^{n} \{y_i - \pi_i + h_i(\frac{1}{2} - \pi_i)\} x_{ir}, \tag{2.6}$$

where the h_i is the i-th diagonal element of the hat matrix $H = W^{1/2}\mathbf{X}(\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T W^{1/2}$ with $W = \text{diag}\{\pi_i(1 - \pi_i)\}$. Then, the Firth-type estimate can be obtained by a Newton-Raphson algorithm

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + I^{-1}(\boldsymbol{\beta}^{(s)}) U(\boldsymbol{\beta}^{(s)})^*, \tag{2.7}$$

where $\boldsymbol{\beta}^{(j)}$ denotes the estimate in the j-th iteration and $U(\cdot)^*$ is Firth's score function (2.6).

Firth's modified score function (2.6) can be rewritten by

$$U(\beta_r)^* = \sum_{i=1}^{n}\{(y_i - \pi_i)(1 + h_i/2) + (1 - y_i - \pi_i)h_i/2\}x_{ir}. \tag{2.8}$$

Assume that each observation $(y_i,\ \mathbf{x}_i)$ is splitting into two new observations $(y_i,\ \mathbf{x}_i)$ and $(1 - y_i,\ \mathbf{x}_i)$, respectively with iteratively updated weights $1 + h_i/2$ and $h_i/2$. In this way, any $\mathbf{x}_i$ in the new data set is corresponding to one response and one non-response. It ensures that the separation phenomenon never exists in the new data set. Consequently, the maximum likelihood estimate based on the new observations is always finite. In addition, it is seen that the ordinary score function $U(\beta_r) = \partial \log L(\boldsymbol{\beta})/\partial\beta_r$ for the new observation $\{(y_i,\ \mathbf{x}_i),\ (1 - y_i,\ \mathbf{x}_i),\ i = 1, 2, \ldots, n\}$ has the same expression as (2.8). It shows that the solutions to Firth's modified score equation are finite. Therefore, Firth's modified score function or Jeffreys invariant prior provides a solution to the problem of separation in the maximum likelihood method.

Other than the maximum likelihood method, the SCAD method is also affected by the separation phenomenon. In the next section, we review the SCAD method and describe its problem caused by the separation phenomenon. Finally, we propose the modified SCAD method to tackle the problem caused by separation.

2.2 The modified SCAD method in logistic regression model

The SCAD method is an effective variable selection approach via penalized likelihood (Fan and Li, 2001). Compared with the classical model selection methods such as subset selection, the SCAD method is more stable and still feasible for high dimensional data. Moreover, the family of smoothly clipped absolute deviation (SCAD) penalty functions results in its estimate with three properties: unbiasedness, sparsity and continuity. In contrast, the estimate by L_q penalty does not have these three properties simultaneously. One more important thing is that the SCAD method enjoys the oracle property with a proper choice of regularization parameters. It means that the SCAD method performs as well as the true model is known in advance. It has been shown with simulation studies that the SCAD method obtains the best performance in identifying significant covariates in comparison with some other penalty likelihood approaches.

In logistic regression, the penalized log-likelihood with the SCAD penalty function is given by

$$\begin{aligned} l_S(\boldsymbol{\beta} \mid \lambda) &= \log L(\boldsymbol{\beta}) - n \sum_{j=1}^{P} p_\lambda(|\beta_j|) \\ &= \sum_{i=1}^{n} \{y_i(\mathbf{x}_i^T \boldsymbol{\beta}) - \log[1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})]\} - n \sum_{j=1}^{P} p_\lambda(|\beta_j|), \end{aligned} \tag{2.9}$$

where

$$p_\lambda(\theta) = \begin{cases} \lambda|\theta| & : \quad |\theta| \leq \lambda, \\ -(\theta^2 - 2a\lambda|\theta| + \lambda^2)/[2(a-1)] & : \quad \lambda < |\theta| \leq a\lambda, \qquad \text{for some } a > 2 \\ (a+1)\lambda^2/2 & : \quad |\theta| > a\lambda. \end{cases} \tag{2.10}$$

is the family of SCAD penalty functions. It can be seen that the SCAD penalty function is bounded by a constant $(a+1)\lambda^2/2$ if the regularization parameters λ and a are given. The first order derivative of the SCAD function (2.10) is expressed by

$$p_\lambda^{'}(\theta) = \lambda\{I(|\theta| \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(|\theta| > \lambda)\}. \tag{2.11}$$

When the estimate is larger than $a\lambda$, the first order derivative of the SCAD penalty is equal to zero.

Given the values of regularization parameters λ and a, the SCAD method selects variables and estimates unknown parameters via maximizing the penalized log-likelihood function (2.9). The penalized log-likelihood function consists of the log-likelihood function and the SCAD penalty function. When the separation phenomenon exists in the dataset, responses and non-responses are separated by one variable or a linear combination of some variables. Therefore, the log-likelihood function is monotone on at least one parameter. This, combined with the fact that the SCAD penalty is bounded, results in at least one infinite estimate. Therefore, the SCAD method is unable to estimate unknown parameters and select variables when the separation phenomenon exists.

To produce finite parameter estimates, we propose the modified SCAD method. The modified SCAD method adds the algorithm of the Jeffreys invariant prior (Jeffreys, 1946) to the original SCAD penalized log-likelihood function. The penalized log-likelihood function of the modified SCAD method is expressed by

$$l_{MS}(\boldsymbol{\beta} \mid \lambda) = \log L(\boldsymbol{\beta}) + \frac{1}{2}\log|I(\boldsymbol{\beta})| - n\sum_{j=1}^{P} p_\lambda(|\beta_j|). \tag{2.12}$$

Three items in the right side of (2.12) are the log-likelihood function, the Jeffreys penalty and the SCAD penalty respectively. The score function based on the penalized likelihood function (2.12) with n observations $\{(y_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$ is

$$\begin{aligned} U_{MS}(\beta_r) &= \sum_{i=1}^{n}\{y_i - \pi_i + h_i(\frac{1}{2} - \pi_i)\}x_{ir} - np_\lambda^{'}(|\beta_r|) \\ &= \sum_{i=1}^{n}\{(y_i - \pi_i)(1 + h_i/2) + (1 - y_i - \pi_i)h_i/2\}x_{ir} - np_\lambda^{'}(|\beta_r|), \end{aligned} \tag{2.13}$$

where h_i is the i-th diagonal element of the hat matrix H. The score function of the original SCAD method is given by

$$U_S(\beta_r) = \sum_{i=1}^{n}(y_i - \pi_i)x_{ir} - np_\lambda^{'}(|\beta_r|). \tag{2.14}$$

Assume that $\{(y_i, \mathbf{x}_i),\ ((1-y_i), \mathbf{x}_i)\ i = 1, \ldots, n\}$ is a new dataset and $(y_i, \mathbf{x}_i)$ and $((1-y_i), \mathbf{x}_i)$ are weighted by $1 + h_i/2$ and $h_i/2$. Then, the score function is expressed by

$$U_S(\beta_r)_{new} = \sum_{i=1}^{n}\{(y_i - \pi_i)(1 + h_i/2) + (1 - y_i - \pi_i)h_i/2\}x_{ir} - np_\lambda^{'}(|\beta_r|). \tag{2.15}$$

Compared (2.13) with (2.15), it is seen that the score function $U_S(\beta_r)$ with $\{(y_i, \mathbf{x}_i),\ ((1-y_i), \mathbf{x}_i)\ i = 1, \ldots, n\}$ has the same expression as the modified score function $U_{MS}(\beta_r)$ with $\{(y_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$. The separation phenomenon never occurs in the new data

set since any $\mathbf{x}_i$ has one response and one non-response. Therefore, the modified SCAD method always avoids infinite estimate caused by separation. In the following, the Newton-Raphson algorithm in the modified SCAD method is described in details.

Assumed that $\boldsymbol{\beta}^{(s)}$ is the estimate at the s-th iteration with an initial value of $\boldsymbol{\beta}^{(0)}$. The estimate at the $(s+1)$-th iteration $\boldsymbol{\beta}^{(s+1)}$ is obtained by

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - (\nabla^2 l_{MS}(\boldsymbol{\beta}^{(s)}))^{-1} \nabla l_{MS}(\boldsymbol{\beta}^{(s)}). \tag{2.16}$$

This is the Newton-Raphson algorithm. Here, the log-likelihood function and the SCAD penalty in (2.12) are locally approximated by

$$l(\beta^{(s)}) + \nabla l(\beta^{(s)})^T(\beta^{(s)}) + \frac{1}{2}(\beta - \beta^{(s)})^T \nabla^2 l(\beta^{(s)})(\beta - \beta^{(s)}) - \frac{1}{2} n\beta^T \Sigma_\lambda(\beta^{(s)})\beta, \tag{2.17}$$

where $l(\boldsymbol{\beta}^{(s)})$ denotes the log-likelihood function of $\boldsymbol{\beta}^{(s)}$, $\nabla(\boldsymbol{\beta}^{(s)})$ denotes the first partial derivative of the likelihood function $\partial l(\boldsymbol{\beta}^{(s)})/\partial(\boldsymbol{\beta})$, $\nabla^2 l(\boldsymbol{\beta}^{(s)})$ denotes the second partial derivative $\partial^2 l(\boldsymbol{\beta}^{(s)})/\partial(\boldsymbol{\beta})\partial(\boldsymbol{\beta})^T$, and $\Sigma_\lambda(\boldsymbol{\beta}^{(s)})$ is the diagonal matrix with diagonal elements $(p'_\lambda(|\beta_1^{(s)}|)/|\beta_1|, \ldots, p'_\lambda(|\beta_P^{(s)}|)/|\beta_P|)$. Since it is difficult to get the second order derivative of the information matrix, we propose to exclude it in $(\nabla^2 l_{MS}(\boldsymbol{\beta}^{(s)}))^{-1}$. When Heinze and Schemper (2002) considered the Firth's penalized likelihood, they also used the information matrix to approximate the second order derivative. Thus, (2.16) can be approximated by

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + \{I(\boldsymbol{\beta}^{(s)}) + n\Sigma_\lambda(\boldsymbol{\beta}^{(s)})\}^{-1} U_{MS}(\boldsymbol{\beta}^{(s)}), \tag{2.18}$$

where $I(\boldsymbol{\beta}^{(s)}) = \mathbf{X}^T W^{(s)} \mathbf{X}$ with $W^{(s)} = diag\{\pi_i^{(s)}(1 - \pi_i^{(s)})\}$ is the Fisher information matrix at $\boldsymbol{\beta}^{(s)}$, $U_{MS}(\beta_r^{(s)}) = \sum_{i=1}^n \{y_i - \pi_i^{(s)} + h_i(\frac{1}{2} - \pi_i^{(s)})\}x_{ir} - np'_\lambda(|\beta_r^{(s)}|)$ denotes the modified

score function at $\boldsymbol{\beta}^{(s)}$, and the h_i is the i-th diagonal element of the hat matrix at the s-th iteration, $H^{(s)} = (W^{(s)})^{1/2}\mathbf{X}(\mathbf{X}^T W^{(s)}\mathbf{X})^{-1}\mathbf{X}^T(W^{(s)})^{1/2}$. Note that $U_{MS}(\beta_r^{(s)})$, $H^{(s)}$ and $W^{(s)}$ are needed to update in each iteration. Moreover, the maximum absolute difference between $\boldsymbol{\beta}^{(s)}$ and $\boldsymbol{\beta}^{(s+1)}$ is controlled by a given value. It can avoid the numerical problems during estimation (Heinze and Ploner, 2003). The modified SCAD estimate is obtained until the algorithm converges, i.e., $\sum_{r=1}^{P} |\beta_r^{(s+1)} - \beta_r^{(s)}|$ or $\max_{r=1,\ldots,P} |\beta_r^{(s+1)} - \beta_r^{(s)}|$ is less than a convergence criterion ϵ.

In Firth's modified score procedure, it was shown that the Jeffreys invariant prior removes $O(n^{-1})$ bias from the maximum likelihood estimate. From this conclusion, it can be seen that the influence of the Jeffreys invariant prior is asymptotically negligible, i.e., the modified SCAD method only adds one negligible item to the SCAD penalized likelihood function. Thus, it appears that the modified SCAD method should maintain the performance of the original SCAD method when the sample size is large. Moreover, as discussed earlier, the modified SCAD method should produce finite estimate in case of the separation phenomenon. In next section, we conduct some simulation studies to evaluate the performance of the modified SCAD method.

2.3 Simulation studies

In the first simulation, we compare the performance of the SCAD method with the modified SCAD method when no separation phenomenon exists in the data sets. The

purpose is to examine whether the Jeffreys invariant prior affects the performance of the modified SCAD method. There are two measures: the positive selection rate (PSR) and the false discovery rate (FDR). The positive selection rate is defined as the proportion of the truly associated covariates selected. The false discovery rate is defined as the proportion of falsely selected covariates among all selected ones. Moreover, the regularization parameter λ in the SCAD penalty function is chosen by Bayesian information criterion (BIC) in both methods. The value of another parameter a in the SCAD penalty function is set to 3.7, since this value was shown to be very reasonable (Fan and Li, 2001).

Example 2.1 : In this example, we simulated 200 data sets consisting of 200 observations from the model $Y \sim Bernoulli\{p(\beta_1 X_1 + \ldots + \beta_8 X_8)\}$, where $p(\mu) = \exp(\mu)/(1 + \exp(\mu))$. The true coefficient vector is set to be

$$(\beta_1, \ldots, \beta_8) = (3,\ 1.5,\ 0,\ 0,\ 2,\ 0,\ 0,\ 0).$$

The first six components of X follow standard normal distribution. The correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$. The last two components of X are independently identically distributed as a Bernoulli distribution with probability of success 0.5. This is a model used in Fan and Li (2001). In Table 2.1, the column labeled "Correct" denotes the average restricted to the true nonzero coefficients, and the column labeled "Incorrect" presents the average of coefficients erroneously set to nonzero. The PSR and FDR denote the average positive selection rate and the average false discovery rate over 200 replications. The standard deviations based on the 200 random samples are

presented in the parentheses, which provides information on simulation errors.

Table 2.1: Simulation results for logistic regression model in case of no separation

Method	Avg. No of nonzero coefficients		PSR	FDR
	Correct	Incorrect		
MSCAD	2.99(0.10)	0.10(0.30)	0.997	0.032
SCAD	2.99(0.10)	0.10(0.30)	0.997	0.032

From Table 2.1, it can be seen that the performance of the modified SCAD method is the same as that of the SCAD method in these two settings. Same performances of these two methods could be attributed to the asymptotic negligible effect of Jeffreys invariant prior. Therefore, simulation results show that the modified SCAD method with Jeffreys invariant prior does not affect the performance of the SCAD penalty function.

In the second simulation, the number of covariates is increased to 100 from 8. In this situation, the sample size is small relative to the number of covariates, so the separation phenomenon likely occurs. The first purpose of this simulation study is to show that the modified SCAD method is still feasible in case of separation. The second one is to evaluate the performance of the modified SCAD method in case of separation in comparison with the L_1 penalty function.

Example 2.2 In this example, we also simulated 200 datasets consisting of 200 observations from the model $Y \sim Bernoulli\{p(\beta_1 X_1 + \ldots + \beta_{100} X_{100})\}$, where $p(\mu) = \exp(\mu)/(1 + \exp(\mu))$. The number of covariates is increased to 100. The true coeffi-

cient vector is set to be

$$(\beta_1, \ldots, \beta_{100}) = (3,\ 1.5,\ 0,\ 0,\ 2,\ \underbrace{0, \ldots, 0}_{95}).$$

Its first 8 components of the true coefficients $\boldsymbol{\beta}$ is the same as the first setting of Example 2.1 and the other components are set to 0. The first 6 components of X are standard normal. The correlation between X_i and X_j is $\rho^{|\,i-j|}$ with $\rho = 0.5$. The last 94 components of X are independently identically distributed as a Bernoulli distribution with probability of success 0.5.

Table 2.2: Simulation results for logistic regression model in case of separation

Method	Avg. No of nonzero coefficients		PSR	FDR
	Correct	Incorrect		
MSCAD	2.99(0.12)	1.59(1.39)	0.997	0.347
L_1 − penalty	2.99(0.10)	2.05(2.16)	0.997	0.407

In Example 2.2, although the separation phenomenon occurs in some replications, the modified SCAD method still works normally. As shown in Table 2.2, the modified SCAD method has the same positive selection rate as the L_1 penalty method. However, the modified SCAD method produces a lower false discovery rate in comparison with the L_1 penalty method, i.e., the L_1 penalty method selects more spurious covariates. In terms of the positive selection rate and false discovery rate, the simulation results show that the modified SCAD method performs better than the L_1 penalty method.

Although the false discovery rate of the modified SCAD method is lower than that

of the L_1 penalty method, it is still unsatisfactorily high (In Example 2.2, it reaches 0.347). Nevertheless, when the number of candidate covariates is 8 (Example 2.1), its false discovery rate is only 0.032. It suggests that the covariates selected by Bayesian information criterion may have an increasing trend as the number of candidate covariates increases. Moreover, it is likely that Bayesian information criterion is inappropriate in medium or high dimensional model space.

2.4 Summary

We have proposed the modified SCAD method to solve the problem of infinite SCAD estimators in case of separation. When the separation phenomenon exists, at least one SCAD estimate is infinite because the log-likelihood function is monotone in this situation and the SCAD penalty function is bounded. Adding the Jeffreys invariant prior guarantees that the modified SCAD method always yield finite estimate. Moreover, the modified SCAD method performs as well as the original SCAD method when the sample size is large because the Jeffreys invariant prior is asymptotic negligible.

Separation is a non-negligible phenomenon, especially in the data sets with a small sample size relative to the number of candidate factors and some highly predictive factors. These situations are common in genome-wide association studies or medical studies. Thus, this chapter develops a necessary modification for the original SCAD method.

Chapter 3

Model Selection Criteria in Generalized Linear Models

In model selection, optimization of a model selection criterion is one approach to identify the best model from all candidates. Undoubtedly, it is a critical problem to explore an appropriate criterion. The extended Bayesian information criteria (EBIC) were proposed by Chen and Chen (2007) as model selection criteria for high dimensional feature space. However, the EBIC method was only studied in linear model with main effects. In this chapter, we discuss the extended Bayesian information criteria in generalized linear models with some justifications and simulations. If the generalized linear model of interest contains all possible two-covariate interactions as well as main effect terms, we modify the EBIC method by penalizing main effects and interactions with two different penalty functions. The main reason is that the effect of selecting one interaction

is to involve two variables in the model. In the context of logistic regression model, we compare the performance of EBIC with different parameter values in terms of the positive selection rate and the false discovery rate. Through some simulated datasets, we demonstrate that the BIC method suffers high false discovery rate while the EBIC method with the most stringent parameter value effectively controls it. Therefore, the EBIC method could be more appropriate than the BIC method in generalized linear models with high dimensional model space. It is consistent with the results in linear model by Chen and Chen (2007). Simulation studies in this chapter only discuss the cases when the number of covariates is moderate and less than the sample size. The performance of the extended Bayesian information criteria in high dimensional generalized linear models will be further evaluated in Chapter 5.

We review several model selection criteria in the context of linear regression model in Section 3.1. In Section 3.2, we describe the extended Bayesian information criteria in generalized linear regression models. In Section 3.3, The EBIC method is illustrated with some simulated datasets. We conclude with a summary in Section 3.4.

3.1 Introduction to model selection criteria

Suppose that Y is the response variable and $X = (1, X_1, \ldots, X_P)^T$ is the vector of covariates. The linear model is expressed by

$$Y = X^T\boldsymbol{\beta} + \varepsilon = \beta_0\mathbf{1} + \beta_1 X_1 + \ldots + \beta_P X_P + \varepsilon, \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_P)^T$ are unknown parameters, and the random error vector ε is independently and identically distributed with mean 0 and variance σ^2. Model M denotes any candidate model. Let $X(M)$ and $\boldsymbol{\beta}(M)$ respectively denote the vector of covariates and corresponding unknown parameters in model M. If model M includes all covariates, model M is the full model.

Various model selection criteria have been proposed, e.g., the C_p criterion (Mallows, 1973), the cross-validation (CV) method (Stone, 1974; Allen, 1974) and the generalized cross-validation (GCV) method (Craven and Wahba, 1979), the Akaike's information criterion (AIC) method (Akaike, 1970) and AIC_c method (Sugiura, 1978; Hurvich and Tsai, 1989), the Bayesian information criterion (BIC) method (Schwarz, 1978; Hannan and Quinn, 1979). Again, some literatures also discussed the asymptotic performance for evaluating these model selection criteria (Nishii, 1984; Rao and Wu, 1989; Stone, 1979; Shibata, 1981; Li, 1987; Shao, 1993). However, the power of any model selection criterion depends on the underlying circumstance and applied fields. Shao (1997) classified these model selection criteria on the basis of asymptotic performance. When the dimension of model space is high, these ordinary model selection criteria are too liberal since they tend to select too many spurious covariates. Fortunately, on the basis of the BIC method, the extended family of Bayesian information criteria (EBIC; Chen and Chen, 2007) were proposed to solve this problem raised by high dimensional model space. Without loss of generality, various model selection criteria are reviewed in the context of linear model without the intercept item β_0 in the remainder of this section.

Mallows (1973) proposed the C_p criterion which is based on some form of mean squared error (MSE) or mean squared prediction error (MSPE). It is well known that MSE is a common way to measure the performance of a prediction. The MSE of model M is defined by

$$MSE(M) = \frac{\sum_{i=1}^{n} E\{\mathbf{x}_i(M)^T\boldsymbol{\beta}(M) - \mathbf{x}_i(M)^T\hat{\boldsymbol{\beta}}(M)\}^2}{n}, \tag{3.2}$$

where $\hat{\boldsymbol{\beta}}(M) = (\mathbf{X}(M)^T\mathbf{X}(M))^{-1}\mathbf{X}(M)^T Y$ is the least square estimate of the unknown parameter vector $\boldsymbol{\beta}(M)$ in model M and $\mathbf{X}(M)$ is the design matrix of model M. However, it can be seen from (3.2) that $MSE(M)$ contains the unknown parameter vector $\boldsymbol{\beta}(M)$, so the mean square error cannot be used directly as a model selection criterion. A natural idea is to replace MSE with its unbiased estimator. Assume that the variance of the random error σ^2 is known. In this situation, the unbiased estimator of $MSE(M)$ is given by

$$\frac{SSE(M)}{n} + \frac{2\sigma^2 v(M)}{n} - \sigma^2, \tag{3.3}$$

where $SSE(M) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i(M)^T\hat{\boldsymbol{\beta}}(M))^2$ is the sum of squares error of model M, and $v(M)$ denotes the number of covariates contained in model M. In terms of prediction performance, the best model is the one that minimizes the unbiased estimator of mean squared error (3.3). Since the last item σ^2 is independent of model M, the best model can be obtained by minimizing the first two items in (3.3). This is the derivation of the C_p criterion, which is expressed by

$$C_p(M) = \frac{SSE(M)}{n} + \frac{2\sigma^2 v(M)}{n}. \tag{3.4}$$

If the variance of the random error σ^2 is an unknown parameter, it is replaced by the least square estimate $\hat{\sigma}^2$ based on the full model. The best model is the one that minimizes the C_p criterion (3.4).

Another popular criterion is cross-validation (CV) proposed by Stone (1974). In CV, the dataset is partitioned into some subsets: training set and testing sets. The idea of CV is to fit a model on the training set. When the training is done, the testing sets are used to validate the performance of the model. In K-fold cross-validation, the original data is divided into K subsets. Each time, one of K subsets is retained as the validation data for testing the model, and the remaining $K - 1$ subsets put together to fit a model. The cross-validation process is then repeated K times (the folds), with each of the K subsets used exactly once as the validation data. The K results from the folds are averaged as a single estimation.

Leave-one-out cross validation (LOOCV) can be seen as the extreme of K-fold cross-validation, with K being equal to the number of observations in the original sample. As the name suggests, leave-one-out cross-validation (LOOCV) involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. For model M, let $\hat{\boldsymbol{\beta}}_{-j}(\mathbf{x}_i, y_i)$ be the ordinary least square estimate of unknown parameter vector $\boldsymbol{\beta}$ with the training data

$\{(\mathbf{x}_i, y_i), i = 1, \ldots, j-1, j+1, \ldots, n\}$. The LOOCV criterion is expressed by

$$LOOCV(M) = \frac{\sum_{j=1}^{n} \sum_{i \neq j} \{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{-j}(\mathbf{x}_i, y_i)\}^2}{n}. \tag{3.5}$$

(3.5) can be written as

$$LOOCV(M) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\mathbf{x}_i, y_i)\}^2}{\{1 - h_i(M)\}^2}, \tag{3.6}$$

where $h_i(M)$'s are the diagonal elements of $H(M) = \mathbf{X}(M)\ (\mathbf{X}(M)^T \mathbf{X}(M))^{-1} \mathbf{X}(M)$, the so-called hat matrix. In this way, one only needs to fit the model once with the full data and compute the diagonal elements of the hat matrix $H(M)$. If the $h_i^{'}s$ are replaced by the average of all diagonal elements of the hat matrix, the criterion is the generalized cross-validation criterion (GCV; Craven and Wahba, 1979)

$$GCV(M) = \frac{1}{n} \sum_{i=1}^{n} \frac{\{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\mathbf{x}_i, y_i)\}^2}{\{1 - \text{trace}(H(M)/n)\}^2}, \tag{3.7}$$

where $tr(H(M)/n)$ denotes the trace of the matrix $H(M)/n$.

Two other popular and well-studied model selection criteria are the Akaike's information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978), both of which are likelihood-based methods. The AIC method is derived on the basis of Kullback-Leibler $(K - L)$ distance between the true model and the approximating candidate model. The BIC method is derived in a Bayesian context with the same prior probability on each candidate model.

The Akaike's information criterion (AIC) is defined as minus twice the maximized log-likelihood for model M, penalized with twice the number of covariates contained in M.

It can be expressed by

$$AIC(M) = -2\log(L(\hat{\boldsymbol{\beta}}(M)) + 2v(M) \tag{3.8}$$

In (3.8), $L(\hat{\boldsymbol{\beta}}(M))$ is the maximized likelihood function on model M. Specifically, if the random error vector ε in (3.1) independently and identically follows the normal distribution with mean 0 and variance σ^2, AIC of model M is equivalent to

$$\frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i(M)^T\hat{\boldsymbol{\beta}}(M))^2}{n\sigma^2} + \frac{2v(M)}{n},$$

which is equivalent to the $C_p(M)$ criterion (3.4). Since AIC is an estimate of $K-L$ distance between the true model and the candidate model, the best model is that minimizes the quantity of AIC. If the sample size is small relative to the number of unknown parameters, it was found that the AIC method performs poorly (Sugiura,1978; Sakamoto *et al*., 1986). On the basis of the AIC method, a refined criterion AIC_c was proposed by Hurvich and Tsai (1989). It can be expressed by

$$AIC_c(M) = AIC + \frac{2v(M)(v(M)-1)}{n - v(M) - 1}.$$

Burnham *et al*.(1994) suggested that the AIC_c is more appropriate when the ratio n/P is small. When the sample size is sufficiently large, the performance of AIC_c is similar to that of the AIC method.

The Bayesian information criterion (BIC) of Schwarz (1978) penalizes instead with the logarithm of the sample size. It is given by

$$BIC(M) = -2\log(L(\hat{\boldsymbol{\beta}}(M)) + v(M)\log(n). \tag{3.9}$$

Let $\pi(\boldsymbol{\beta}(M))$ denote the prior density function of $\boldsymbol{\beta}(M)$ and $p(M)$ denote the prior probability of model M. The likelihood function of model M can be expressed by

$$m(Y|M) = \int f(Y;\boldsymbol{\beta}(M))\pi(\boldsymbol{\beta}(M))d\boldsymbol{\beta}(M).$$

The posterior probability of model M is

$$p(M|Y) = \frac{m(Y|M)p(M)}{\sum_M m(Y|M)p(M)}. \tag{3.10}$$

Under the Bayesian paradigm, the best model is what maximizes the posterior probability (3.10). Since the denominator in (3.10) is a constant, maximization of the posterior probability is equivalent to maximize the numerator of the posterior probability. Under some regularity conditions on $f(Y;\boldsymbol{\beta})$, $-2\log(m(Y|M))$ has a Laplace approximation given by BIC(M) up to an additive constant. Moreover, it is known that the prior probabilities assigned to all candidate models are the same and equal to the reciprocal of the number of all candidate models. Thus, maximizing the posterior probability is equivalent to minimizing the Bayesian information criterion (3.9).

The penalty functions of both AIC and BIC are non-decreasing to the number of covariates involved in the selected model, so they discourage the selection of models with excessive number of covariates. However, in comparison with the BIC method, the AIC method tends to select the model with more covariates, since the penalty function of BIC is larger than AIC. In addition, the BIC method is dimension consistent under the assumptions that the true model exists and that the true model is in the set of candidate models.

To evaluate various model selection criteria, some studies discussed their asymptotic performance while their assumptions are different and impact on the results (Nishii, 1984; Rao and Wu, 1989; Stone, 1979; Shibata, 1981; Li, 1987; Shao, 1993). It was found that the results are different, even contrary to each other. The main factors determining the asymptotic performance of various model selection criteria are whether the true models are among the candidate models and whether the dimension of unknown parameters increases with the sample size n.

Shao (1997) provided a clear picture of the asymptotic performance of various model selection criteria in terms of consistency and proposed the generalized information criterion (GIC) to summarize these model selection criteria. Under the assumption that the random error vector independently and identically follows the normal distribution with mean 0 and variance σ^2, the generalized information criterion is given by:

$$GIC_{\lambda_n}(M) = \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i(M)^T\hat{\boldsymbol{\beta}}(M))^2}{n} + \frac{\lambda_n\hat{\sigma}^2 v(M)}{n}, \tag{3.11}$$

where $\{\lambda_n\}$ is a sequence of non-random number ≥ 2 and satisfies $\lambda_n/n \to 0$, as $n \to \infty$. According to the asymptotic behavior, model selection criteria were classified into three classes:

C1. The GIC_2, the C_p, the AIC, the LOOCV and the GCV.

C2. The GIC_{λ_n} with $\lambda_n \to \infty$ and the delete-d CV with $d/n \to 1$.

C3. The GIC_λ with a fixed $\lambda > 2$ and the delete-d with $d/n \to \eta \in (0, 1)$.

The methods in group 1 are useful in the case where there is no fixed-dimension correct model. With a suitable choice of λ_n or d, the methods in group 2 are useful in the case where there exist fixed-dimension correct models. The methods in group 3 are compromised versions of the methods in group 1 and group 2; but their asymptotic performances are not as good as those of the methods in group 1 in the case where no fixed-dimension correct model exists, and not as good as those of the methods in class 2 when there are fixed-dimension correct models.

With the development of modern technologies, high dimensional data frequently appear in many fields and provide more information; on the other hand, these high dimensional data also pose great challenges to model selection. One of the challenges is that the conventional model selection criteria tend to select models with too many spurious covariates as the best model (Broman and Speed, 2002; Chen and Chen, 2008). Chen and Chen (2008) re-examined the Bayesian paradigm implemented in the ordinary Bayesian information criterion and proposed the extended family of Bayesian information criteria (EBIC) to solve this problem.

Let $\mathcal{S}_i$ denote the set of models containing i covariates. As mentioned earlier, the prior probability of any candidate model is the same in the ordinary Bayesian information criterion. Consequently, the total probability assigned to the set $\mathcal{S}_i$ is proportional to the size of $\mathcal{S}_i$. Assume that the number of covariates in the full model is P. The sizes of $\mathcal{S}_1$ and $\mathcal{S}_2$ are respectively P and $P(P-1)/2$. Thus, the probability assigned to $\mathcal{S}_2$ is

$(P-1)/2$ times as that assigned to the set $\mathcal{S}_1$. Chen and Chen (2007) pointed out that it is unreasonable to assign much higher probability to the set of models containing more covariates. Let $\tau(\mathcal{S}_j)$ denote the size of the set $\mathcal{S}_j$. The extended Bayesian information criteria (EBIC) suggest to replace the proportional probability $\tau(\mathcal{S}_j)$ with the probability $\tau(\mathcal{S}_j)^\xi$ for some ξ less than 1 and more than or equal to 0. The extended Bayesian information criteria are based on a set of new prior probabilities proportional to $\tau(\mathcal{S}_j)^\xi$ and expressed by

$$EBIC_\gamma(M) = -2\log L(\hat{\boldsymbol{\beta}}(M)) + v(M)\log(n) + 2\gamma\log[\tau(\mathcal{S}_{v(M)})],\ 0 \le \gamma \le 1. \quad (3.12)$$

In (3.12), when the parameter γ equals to zero, the extended Bayesian information criterion is equivalent to the original Bayesian information criterion. EBIC with $\gamma = 1$ is the most stringent criterion in its family. In the context of linear regression, the extended Bayesian information criteria are consistent if $P = O(n^k)$ and $\gamma > 1 - (1/2k)$. As a result, an appropriate choice of γ is $1 - \log(n)/(2\log(P))$ since it is the lower bound of consistency.

In linear regression model only with main effects, the EBIC method suffers slightly lower positive selection rate (PSR) but effectively controls false discovery rate (FDR) in comparison with the ordinary Bayesian information criteria. Due to high false discovery rate, the ordinary Bayesian information method may be inappropriate in high dimensional model space. Therefore, the EBIC method could be an useful tool in model selection with high dimensional model space. However, the EBIC method was only discussed and evaluated in the linear regression model with main effects. In practice,

generalized linear regression models are more flexible and suitable to describe the relationship between the response and covariates. Moreover, it was found in Example 2.2 (Chapter 2) that the Bayesian information criterion may tend to select more spurious covariates when the number of candidate covariates increases. In the next section, the EBIC method is extended to generalized linear models with both main effects and two-covariate interactions.

3.2 The extended Bayesian information criteria in generalized linear models

In a generalized linear model(GLM), the response variable Y follows a particular distribution function in exponential family with mean $\mu = E(Y)$ and variance $V = Var(Y)$. The probability density function of Y is given by

$$f(y|\theta,\phi) = \exp\{\frac{y\,\theta - b(\theta)}{a(\phi)} + c(\phi,y)\}I_A(y), \tag{3.13}$$

where A does not depend on θ (canonical parameter) and ϕ (dispersion parameter). If the model of interest only contains main effects, the link function $g(\cdot)$ is expressed by

$$g(E(Y)) = \eta = \beta_0 + \beta_1 X_1 + ... + \beta_P X_P.$$

Suppose that the dispersion parameter ϕ in (3.13) is known, the joint probability density function of $(Y_1,\ldots,Y_n)$ can be expressed by $f^{(n)}(y_1,\ldots,y_n;\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_P)$

are unknown parameters. The likelihood function of $\boldsymbol{\beta}$ is given by

$$L(\boldsymbol{\beta}) = f^{(n)}(y_1, \ldots, y_n; \boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \boldsymbol{\beta}). \tag{3.14}$$

Let $\mathcal{S}_j$ denote the set of all models containing j covariates and $\tau(\mathcal{S}_j)$ denote the size of set $\mathcal{S}_j$. In this way, the model space $\mathcal{S}$ is the disjoint union of sets $\mathcal{S}_1, \ldots, \mathcal{S}_P$. The extended family of Bayesian information criteria (EBIC) in a GLM is given by

$$\begin{aligned} EBIC_\gamma(M) &= -2\log(L(\hat{\beta}_0, \hat{\boldsymbol{\beta}}(M))) + v(M)\log(n) + 2\gamma\log(\tau(\mathcal{S}_{v(M)})) \\ &= BIC(M) + 2\gamma\log(\tau(\mathcal{S}_{v(M)})),\ 0 \le \gamma \le 1, \end{aligned} \tag{3.15}$$

where $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}(M))$ is the maximum likelihood estimates of unknown parameters based on model M. As demonstrated in (3.15), the EBIC method is equivalent to adding one term $2\gamma\log(\tau(\mathcal{S}_{v(M)}))$ to the BIC method. The item $2\gamma\log(\tau(\mathcal{S}_{v(M)}))$ can be seen as a penalty function. It is expected that the EBIC method would discourage the model with too many spurious covariates.

Besides main effects of single covariates, two-covariate interactions may be also non-negligible factors related to the response variable. For instance, some studies have shown that interactions among loci may contribute broadly to common disease, so gene-gene interactions are suggested to be considered in genome-wide association studies. More specifically, some significant two-covariate interactions may have little or no main effects at each single covariate. In these situations, these covariates could not be detected if we only consider main effects. Hence, it is more powerful to consider the model with both main effects and two-covariate interactions in terms of identification

of significant covariates. When two-covariate interactions are included in the models of interest, the number of unknown parameters increases from P to $P(P+1)/2$. In this situation, the dimension of model space is high even if the number of covariates is not large. To be specific, when the number of covariates P is 20, the dimension of model space $P(P+1)/2$ has reached 210. In view of high dimensional model space, the extended Bayesian information criteria are likely more appropriate. In the following, we modify the extended family of Bayesian information criteria such that it is suitable to model selection in generalized linear regression models with both main effects and interactions.

If the generalized linear model of interest contains not only main effects but also two-covariate interactions, the corresponding link function is expressed by

$$g(E(Y|X)) = \alpha + \sum_{j=1}^{P} \beta_j X_j + \sum_{k=1}^{P} \sum_{l \neq k} \xi_{kl} X_k X_l.$$

Moreover, the likelihood function of $(\alpha,\ \boldsymbol{\beta},\ \boldsymbol{\xi})$ is given by

$$L(\alpha, \boldsymbol{\beta},\ \boldsymbol{\xi}) = f^{(n)}(y_1, \ldots, y_n; \alpha, \boldsymbol{\beta},\ \boldsymbol{\xi}) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i, \alpha, \boldsymbol{\beta},\ \boldsymbol{\xi}),$$

where $f(y_i|\mathbf{x}_i,\ \alpha, \boldsymbol{\beta},\ \boldsymbol{\xi})$ is the probability density function of the response Y_i.

In model selection, either main effects or interactions can be considered as possible factors related with the response variable Y. However, the effects of selecting a main effect and selecting an interaction are different. Selecting a main effect means the corresponding covariate is included in the model, whereas selecting an interaction is equivalent to

involve two corresponding covariates in the model. Thus, it would be more appropriate to give different emphases on main effects and interactions. As mentioned earlier, the extended family of Bayesian information criteria (3.15) is equivalent to put one more penalty function on the ordinary Bayesian information criterion. This penalty function $2\gamma \log(\tau(\mathcal{S}_{\nu(M)}))$ depends on $\nu(M)$, the number of factors in model M. To emphasize different effects of one main effect and one interaction, we modify the extended family of Bayesian information criteria by penalizing model M instead with two parts of penalty functions. One part is the penalty for its main effects and the other is for its interactions.

Let $\mathcal{S}^1_{n_1}$ denote the set of models containing n_1 main effects but no interaction, and $S^2_{n_2}$ denote the set of models containing n_2 interactions but no main effect. Suppose the number of covariates under consideration is P. The size of $\mathcal{S}^1_{n_1}$,$\tau(\mathcal{S}^1_{n_1})$ is equal to $C_P^{n_1}$ and the size of $\mathcal{S}^2_{n_2}$, $\tau(\mathcal{S}^2_{n_2})$ is equal to $C_{P(P-1)/2}^{n_2}$. Let $v_1(M)$ and $v_2(M)$ denote the number of main effects and the number of interactions in model M. In this way, the elements of model M is the combination of $v_1(M)$ main effects and $v_2(M)$ interactions. Instead of the penalty function $2\gamma \log(\tau(\mathcal{S}_{\nu(M)}))$ in the expression (3.15), the EBIC method penalizes model M with $2\gamma_1 \log(\tau(\mathcal{S}^1_{v_1(M)})) + 2\gamma_2 \log(\tau(\mathcal{S}^2_{v_2(M)}))$. The item $2\gamma_1 \log(\tau(\mathcal{S}^1_{v_1(M)}))$ is the penalty function for the part of main effects, while the item $2\gamma_2 \log(\tau(\mathcal{S}^2_{v_2(M)}))$ is the penalty function for the part of interactions. Hence, in generalized linear models with both main effects and two-covariate interactions, the extended Bayesian information

criteria are expressed by

$$\begin{aligned} EBIC(M)_{(\gamma_1,\gamma_2)} &= -2\log L(\hat{\alpha},\hat{\boldsymbol{\beta}}(M),\hat{\boldsymbol{\xi}}(M)) + (\nu_1(M)+\nu_2(M))\log(n) \\ &+ 2\gamma_1\log(\tau(\mathcal{S}^1_{\nu_1(M)})) + 2\gamma_2\log(\tau(\mathcal{S}^2_{\nu_2(M)})),\ 0\le\gamma_1\le 1,\ 0\le\gamma_2\le 1, \quad (3.16) \end{aligned}$$

where $(\hat{\alpha},\ \hat{\boldsymbol{\beta}}(M),\ \hat{\boldsymbol{\xi}}(M))$ is the maximized likelihood estimate of the unknown parameter vector $(\alpha,\ \boldsymbol{\beta}(M),\ \boldsymbol{\xi}(M))$.

In this section, we described the extended Bayesian information criteria in the context of generalized linear models. The expression (3.15) is applicable to generalized linear models only with main effects, while the expression (3.16) is used in case of considering both main effects and two-covariates interactions. In the next section, we illustrate the performances of the extended family of Bayesian information criteria (3.15) and (3.16) by some simulation studies.

3.3 Simulation studies

In simulation studies, we consider logistic regression models. In logistic regression model, the response variable Y follows Bernoulli distribution with mean μ and variance $\mu(1-\mu)$. The link function is $g(\mu) = \log(\mu/(1+\mu))$. Two sets of simulations are conducted to respectively assess EBIC in (3.15) and (3.16). In the first set, the response variable Y follows the Bernoulli distribution with the parameter $\exp(\beta_0+\beta_1X_1+\ldots+\beta_PX_P)/(1+\exp(\beta_0+\beta_1X_1+\ldots+\beta_PX_P))$. In this set of simulations, we present the results

of $EBIC_\gamma$ in (3.15) with three γ values 0, $1 - \log(n)/2\log(P)$ and 1. In the second set of simulations, not only main effects but also two-covariate interactions are considered. In this way, the responses are generated from Bernoulli$\{\exp(\eta)/(1 + \exp(\eta))\}$, where $\eta = \alpha + \sum_{j=1}^{P} X_j\beta_j + \sum_{k=1}^{P}\sum_{l\neq k} X_kX_l\xi_{kl}$. Similarly, the parameter vector (γ_1, γ_2) in (3.16) choose three values: (0, 0), $(1 - \log(n)/2\log(P),\ 1 - \log(n)/2\log(P(P-1)/2))$ and (1, 1). In these two sets of simulations, the number of covariates P is set to be less than the sample size n. The simulation studies for the cases that $P > n$ will be conducted in Chapter 5.

There are two popular methods in the penalized likelihood methodology: the SCAD method and the Lasso method. It has been shown that the SCAD method outperforms the Lasso method in terms of selecting features (Fan and Li, 2001). In the first set of simulations, the modified SCAD method in Chapter 2 is used to select models. The favorable property of the modified SCAD method is that it not only maintains the performance of the original SCAD method but also always produces finite parameter estimates even in case of separation. The penalized likelihood function of the modified SCAD method is given by

$$l_{MS}(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) + \frac{1}{2}\log|I(\boldsymbol{\beta})| - n\sum_{j=1}^{P} p_\lambda(|\beta_j|),$$

where $L(\boldsymbol{\beta})$ is the likelihood function, $I(\boldsymbol{\beta})$ is the Fisher information matrix, $p_\lambda(|\beta_j|)$ is the SCAD penalty function of β_j, and λ is the tuning parameter. In the second setting, the number of total factors is up to $P(P+1)/2$ since both main effects and two-covariate interactions are considered. Although the number of covariates P is less than the sample

size n, the number $P(P+1)/2$ is more than n when P is large enough. The maximum likelihood estimate is a good initial value for the SCAD method but cannot be obtained when the number of unknown parameters is more than the sample size. Moreover, it is known that the performance of the SCAD method depends on the initial value. Thus, in the second setting, the L_1 penalized likelihood method is used to select variables. It is expressed by

$$l_{\lambda_1}(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi}) = \log(L(\alpha, \boldsymbol{\beta}, \boldsymbol{\xi})) - n \sum_{j=1} \lambda_1 |\beta_j| - n \sum_{k=1} \sum_{l \neq k} \lambda_1 |\xi_{kl}|.$$

In these two settings, the tuning parameter λ or λ_1 is increased gradually from 0 such that only a covariate is deleted at one time. Thus, a sequence of nested models is obtained. Then, the un-penalized parameter estimates are obtained by maximizing the log-likelihood function. On the basis of the un-penalized parameter estimates, the extended Bayesian information criteria are calculated for each model. The best model is the one which minimizes the extended Bayesian information criteria.

Example 3.1 : In this example, 200 replicates consisting of 200 observations with 100 cases and 100 controls are simulated. In each simulation replicate, four covariates $X_1, \ldots, X_4$ are set to nonzero coefficients, while others are assigned to 0. The intercept item β_0 is set to -5. The significant covariates $X_1, \ldots, X_4$ independently follows Bernoulli distributions with success probabilities $(0.13, 0.15, 0.16, 0.20)$. For each i from 1 to 4, X_{i+4} is correlated with X_i and generated from the conditional probabilities $p(X_{i+4} = 1|X_i = 1) = 0.93$ and $p(X_{i+4} = 1|X_i = 0) = 0.07$. The other $(P-8)$ covariates $X_9, \ldots, X_P$ are independently distributed as Bernoulli distributions. Their success

probabilities are randomly selected from the uniform distribution $U(0.20, 0.50)$. In this example, there are two settings with different true coefficient vectors, one is

$$(\underbrace{1.70, 1.60, 1.50, 1.40}_{4}, \underbrace{0, \ldots, 0}_{(P-4)}),$$

and the other one is

$$(\underbrace{0.90, 0.85, 0.80, 0.95}_{4}, \underbrace{0, \ldots, 0}_{(P-4)}).$$

The results of these two settings with the number of candidate covariates $P = 50,\ 100$ are respectively reported in Tables 3.1 and 3.2. The column labeled "Correct" denotes the average restricted to the true nonzero coefficients, and the column labeled "Incorrect" presents the average of coefficients erroneously set to nonzero. The standard deviations for "Correct" and "Incorrect" based on the 200 simulation replicates are presented in the parentheses. The columns PSR and FDR denote the positive selection rate and false discovery rate. Note that methods MSCAD^0, MSCAD*, MSCAD^1 are the modified SCAD method cum EBIC (3.15) with γ being 0, $1 - \log(n)/(2\log(P))$, 1.

Table 3.1: Simulation results for logistic model only with main effects-1

Method	Avg. No of nonzero coefficients		PSR	FDR
	Correct	Incorrect		
P=50				
MSCAD^0	3.72(0.55)	1.53(1.51)	0.930	0.291
MSCAD*	3.67(0.59)	0.85(1.16)	0.918	0.188
MSCAD^1	3.57(0.63)	0.35(0.64)	0.893	0.078
P=100				
MSCAD^0	3.55(0.64)	2.59(2.54)	0.888	0.422
MSCAD*	3.53(0.63)	0.91(1.20)	0.883	0.205
MSCAD^1	3.37(0.73)	0.39(0.67)	0.843	0.104

Note: the number in the parentheses denotes the standard deviation based on 200 replications

Table 3.2: Simulation results for logistic model only with main effects-2

Method	Avg. No of nonzero coefficients		PSR	FDR
	Correct	Incorrect		
P=50				
MSCAD^0	2.68(0.92)	1.49(1.49)	0.670	0.357
MSCAD*	2.42(0.93)	0.85(1.04)	0.605	0.260
MSCAD^1	1.67(0.93)	0.26(0.51)	0.448	0.131
P=100				
MSCAD^0	2.65(0.89)	2.73(2.27)	0.663	0.507
MSCAD*	2.38(0.96)	1.01(1.78)	0.595	0.298
MSCAD^1	1.54(0.88)	0.23(0.49)	0.385	0.130

From Tables 3.1 and 3.2, it can be seen that both the positive selection rate (PSR) and the false discovery rate (FDR) decrease as the parameter γ in (3.15) increases. This is our expected result. The main reason for PSR and FDR decreasing is that the large penalty function discourages to select the model with too many variables. Since the size of the set $\mathcal{S}_{v(M)}$, $\tau(\mathcal{S}_{v(M)})$ is a constant when model M is given, the penalty function of EBIC in (3.15) is an increasing function in the parameter γ. Moreover, the methods MSCAD^0, MSCAD* and MSCAD^0 correspond to the values of the parameter γ in EBIC being 0, $0 < 1 - \log(n)/(2\log(P)) < 1$ and 1 respectively. Therefore, the penalty function in these three methods have a declining trend.

It is known that EBIC with $\gamma = 0$ is equivalent to the ordinary BIC method. As shown in Tables 3.1 and 3.2, the false discovery rates (FDRs) of the modified SCAD method cum BIC are high in all cases considered. This phenomenon is more prominent when the number of candidate covariates (P) increases to 100. In the second setting, the FDR

of the modified SCAD method cum BIC in case of $P = 100$ has reached to 0.507, which means more than one half features selected by MSCAD^0 are spurious. In contrast, the false discovery rate is effectively controlled around 0.10 by MSCAD^1 even when P increases to 100. In general, the positive selection rate (PSR) of the modified SCAD method cum BIC (MSCAD^0) is higher than those of the SCAD method cum EBIC (MSCAD* and MSCAD^1). However, in the first setting, the PSR of MSCAD^1 is only slightly lower than that of MSCAD^0.

Example 3.2 : In this example, 200 datasets consisting of 400 observations with 200 cases and 200 controls are generated from the model $Y \sim Bernoulli\{p(\beta_0 + \sum_{j=1}^{P} \mathbf{x}_j\beta_j + \sum_{k=1}^{P}\sum_{l\neq k} \mathbf{x}_k\mathbf{x}_l\xi_{kl})\}$, where $p(\mu) = \exp(\mu)/(1 + \exp(\mu))$, the intercept β_0 is set to -5. The structure of covariates is similar to that in Example 3.1 whereas the number of significant covariates is increased to 6. The last two covariate (X_5, X_6) are significant in main effects, while the first four ($X_1, \ldots, X_4$) are significant in their two-covariate interactions X_{12} and X_{34}. The success probabilities of six significant covariates are (0.140, 0.145, 0.150, 0.155, 0.160, 0.165). The conditional distribution of correlated covariates is same as that in Example 3.1. In this example, there are also two settings, one with nonzero coefficients

$$\beta_5 = 1.2,\ \beta_6 = 1.1,\ \xi_{12} = 1.6,\ \xi_{34} = 1.4,$$

the other with

$$\beta_5 = 0.3,\ \beta_6 = 1.2,\ \xi_{12} = 0.5,\ \xi_{34} = 1.4.$$

The results of these two settings with the number of candidate covariates $P = 50,\ 100$ are reports in Table 3.3 and 3.4. Note that methods $L_1 - \text{penalty}^0$, $L_1 - \text{penalty}^*$, $L_1 - \text{penalty}^1$ are the L_1 penalized likelihood method with (γ_1, γ_2) in EBIC (2.16) being $(0, 0)$, $(1 - \log(n)/(2\log(P)), 1 - \log(n)/(2\log(P(P-1)/2)))$ and $(1, 1)$.

Table 3.3: Simulation results for logistic model with main effects and interactions-1

Method	Avg. No of nonzero coefficients		PSR	FDR
	Correct	Incorrect		
P=50				
$L_1 - \text{penalty}^0$	5.72(0.52)	3.58(3.20)	0.953	0.385
$L_1 - \text{penalty}^*$	5.31(0.91)	1.45(1.09)	0.885	0.214
$L_1 - \text{penalty}^1$	4.68(1.22)	0.90(0.81)	0.780	0.161
P=100				
$L_1 - \text{penalty}^0$	5.52(0.84)	6.92(5.70)	0.920	0.556
$L_1 - \text{penalty}^*$	4.74(1.21)	1.20(1.10)	0.790	0.202
$L_1 - \text{penalty}^1$	4.16(1.48)	0.72(0.73)	0.693	0.148

Table 3.4: Simulation results for logistic model with main effects and interactions-2

Method	Avg. No of nonzero coefficients		PSR	FDR
	Correct	Incorrect		
P=50				
$L_1 - \text{penalty}^0$	3.37(0.81)	5.19(5.34)	0.562	0.606
$L_1 - \text{penalty}^*$	2.88(0.56)	0.61(0.80)	0.480	0.175
$L_1 - \text{penalty}^1$	2.69(0.67)	0.41(0.63)	0.448	0.132
P=100				
$L_1 - \text{penalty}^0$	3.54(0.91)	11.15(7.43)	0.590	0.759
$L_1 - \text{penalty}^*$	2.78(0.64)	0.68(0.91)	0.463	0.197
$L_1 - \text{penalty}^1$	2.50(0.82)	0.42(0.66)	0.417	0.144

Tables 3.3 and 3.4 also show that either the positive selection rate (PSR) or the false discovery rate (FDR) have a decreasing tendency as γ_1 and γ_2) in EBIC (3.16) are in-

creased from 0 to 1. The FDRs of $L_1-\text{penalty}^0$ are intolerably high in both two settings. However, the FDR of $L_1-\text{penalty}^1$ does not exceed 0.17 even in the worst case. The performance of $L_1-\text{penalty}^*$ is between those of $L_1-\text{penalty}^0$ and $L_1-\text{penalty}^1$ in both the positive selection rate and the false discovery rate.

Simulation results described in Table 3.1-3.4 suggest that the ordinary Bayesian information criterion (Schwarz, 1978) may not be appropriate in a generalized linear model with mediate-size model space. The main reason is that it tends to select too many spurious covariates. Moreover, the performance of the extended Bayesian information criteria in generalized linear regression models is consistent with the earlier finding suggesting that the EBIC method performs better than the ordinary BIC method in linear regression model with medium dimensional model space (Chen and Chen, 2007).

3.4 Summary

We have discussed the extended Bayesian information criteria in the context of generalized linear regression models. If both main effects and two-covariates interaction are considered as possible factors, we suggest that the extended Bayesian information criteria impose different penalties on main effects and interactions. This modification allows the main effect item to enter the model more easily in comparison with the interaction item. It supports the fact that the effect of selecting one interaction is to involve two corresponding covariates in the model.

Although the extended Bayesian information criteria were originally proposed by Chen and Chen (2007), the model of interest was limited to linear regression model only with main effects. This chapter has provided clear evidence that the EBIC method is more appropriate than BIC in generalized linear models when the dimension of the model space is high. Moreover, this work would make the EBIC method more popular as an appropriate criterion in high dimensional model selection.

Chapter 4

The Generalized Tournament Screening Cum EBIC Approach

In this chapter, we introduce the generalized tournament screening cum EBIC approach for high dimensional feature selection in generalized linear models. Its basic idea is to combine the dimension reduction with model selection. The generalized tournament approach can deal with feature space consisting of main effects and interaction effects. One characteristic of the generalized tournament approach is that it transfers a high dimensional feature selection problem to some low dimensional feature selection problems. This characteristic ensures that it is applicable whatever the dimension of feature space is. In addition, the generalized tournament approach is computationally feasible since it selects features not individually but in groups. We suggest using the penalized likelihood methodology for selecting features in the whole procedure and using the ex-

tended Bayesian information criteria (EBIC; Chapter 3) as model selection criterion to determine the best model, i.e., the significant variables.

In the following sections, the generalized tournament screening cum EBIC approach is described in more details. In Section 4.1, we briefly introduce the generalized tournament approach and propose two strategies for tackling interaction effects. The first step, pre-screening, is described in Section 4.2. The second step, final selection, is introduced in Section 4.3. We give a summary for the generalized tournament screening cum EBIC approach in Section 4.4.

4.1 Introduction to the generalized tournament screening cum EBIC approach

Chen and Chen (2007) developed the tournament screening cum EBIC approach for high dimensional feature selection. They only discussed the approach in the context of linear models with main effects. However, high dimensional generalized linear models are widely used in many fields such as medical and genome-wide association studies. Moreover, besides main effects, interaction effects are likely to play an important role in explaining the response variable. Thus, we propose the generalized tournament screening cum EBIC approach to provide a solution to high dimensional feature selection in the context of generalized linear models. The method effectively copes with both main

effects and interaction effects.

The generalized tournament screening cum EBIC approach combines the dimension reduction with model selection and has two corresponding steps. The first step is called pre-screening. The aim is to reduce a high dimensional feature space to a low dimensional feature space, i.e., dimension reduction. The second step is called final selection. Its aim is to select significant variables by identifying the best model. Compared to the number of main effects, the number of interaction effects is much larger. This is more prominent when the number of variables under consideration increases. Hence, we propose two strategies to deal with interaction effects.

In the pre-screening step, we consider two strategies for interaction effects: the two-stage strategy and the full strategy. For the two-stage strategy, we only consider main effects in the first stage and select a pre-specified number of them. In the second stage, only main effects and interactions involving the variables selected in the first stage are considered. The two-stage strategy should work well if interactions between variables are such that the marginal effects of the variables are still sizable though the interaction effects dominate. However, it will miss the variables that have significant interaction effects but no or little marginal effects. The full strategy is to consider all possible main effects and interaction effects at the beginning of screening. Compared with the two-stage strategy, the full strategy will take more computational time in the pre-screening step. However, the full strategy should be able to pick up the variables with significant

interaction effects but no or little marginal effects. In the following part, we will only describe the procedure for the full strategy. For the two-stage strategy, the only difference is that only main effects are considered in the first stage.

The basic idea of the generalized tournament approach is as follows. In the full strategy, main effects and interactions are selected separately in the pre-matches. Both main effects and interaction effects are screened by using a penalized likelihood method in a sequence of stages. Main effects are randomly divided into groups and the screening procedure is carried out from group to group. The procedure is repeated until the number of main effects is reduced to a desirable level. And then, interaction effects are selected in the same way. All variables corresponding to either the selected main effects or the selected interaction effects in the pre-matches are tentatively selected for the semi-final stage. In the semi-final, only a part of them survive and enter the final stage. At the final stage, a sequence of nested models are generated by a penalized likelihood method and the best model is determined by optimizing a model selection criterion.

4.2 The procedure of the pre-screening step

Pre-matches for main effects

- **Round 1**: Let $\mathcal{X}^1$ be the variables that represent all main effects. Partition $\mathcal{X}^1$ at random into subsets of nearly equal size M to yield

$$\mathcal{X}^1 = \mathcal{X}^1_1 \cup ... \cup \mathcal{X}^1_J,$$

where J is the integer such that $[JM]$ equals to the total number of main effects. For each subset $\mathcal{X}^1_j$, let $\mathbf{x}_j$ denote the vector of main effects in $\mathcal{X}^1_j$ and $\boldsymbol{\beta}_j$ denote unknown parameters corresponding to $\mathbf{x}_j$. Carry out a main effect selection step as follows. With a properly tuned value λ^*_j, maximize

$$l_p(\boldsymbol{\beta}_j, \mathbf{0} \mid \mathbf{x}_j, \mathbf{0}) = l(\boldsymbol{\beta}_j, \mathbf{0} \mid \mathbf{x}_j, \mathbf{0}) - n \sum_k p_{\lambda^*_j}(|\beta_k|)$$

to yield m_1 nonzero fitted components of $\boldsymbol{\beta}_j$. A variable is tentatively selected and entered the next round of the matches, if its corresponding parameter estimate is nonzero.

- **Round $\mathbf{r_1}$**: Repeat the same procedure as in the round 1 with the set $\mathcal{X}^{r_1}$ generated from the previous round until the number of variables representing main effects reaches a desirable level. Assume that after the round r_1, m variables of main effects are selected and formed the set $\mathcal{X}^* = \{X^*_1, \ldots, X^*_m\}$.

The previous r_1 rounds is the pre-matches for the variables representing main effects. Similarly, we do the pre-matches for interaction effects in the following rounds. The difference is that variables in the set $\mathcal{X}^*$ will be included in each step.

Pre-matches for interaction effects

- **Round $\mathbf{r_1 + 1}$**: Let $\mathcal{V}^1$ be the set of variables representing interaction effects between any two variables in the data set. Partition $\mathcal{V}^1$ at random into subsets of nearly equal size H to yield

$$\mathcal{V}^1 = \mathcal{V}^1_1 \cup ... \cup \mathcal{V}^1_K,$$

  where K is the integer such that $[KH]$ is the total number of interactions. For each subset $\mathcal{V}^1_j$, let $\mathbf{v}_j$ denote the vector of interaction effects in $\mathcal{V}^1$ and $\boldsymbol{\xi}_j$ denote unknown parameters corresponding to $\mathbf{v}_j$. Carry out an interaction effect selection step as follows. With a properly tuned value τ^*_j, maximize

$$l_p(\boldsymbol{\beta}^*, \xi_j \mid \mathbf{x}^*, \mathbf{v}_j) = l(\boldsymbol{\beta}^*, \xi_j \mid \mathbf{x}^*, \mathbf{v}_j) - n \sum_k p_{\tau^*_j}(|\beta^*_k|) - n \sum_l p_{\tau^*_j}(|\xi_l|)$$

  to yield h_1 nonzero fitted components of $\boldsymbol{\xi}_j$.

- **Round $\mathbf{r_2}$**: Repeat the same procedure as in the round $r_1 + 1$ with the set $\mathcal{V}^{r_2}$ generated from the previous round until the number of interaction effects reaches a desired level. Assume that after the round r_2, h interaction effects are selected and formed the set $\mathcal{V}^* = \{V^*_1, \ldots, V^*_h\}$.

After the round r_2, there are m main effects and h interaction effects selected from the feature space. One variable is tentatively selected, if either its main effect or its interaction effect with any other variable is selected in the pre-matches. All the tentatively

selected variables are collected in a set entering the semi-final stage.

Semi-final

- The semi-final stage begins with the variables left from the pre-matches. Denote the sets of main-effect variables and interaction-effect variables respectively by $\mathbf{x}^*$ and $\mathbf{v}^*$. Denote their corresponding unknown parameters by $\boldsymbol{\beta}^*$ and $\boldsymbol{\xi}^*$. Maximize

$$l_p(\boldsymbol{\beta}^*, \boldsymbol{\xi}^* \mid \mathbf{x}^*, \mathbf{v}^*) = l(\boldsymbol{\beta}^*, \boldsymbol{\xi}^* \mid \mathbf{x}^*, \mathbf{v}^*) - n \sum_k p_\lambda(|\beta_k^*|) - n \sum_l p_\lambda(|\xi_l^*|)$$

  respectively with respect $\boldsymbol{\beta}^*$ and $\boldsymbol{\xi}^*$ with a properly tuned value of λ to produce a pre-specified number $K(< n)$ of variables. Note that the same tuning parameter is used for both main effects and interaction effects in the semi-final stage.

In each stage of pre-screening, the penalized likelihood methodology is applied to select main effects and interaction effects. There are many existing penalized likelihood methods such as the SCAD method, the L_1 penalized method. The performances of these methods in feature selection are different due to different properties of penalty functions. In the pre-screening step, a huge number of variables representing both main effects and interaction effects needed to consider. Thus, we suggest to use the L_1 penalized method due to its high computational efficiency. The GLM path algorithm (Park *et al.*, 2007) is based on the L_1 penalty function and effectively computes solutions among the entire regularization path for generalized linear models. Thus, it is efficient for the GLM path algorithm to select a specified number of variables.

4.3 The procedure of the final selection step

After the pre-screening step, the generalized tournament screening has already reduced the dimension of feature space from $P(P+1)/2$ to $K(< n)$. It is known that when the dimension of feature space is smaller than the sample size, conventional penalized likelihood methods can be directly used to select variables. Unlike the pre-screening step, the number of variables in the final selection is small, so the computational burden is not a challenge in this step. In contrast, the selection performance is more important than the computational efficiency. It has been shown that the SCAD method enjoys many favorable theoretical properties and a good performance on model selection (Fan and Li, 2001). Hence, the SCAD method is suggested to use in the final selection. When the generalized tournament approach is applied in a logistic regression model, the SCAD method will be replaced by the modified SCAD method in Chapter 2. The modified SCAD method always produces the finite parameter estimate even in case of complete separation. The SCAD method/the modified SCAD method is used to generate a sequence of models by tuning parameter value and some model selection criterion is used to determine the best model. Finally, all variables entered into the best model are considered to be significant whether they appear in the form of main effects or interaction effects. Although the dimension of feature space has been reduced to a number less than the sample size, the best model is selected from the all candidate models from the feature space with P main effects and the total $P(P-1)/2$ interaction effects. We have discussed the extended Bayesian information criteria (Chapter 3) could be more

appropriate in high dimensional feature selection for generalized linear models. Thus, the extended Bayesian information criteria are suggested to be model selection criterion in the final selection step. In the remainder of this section, we describe the procedures of generating a sequence of nested models and identifying the best model in details.

A sequence of nested models

- Let the full model be the first model in the sequence. It consists of all the main effects and interactions selected from the semi-final stage. Use the same penalty parameter for both the main effects and interactions in the SCAD penalized log-likelihood function. The tuning parameter λ is increased to λ_1 which is the smallest value to make one component of the parameter estimate obtained by minimizing to be zero. The corresponding model is the second model in the sequence. In the following, by increasing the penalty parameter λ, delete the effects one at a time sequentially until that the last model only contains the intercept term. The effect of yielding a sequence of nested models is to rank all effects from the semi-final stage.

Model fitting and selection

- Refit all models by maximizing un-penalized log-likelihood function. Compute for each model the model selection criterion EBIC and choose the best model. The best model is the one which optimizes the extended Bayesian information

criterion. Finally, one variable is declared to be associated with the response variable, if its main effects or its interaction with any other variable is contained in the best model.

The modified SCAD method and the extended Bayesian information criteria have been introduced in Chapter 2 and 3 respectively, so we omit the details here.

4.4 Summary

In this chapter, we have introduced the generalized tournament screening cum EBIC approach for high dimensional feature selection in generalized linear models. Compared with feature selection methods based on multiple testing, it not only avoids dilemma of choosing the overall threshold but also considers multiple joint effects. Therefore, it is expected that the generalized tournament screening cum EBIC is a more appropriate approach than multiple testing for genome-wide association studies.

We need to tackle a huge number of variables representing main effects and interaction effects in high dimensional feature selection. These variables are evaluated in groups, not individually, in the generalized tournament approach. Moreover, the GLM path algorithm computes the entire regularization path sequentially, thereby avoiding independent optimization at different values of the tuning parameter λ.

In the final selection step, we suggest to utilize the SCAD penalty function to yield the sequence of nested models due to its favorable theoretical properties and good selection performance. If the model of interest is a logistic model, the original SCAD method will be replaced by the modified SCAD method (Chapter 2) since it always produces finite parameter estimates even in case of separation. In addition, the number of spurious variables contained in the best model would be effectively controlled by using the extended information criteria (EBIC) (Chapter 3).

Chapter 5

The Application of the Generalized Tournament Approach in Genome-wide Association Studies

In this chapter, we apply the generalized tournament screening cum EBIC (Chapter 4) in genome-wide association studies for detecting genetic variants. Empirical evidence suggests that interactions among loci may play an important role in explaining many common diseases. The generalized tournament approach not only identifies gene-gene interactions but also is beneficial to the data consisting of high dimensional genotypes and a binary disease status. We compare the performance of the generalized tournament approach with that of the multiple testing of all possible pairwise gene-gene interactions. Through simulated datasets, we demonstrate that the generalized tournament

approach outperforms the multiple testing in terms of the positive selection rate as well as the false discovery rate. In addition, our method enjoys high computational efficiency in comparison with the multiple testing.

We review the multiple testing of all possible pairwise gene-gene interactions in Section 5.1. We explore the use of the generalized tournament approach in genome-wide association studies in Section 5.2. Some aspects of genetics related to numerical studies are introduced in Section 5.3. Our method are illustrated with simulated datasets in Section 5.4. We conclude with a summary in Section 5.5.

5.1 Introduction to the multiple testing for genome-wide association studies

More and more studies demonstrate possible importance of interactions among loci in genome-wide association studies, but most traditional analytical methods only consider each genetic marker or haplotype individually. The multiple testing of all possible pairwise gene-gene interactions proposed by Marchini, *et al*. (2005) is a recent technique for detecting gene-gene interactions associated with many common complex diseases. The log-likelihood ratio test is used to evaluate interactions (Balding, *et al*., 2001) and Bonferroni correction is used to declare the genome-wide significance. The multiple testing assesses all possible interactions by using the following three strategies.

Strategy I: Single locus. Fit a logistic model at each locus. For example, for locus A, define $\delta_1 = \begin{cases} 1 & \text{if AA} \\ 0 & \text{otherwise} \end{cases}$ and $\delta_2 = \begin{cases} 1 & \text{if Aa} \\ 0 & \text{otherwise} \end{cases}$. The linear predictor in the logistic model is $\eta = \beta_0 + \beta_1\delta_1 + \beta_2\delta_2$. Use a Bonferroni correction to set the significance level to be α/L to nominally control the overall type I error at level α. Evaluate this strategy by two criteria: (i) requiring that at least one of the two loci meet the significance threshold, irrespective of the other locus, or (ii) requiring that both loci are significant.

Strategy II: Full interaction. Fit a logistic regression model at each pair of loci. Use a Bonferroni correction to set the significance level at α/C_L^2.

Strategy III: Two-stage. Identify all loci that are significant in single-locus tests (as strategy I) at a liberal level α_1 in the first stage. Call this set of loci $I_1 \subseteq \{1, \ldots, L\}$. Let d_1 be the degree of freedom of the single-locus model fitted at stage one for locus l (maximum 2 degrees of freedom if all three genotypes are present) and define k_l such that $P(\chi^2_{d_i} > k_l) = \alpha_1$ for $l \subseteq I_1$. In the second stage, for each pair of loci l and m identified in stage one, calculate the log-likelihood ratio statistic $R_{(l,m)}$ for the full interaction model. Use a new statistic $R'_{(l,m)} = R_{(l,m)} - (k_l + k_m)$ to assess the significance of this statistic against a $\chi^2_{d'}$ distribution in which d' is the degrees of freedom of the full model fitted at the two loci. Set the level of significance using a Bonferroni correction based

on the expected number of tests to be done ($\alpha/C^2_{\alpha_1 \times L}$).

The multiple testing of all possible pairwise gene-gene interactions is a simple analytical method and is more powerful to detect disease gene than single locus method when interactions exist. However, it requires an extremely small p-value to claim the genome-wide significance, so the power of detecting the influenced loci is likely to be low. In the multiple testing, only the last two strategies are based on gene-gene interactions and have been shown to outperform the first strategy, so we only compare our method with these two strategies in numerical studies.

5.2 The generalized tournament screening cum EBIC approach for genome-wide association studies

The data from genome-wide association studies consist of genotype measurements and a binary disease status indicating whether a subject is the affected or unaffected. In this situation, logistic regression model is a natural tool for describing the relationship between the disease status and the locus genotypes. Let the binary response variable Y

be defined as,

$$Y = \begin{cases} 1 & \text{if diseased, i.e., case} \\ 0 & \text{otherwise, i.e., control} \end{cases}$$

Let $\mathbf{X}$ denote the vector of variables representing the genotypes of all considered SNPs. Let $\mathbf{V}$ denote the vector of variables representing the product terms of the variables in $\mathbf{X}$. Assume that there are n independent samples with $n/2$ cases and $n/2$ controls. The logistic model formulates that the probability density function of the joint distribution of $\{Y_i,\ i = 1, .., n\}$ is given by

$$f(y_1, \ldots, y_n) = \Pi_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)},$$

where

$$\pi_i = \frac{\exp\{\alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{v}_i^T \boldsymbol{\xi}\}}{1 + \exp\{\alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{v}_i^T \boldsymbol{\xi}\}},$$

or

$$\log \frac{\pi_i}{1 - \pi_i} = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{v}_i^T \boldsymbol{\xi}. \tag{5.1}$$

In the following, we propose a penalized logistic model imposed on both main effects and interaction effects. Let $l(\boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{X}, \mathbf{V}) = \log L(\boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{X}, \mathbf{V})$. Explicitly,

$$l(\boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{X}, \mathbf{V}) = \sum_{i=1}^{n} [y_i(\alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{v}_i^T \boldsymbol{\xi}) - \log(1 + \exp\{\alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{v}_i^T \boldsymbol{\xi}\})]. \tag{5.2}$$

The penalized log-likelihood function is given by

$$l_p(\boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{X}, \mathbf{V}) = l(\boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{X}, \mathbf{V}) - n \sum_k p_\lambda(|\beta_k|) - n \sum_l \sum_m p_\tau(|\xi_{lm}|), \tag{5.3}$$

where $p_\lambda(.)$ is the penalty function of main affects and $p_\tau(.)$ is the penalty function of interaction effects, β_k are components of $\boldsymbol{\beta}$ and ξ_{lm} are the component of $\boldsymbol{\xi}$.

The goal of genome-wide association studies is to identify genetic variants associated with a particular disease. However, one challenge is that only a few out of a huge number of considered SNPs contributed to the disease. It is equivalent to select significant features from all features contained in model (5.1). The generalized tournament screening cum EBIC approach (Chapter 4) can effectively select features from high dimensional feature space. In genome-wide association studies, the separation phenomenon always occurs in the data. The Jeffreys prior penalty can be incorporated into the penalized likelihood function to handle this phenomenon. Therefore, the generalized tournament screening cum EBIC approach is appropriate in genome-wide association studies. In the remainder of this section, we briefly describe its basic procedure in the context of genome-wide association studies.

First, all the variables representing the genotype of all the SNPs and their products screened by the generalized tournament procedure with L_1-penalized likelihood function until that the number of variables is reduced to a desired level. Then, these selected variables are ranked by the modified SCAD method, and a sequence of nested models are generated. Finally, the extended Bayesian information criteria are used to identify the best model. If either its main effect or any interaction with an other SNP is in the best model, this SNP is considered to be significant. The details of the generalized tournament screening cum EBIC has been introduced in Chapter 4.

5.3 Some genetical aspects

In this section, we introduce some aspects of genetics: the marginal effect, the prevalence of a disease, the linkage disequilibrium and the statistical models representing gene-gene interactions. These definitions and models will be used in numerical studies.

Assume that the two causal loci (A and B) are under linkage equilibrium and the disease allele frequencies at loci A and B are π_A and π_B. The prevalence of a disease is the probability of a disease in the population. Let p denote the prevalence of a disease and be expressed by

$$p = p(D) = \sum_{g_A, g_B} p(D|g_A, g_B)p(g_A, g_B), \tag{5.4}$$

where g_A and g_B denote the genotype at locus A and B, $p(D|g_A, g_B)$ is the conditional probability that an individual has the disease given that they have genotype g_A at locus A and genotype g_B at locus B, $p(g_A, g_B)$ denotes the joint probability of genotypes g_A and g_B and is equal to the product of $p(g_A)$ and $p(g_B)$ if loci A and B are under linkage equilibrium.

The marginal odds ratio at locus A is given by

$$\frac{p(D|g_A)}{p(\bar{D}|g_A)} = \frac{\sum_{g_B} p(D|g_A, g_B)p(g_B)}{\sum_{g_B} p(\bar{D}|g_A, g_B)p(g_B)}, \tag{5.5}$$

where $p(D|g_A)$ is the conditional probability that an individual has the disease given that they have genotype g_A at locus A, $p(\bar{D}|g_A)$ is the conditional probability that an individual does not have the disease given that they have genotype g_A at locus A. The

parameter λ_A represents the marginal effect of locus A. It is given by

$$\lambda_A = \frac{p(D|1_A)}{p(\bar{D}|1_A)} / \frac{p(D|0_A)}{p(\bar{D}|0_A)} - 1, \tag{5.6}$$

where 0_A and 1_A respectively denote the genotypes aa and Aa in the locus A. From (5.5), it can be seen that the parameter λ_A depends on the disease allele frequency of locus B, the conditional probabilities of disease and non-disease given the genotypes.

Linkage disequilibrium (LD) is a measure of association between alleles of two different genes. For a general discussion, suppose a disease locus A has alleles A, a and a marker locus X has two alleles X and x. The linkage disequilibrium can be reflected by the conditional haplotype probability. The allele frequencies of A and X are π_A and π_X. There are three different constraints for the conditional probabilities $p(X|A)$ and $p(X|a)$:

(Constraint 1) $p(X|A) = q$, $p(X|a) = 1 - q$

(Constraint 2) $p(X|A) = 1$, $p(X|a) = q$

(Constraint 3) $p(X|A) = q$, $p(X|a) = 0$

The square correlation coefficients r^2 (Pritchard and Przeworski, 2001) is a parameter to measure the magnitude of LD. It is expressed by

$$r^2_{AX} = [p(X|A) - p(X|a)]^2 \frac{\pi_A(1-\pi_A)}{\pi_X(1-\pi_X)} \tag{5.7}$$

Given the value of the square correlation coefficients r^2, the conditional probabilities $p(X|A)$, $p(X|a)$ can be calculated by (5.7). In the following, we introduce four statistical models for two-locus interactions and then calculate the expression of the marginal effect parameter λ_A for each model.

There are a variety of general two-locus models mimicking simple biological mechanisms. In numerical studies, we choose four models for the comparison of the generalized tournament approach with the multiple testing. These four models range from situation in which both marginal effects and interaction effect of the two loci exist to those in which only interaction effect exists without marginal effects presenting. The first three models are demonstrated by the conditional odds given the genotypes of two disease loci under the epistatic scenarios considered by Marchini, *et al*. (2005). The last interaction model that have no or little marginal effects at each locus has been studied by some recent work (Hoh and Ott, 2003; Culverhouse, *et al*., 2002; Moore and Ritchie, 2004).

Model 1: two-locus interaction multiplicative effects

	aa	Aa	AA
bb	α	α	α
Bb	α	$\alpha(1+\theta)$	$\alpha(1+\theta)^2$
BB	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

The entries in the table are conditional odds given the genotypes of both loci, e.g., $p(D|\text{aa, bb})/p(\bar{D}|\text{aa, bb}) = \alpha$ etc. In model 1, the conditional odds have a baseline value (α) unless both loci have at least one disease allele. From the conditional odds, it can be seen that loci A and B affect disease in their interaction and they have the same marginal effects. The log conditional odds of the interaction effect can be expressed

by $\eta(g_A, g_B) = \log(\alpha) + \log(1+\theta)N_A N_B$, where N_A denote the number of allele A in genotype g_A and N_B denote the number of allele B in genotype g_B. On the basis of the definition (5.5), we calculate the marginal odds ratio of locus A.

$$\begin{aligned}
\frac{p(D|0_A)}{p(\bar{D}|0_A)} &= \frac{p(D|0_A,0_B)p(0_B)+p(D|0_A,1_B)p(1_B)+p(D|0_A,2_B)p(2_B)}{p(\bar{D}|0_A,0_B)p(0_B)+p(\bar{D}|0_A,1_B)p(1_B)+p(\bar{D}|0_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha}{1+\alpha}(1-\pi_B)^2+\frac{\alpha}{1+\alpha}2\pi_B(1-\pi_B)+\frac{\alpha}{1+\alpha}\pi_B^2}{\frac{1}{1+\alpha}(1-\pi_B)^2+\frac{1}{1+\alpha}2\pi_B(1-\pi_B)+\frac{1}{1+\alpha}\pi_B^2} \\
&= \alpha
\end{aligned}$$

$$\begin{aligned}
\frac{p(D|1_A)}{p(\bar{D}|1_A)} &= \frac{p(D|1_A,0_B)p(0_B)+p(D|1_A,1_B)p(1_B)+p(D|1_A,2_B)p(2_B)}{p(\bar{D}|1_A,0_B)p(0_B)+p(\bar{D}|1_A,1_B)p(1_B)+p(\bar{D}|1_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha}{1+\alpha}(1-\pi_B)^2+\frac{\alpha(1+\theta)}{1+\alpha(1+\theta)}2\pi_B(1-\pi_B)+\frac{\alpha(1+\theta)^2}{1+\alpha(1+\theta)^2}\pi_B^2}{\frac{1}{1+\alpha}(1-\pi_B)^2+\frac{1}{1+\alpha(1+\theta)}2\pi_B(1-\pi_B)+\frac{1}{1+\alpha(1+\theta)^2}\pi_B^2} \\
&= \alpha(1+\frac{\frac{2\theta\pi_B(1-\pi_B)}{1+\alpha(1+\theta)}+\frac{\theta(2+\theta)\pi_B^2}{1+\alpha(1+\theta)^2}}{\frac{(1-\pi_B)^2}{1+\alpha}+\frac{2\pi_B(1-\pi_B)}{1+\alpha(1+\theta)}+\frac{\pi_B^2}{1+\alpha(1+\theta)^2}}) \\
&= \alpha(1+\lambda_1)
\end{aligned}$$

$$\begin{aligned}
\frac{p(D|2_A)}{p(\bar{D}|2_A)} &= \frac{p(D|2_A,0_B)p(0_B)+p(D|2_A,1_B)p(1_B)+p(D|2_A,2_B)p(2_B)}{p(\bar{D}|2_A,0_B)p(0_B)+p(\bar{D}|2_A,1_B)p(1_B)+p(\bar{D}|2_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha}{1+\alpha}(1-\pi_B)^2+\frac{\alpha(1+\theta)^2}{1+\alpha(1+\theta)^2}2\pi_B(1-\pi_B)+\frac{\alpha(1+\theta)^4}{1+\alpha(1+\theta)^4}\pi_B^2}{\frac{1}{1+\alpha}(1-\pi_B)^2+\frac{1}{1+\alpha(1+\theta)^2}2\pi_B(1-\pi_B)+\frac{1}{1+\alpha(1+\theta)^4}\pi_B^2} \\
&= \alpha(1+\frac{\frac{2\theta(\theta+2)\pi_B(1-\pi_B)}{1+\alpha(1+\theta)^2}+\frac{((1+\theta)^4-1)\pi_B^2}{1+\alpha(1+\theta)^2}}{\frac{(1-\pi_B)^2}{1+\alpha}+\frac{2\pi_B(1-\pi_B)}{1+\alpha(1+\theta)^2}+\frac{\pi_B^2}{1+\alpha(1+\theta)^4}}) \\
&= \alpha(1+\lambda_2)
\end{aligned}$$

The parameter λ_A is expressed by

$$\lambda_A = \lambda_1 = \frac{\frac{2\theta\pi_B(1-\pi_B)}{1+\alpha(1+\theta)}+\frac{\theta(2+\theta)\pi_B^2}{1+\alpha(1+\theta)^2}}{\frac{(1-\pi_B)^2}{1+\alpha}+\frac{2\pi_B(1-\pi_B)}{1+\alpha(1+\theta)}+\frac{\pi_B^2}{1+\alpha(1+\theta)^2}} \tag{5.8}$$

The parameter λ_B for locus B can be given by the symmetric formula of (5.8).

Model 2: two-locus interaction threshold effects

	aa	Aa	AA
bb	α	α	α
Bb	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
BB	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$

In this model, at least one disease alleles is required to have a effect beyond α. However, unlike model 1, this model does not increase the risk as the number of disease alleles increases. Hence, this model specifies a threshold of disease effects rather than multiplicative gene action. The log conditional odds can be expressed by $\eta(g_A, g_B) = \log(\alpha) + \log(1+\theta) I_A I_B$, where I_A is the indicator whether the genotype g_A involves the allele A and I_B is the indicator whether the genotype g_B involves the allele B.

$$\begin{aligned}
\frac{p(D|0_A)}{p(\bar{D}|0_A)} &= \frac{p(D|0_A,0_B)p(0_B)+p(D|0_A,1_B)p(1_B)+p(D|0_A,2_B)p(2_B)}{p(\bar{D}|0_A,0_B)p(0_B)+p(\bar{D}|0_A,1_B)p(1_B)+p(\bar{D}|0_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha}{1+\alpha}(1-\pi_B)^2+\frac{\alpha}{1+\alpha}2\pi_B(1-\pi_B)+\frac{\alpha}{1+\alpha}\pi_B^2}{\frac{1}{1+\alpha}(1-\pi_B)^2+\frac{1}{1+\alpha}2\pi_B(1-\pi_B)+\frac{1}{1+\alpha}\pi_B^2} \\
&= \alpha
\end{aligned}$$

$$\begin{aligned}
\frac{p(D|1_A)}{p(\bar{D}|1_A)} &= \frac{p(D|1_A,0_B)p(0_B)+p(D|1_A,1_B)p(1_B)+p(D|1_A,2_B)p(2_B)}{p(\bar{D}|1_A,0_B)p(0_B)+p(\bar{D}|1_A,1_B)p(1_B)+p(\bar{D}|1_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha}{1+\alpha}(1-\pi_B)^2+\frac{\alpha(1+\theta)}{1+\alpha(1+\theta)}\pi_B(2-\pi_B)+\frac{\alpha(1+\theta)^2}{1+\alpha(1+\theta)^2}\pi_B^2}{\frac{1}{1+\alpha}(1-\pi_B)^2+\frac{1}{1+\alpha(1+\theta)}\pi_B(2-\pi_B)} \\
&= \alpha\left(1+\frac{\frac{\theta\pi_B(2-\pi_B)}{1+\alpha(1+\theta)}}{\frac{(1-\pi_B)^2}{1+\alpha}+\frac{\pi_B(2-\pi_B)}{1+\alpha(1+\theta)}}\right) \\
&= \alpha(1+\lambda_1)
\end{aligned}$$

$$\begin{aligned}
\frac{p(D|2_A)}{p(\bar{D}|2_A)} &= \frac{p(D|2_A,0_B)p(0_B)+p(D|2_A,1_B)p(1_B)+p(D|2_A,2_B)p(2_B)}{p(\bar{D}|2_A,0_B)p(0_B)+p(\bar{D}|2_A,1_B)p(1_B)+p(\bar{D}|2_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha}{1+\alpha}(1-\pi_B)^2+\frac{\alpha(1+\theta)}{1+\alpha(1+\theta)}\pi_B(2-\pi_B)+\frac{\alpha(1+\theta)^2}{1+\alpha(1+\theta)^2}\pi_B^2}{\frac{1}{1+\alpha}(1-\pi_B)^2+\frac{1}{1+\alpha(1+\theta)}\pi_B(2-\pi_B)} \\
&= \alpha(1+\frac{\frac{\theta\pi_B(2-\pi_B)}{1+\alpha(1+\theta)}}{\frac{(1-\pi_B)^2}{1+\alpha}+\frac{\pi_B(2-\pi_B)}{1+\alpha(1+\theta)}}) \\
&= \alpha(1+\lambda_2)
\end{aligned}$$

The parameter λ_A can be expressed by

$$\lambda_A = \lambda_1 = \frac{\frac{\theta\pi_B(2-\pi_B)}{1+\alpha(1+\theta)}}{\frac{(1-\pi_B)^2}{1+\alpha}+\frac{\pi_B(2-\pi_B)}{1+\alpha(1+\theta)}}. \tag{5.9}$$

Model 3: multiplicative within and between loci model

	aa	Aa	AA
bb	α	$\alpha(1+\theta_A)$	$\alpha(1+\theta_A)^2$
Bb	$\alpha(1+\theta_B)$	$\alpha(1+\theta_A)(1+\theta_B)$	$\alpha(1+\theta_A)^2(1+\theta_B)$
BB	$\alpha(1+\theta_B)^2$	$\alpha(1+\theta_A)(1+\theta_B)^2$	$\alpha(1+\theta_A)^2(1+\theta_B)^2$

In this model, the odds increase multiplicatively with the number of disease alleles both within and between loci. From the odds of disease, it can be seen that two loci A and B affect disease independently. The log conditional odds of the interaction effect can be expressed by $\eta(g_A,g_B)=\log(\alpha)+\log(1+\theta_A)N_A+\log(1+\theta_B)N_B$, where N_A denote the number of allele A in genotype g_A and N_B denote the number of allele B in genotype g_B.

$$\begin{aligned}
\frac{p(D|0_A)}{p(\bar{D}|0_A)} &= \frac{p(D|0_A,0_B)p(0_B)+p(D|0_A,1_B)p(1_B)+p(D|0_A,2_B)p(2_B)}{p(\bar{D}|0_A,0_B)p(0_B)+p(\bar{D}|0_A,1_B)p(1_B)+p(\bar{D}|0_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha}{1+\alpha}(1-\pi_B)^2+\frac{\alpha(1+\theta_B)}{1+\alpha(1+\theta_B)}2\pi_B(1-\pi_B)+\frac{\alpha(1+\theta_B)^2}{1+\alpha(1+\theta_B)^2}\pi_B^2}{\frac{1}{1+\alpha}(1-\pi_B)^2+\frac{1}{1+\alpha(1+\theta_B)}2\pi_B(1-\pi_B)+\frac{1}{1+\alpha(1+\theta_B)^2}\pi_B^2} \\
&= \alpha(1+\frac{\frac{2\theta_B\pi_B(1-\pi_B)}{1+\alpha(1+\theta_B)}+\frac{\theta_B(2+\theta_B)\pi_B^2}{1+\alpha(1+\theta_B)^2}}{\frac{(1-\pi_B)^2}{1+\alpha}+\frac{2\pi_B(1-\pi_B)}{1+\alpha(1+\theta_B)}+\frac{\pi_B^2}{1+\alpha(1+\theta_B)^2}}) \\
&= \alpha(1+\mu_1)
\end{aligned}$$

$$\begin{aligned}
\frac{p(D|1_A)}{p(\bar{D}|1_A)} &= \frac{p(D|1_A,0_B)p(0_B)+p(D|1_A,1_B)p(1_B)+p(D|1_A,2_B)p(2_B)}{p(\bar{D}|1_A,0_B)p(0_B)+p(\bar{D}|1_A,1_B)p(1_B)+p(\bar{D}|1_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha(1+\theta_A)}{1+\alpha(1+\theta_A)}(1-\pi_B)^2+\frac{\alpha(1+\theta_A)(1+\theta_B)}{1+\alpha(1+\theta_A)(1+\theta_B)}2\pi_B(1-\pi_B)+\frac{\alpha(1+\theta_A)(1+\theta_B)^2}{1+\alpha(1+\theta_A)(1+\theta_B)^2}\pi_B^2}{\frac{1}{1+\alpha(1+\theta_A)}(1-\pi_B)^2+\frac{1}{1+\alpha(1+\theta_A)(1+\theta_B)}2\pi_B(1-\pi_B)+\frac{1}{1+\alpha(1+\theta_A)(1+\theta_B)^2}\pi_B^2} \\
&= \alpha(1+\frac{\frac{2\theta_B\pi_B(1-\pi_B)}{1+\alpha(1+\theta_A)(1+\theta_B)}+\frac{\theta_B(2+\theta_B)\pi_B^2}{1+\alpha(1+\theta_A)(1+\theta_B)^2}}{\frac{(1-\pi_B)^2}{1+(1+\theta_A)\alpha}+\frac{2\pi_B(1-\pi_B)}{1+\alpha(1+\theta_A)(1+\theta_B)}+\frac{\pi_B^2}{1+\alpha(1+\theta_A)(1+\theta_B)^2}}) \\
&= \alpha(1+\theta_A)(1+\mu_2)
\end{aligned}$$

$$\begin{aligned}
\frac{p(D|2_A)}{p(\bar{D}|2_A)} &= \frac{p(D|2_A,0_B)p(0_B)+p(D|2_A,1_B)p(1_B)+p(D|2_A,2_B)p(2_B)}{p(\bar{D}|2_A,0_B)p(0_B)+p(\bar{D}|2_A,1_B)p(1_B)+p(\bar{D}|2_A,2_B)p(2_B)} \\
&= \frac{\frac{\alpha(1+\theta_A)^2}{1+\alpha(1+\theta_A)^2}(1-\pi_B)^2+\frac{\alpha(1+\theta_A)^2(1+\theta_B)}{1+\alpha(1+\theta_A)^2(1+\theta_B)}2\pi_B(1-\pi_B)+\frac{\alpha(1+\theta_A)^2(1+\theta_B)^2}{1+\alpha(1+\theta_A)^2(1+\theta_B)^2}\pi_B^2}{\frac{1}{1+\alpha(1+\theta_A)^2}(1-\pi_B)^2+\frac{1}{1+\alpha(1+\theta_A)^2(1+\theta_B)}2\pi_B(1-\pi_B)+\frac{1}{1+\alpha(1+\theta_A)^2(1+\theta_B)^2}\pi_B^2} \\
&= \alpha(1+\frac{\frac{2\theta_B\pi_B(1-\pi_B)}{1+\alpha(1+\theta_A)^2(1+\theta_B)}+\frac{\theta_B(2+\theta_B)\pi_B^2}{1+\alpha(1+\theta_A)^2(1+\theta_B)^2}}{\frac{(1-\pi_B)^2}{1+(1+\theta_A)^2\alpha}+\frac{2\pi_B(1-\pi_B)}{1+\alpha(1+\theta_A)^2(1+\theta_B)}+\frac{\pi_B^2}{1+\alpha(1+\theta_A)^2(1+\theta_B)^2}}) \\
&= \alpha(1+\theta_A)^2(1+\mu_3)
\end{aligned}$$

The parameter λ_A can be expressed by

$$\lambda_A = \frac{(1+\mu_2)(1+\theta_A)}{(1+\mu_1)} - 1 \tag{5.10}$$

Model 4: significant interaction effect with negligible marginal effects

In this model, the log conditional odds of the interaction effect can be expressed by

$$\eta(g_A, g_B) = \beta_0 + \beta_1 N_A + \beta_2 N_B + \xi_{12} N_A N_B \tag{5.11}$$

subject to

$$\sum_{g_B} \eta(AA, g_B)p(g_B) = \sum_{g_B} \eta(Aa, g_B)p(g_B) = \sum_{g_B} \eta(aa, g_B)p(g_B) \tag{5.12}$$

and

$$\sum_{g_A} \eta(g_A, BB)p(g_A) = \sum_{g_A} \eta(g_A, Bb)p(g_A) = \sum_{g_A} \eta(g_A, bb)p(g_A). \tag{5.13}$$

Given the value of ξ_{12}, the parameters β_1 and β_2 can be calculated by (5.11)-(5.13). The marginal effect of locus A being zero is equivalent to that the marginal odds at different genotype g_A is a constant.

In the numerical studies, we will generate the simulated datasets on the basis of these four models. For the first three models, given disease allele frequencies π_A and π_B, the prevalence of a disease p and the marginal effect parameter λ_A and λ_B, we can calculate the value of α and θ (θ_A, θ_B) contained in the condition odds. However, the expected marginal effect of the last model is equal to zero. Hence, for the last model, we give the unknown parameter ξ_{12} and the prevalence of a disease, and calculate the other unknown parameters $\beta_0, \beta_1, \beta_2$ by (5.11)-(5.13).

5.4 Numerical Studies

In this section, the results of two sets of numerical studies are presented. In the first set, the generalized tournament screening cum EBIC approach with different penalties are compared with each other, and they are also compared with the multiple testing on their performances of identifying causal loci. In the second set, we report the performance of the generalized tournament approach in simulated datasets with some more complex structures of SNPs.

5.4.1 Numerical study 1

The first goal of this numerical study is to compare the performances of the extended Bayesian information criteria with three different parameter values, i.e., $(\gamma_1, \gamma_2) = (0,0), (1-\log(n)/(2*\log(P)), 1-\log(n)/(2*\log(P(P-1)/2)))$ and $(1,1)$. The value (0,0) corresponding to the Bayesian information criterion. The value (1,1) is the most stringent value for EBIC. The value $(1-\log(n)/(2*\log(P)), 1-\log(n)/(2*\log(P(P-1)/2)))$ is another choice suggested by Chen and Chen (2007). The second goal is to compare the performances of the generalized tournament screening cum EBIC and the multiple testing for gene-gene interactions. We select four gene-gene interaction models introduced in Section 5.2 for our comparison. The first three models were used by Marchini *et al.* (2005) to compare the multiple testing for gene-gene interactions with other traditional single marker analyses. In these three gene-gene interaction models, the significant interactions between genetic loci have non-negligible marginal effects in two individual loci. The last gene-gene interaction model has no or little marginal effects at each locus.

Each dataset contains $n = 800$ samples (400 cases and 400 controls). In this set of numerical studies, $P = 1000$ and $P = 5000$ SNPs are considered. In all SNPs under consideration, there is only one pair of SNPs that contribute to the disease status. Assume that two causal loci have same effects on the disease. For the first three models, the effect of each causal loci is specified by the marginal effect parameter λ in (5.6). Since the marginal effect is zero in the last model, the effect of each causal loci is spec-

ified by the parameter ξ_{12} in (5.11). The susceptibility allele at each causal locus is in linkage disequilibrium (LD) with a particular single marker allele. The remaining SNPs are assumed to be independent and in Hardy-Weinberg equilibrium. When we generate the genotypes of two causal loci and two linked markers, linkage disequilibrium between a disease allele and one marker allele is implemented by specifying the square correlation coefficients $r^2 = 0.5$. For all data sets, the overall proportion of the disease population, i.e., the prevalence is set to be 0.01. We compare the performances of different methods by using two measures: the positive selection rate (PSR) and the false discovery rate (FDR) (Banjamini and Hochberg, 1995). The positive selection rate is defined as the proportion of the truly associated covariates selected. The false discovery rate is defined as the proportion of falsely selected covariates among all selected ones. In genome-wide associated studies, the positive selection rate is defined as the proportion of the truly associated SNPs selected. It is similar to the power in hypothesis testing. The false discovery rate is defined as the proportion of falsely selected SNPs among all selected ones, which is an alterative measure to the probability of type I error in the hypothesis testing. All simulations are conducted by using R package.

The average positive selection rate (PSR) and the average false discovery rate (FDR) based on 100 replications for four models are respectively summarized in the following tables. Let GT1 denote the generalized tournament screening cum EBIC with $(\gamma_1, \gamma_2) = (0, 0)$, GT2 denote the generalized tournament screening cum EBIC with $(\gamma_1, \gamma_2) = (1 - \log(n)/(2 * \log(P)), 1 - \log(n)/(2 * \log(P(P-1)/2)))$, GT3 denote the

generalized tournament screening cum EBIC with $(\gamma_1, \gamma_2) = (1, 1)$ and MT denote the multiple testing method.

Table 5.1: The average PSR for "Two-locus interaction multiplicative effects" model

(n, P)	λ	disease allele frequency	PSR			
			GT1	GT2	GT3	MT
(800,1000)	0.5	0.1	0.735	0.450	0.265	0.175
	0.5	0.2	0.890	0.805	0.650	0.550
	0.7	0.1	0.905	0.860	0.790	0.710
	0.7	0.2	0.980	0.950	0.950	1.000
(800,5000)	0.5	0.1	0.665	0.335	0.175	0.085
	0.5	0.2	0.880	0.695	0.610	0.405
	0.7	0.1	0.915	0.815	0.720	0.480
	0.7	0.2	0.965	0.940	0.940	0.930

Table 5.2: The average FDR for "Two-locus interaction multiplicative effects" model

(n, P)	λ	disease allele frequency	FDR			
			GT1	GT2	GT3	MT
(800,1000)	0.5	0.1	0.958	0.286	0.086	0.352
	0.5	0.2	0.939	0.134	0.071	0.763
	0.7	0.1	0.941	0.149	0.048	0.758
	0.7	0.2	0.927	0.095	0.050	0.954
(800,5000)	0.5	0.1	0.977	0.221	0.079	0.595
	0.5	0.2	0.967	0.151	0.077	0.928
	0.7	0.1	0.968	0.163	0.062	0.776
	0.7	0.2	0.963	0.121	0.051	0.980

Table 5.1 and 5.2 describe the average positive selection rate (PSR) and false discovery rate (FDR) of two-locus interaction multiplicative effects model. It can be seen that both the positive selection rate (PSR) and the false discovery rate (FDR) decrease as

the parameter γ in EBIC increases. The reason is that the large penalty function discourages to select the model with too many variables. Table 5.2 shows that the false discovery rates (FDRs) of the generalized tournament screening cum BIC (GT1) are intolerably high in all cases considered. The lowest one has already reached to 0.941, which means 94.1 percentage of SNPs selected by GT1 are spurious. Thus, it is likely that the Bayesian information criterion is not appropriate in high dimensional model space mainly because it tends to select too many spurious variables. However, the false discovery rate is effectively controlled around 0.20 by GT2. Furthermore, the false discovery rate of GT3 does not exceed 0.10 even in the worst case. As shown in Table 5.1, the positive selection rate of the generalized tournament screening cum BIC (GT1) is higher than those of the generalized tournament screening cum EBIC (GT2 and GT3) in general. However, the positive selection rate (PSR) of GT2 is slightly lower than that of GT1 in some cases. Especially, in the sixth case, the PSR of GT2 is 0.940, which is very close to that of GT1, 0.965. Thus, the extended Bayesian information criteria could be more reasonable than the original Bayesian information (Schward, 1978) in high dimensional generalized linear models with main effects and interactions. It is consistent with the earlier finding suggesting that the EBIC method performs better than the ordinary BIC method in linear model with main effects (Chen and Chen, 2007).

Table 5.1 and 5.2 also demonstrate that the generalized tournament screening cum EBIC (GT2 and GT3) enjoys high positive selection rate (PSR) and low false discovery rate (FDR) in comparison with the multiple testing method (MT). Bonferroni correction is

very conservative when the number of hypothesis tests is huge. It accounts for the low positive selection rate of the multiple testing. The low false discovery rate of the generalized tournament approach could be attributed to the penalized likelihood methodology and the extended Bayesian information criteria. One causal SNP may make its interactions with many non-causal SNPs highly correlated with the response variable. The penalized likelihood methodology assesses interaction effects in groups, so the joint effects among interactions are considered. However, the multiple testing evaluates interaction effects individually, which likely incurs too many spurious SNPs. The results of lower PSR and higher FDR of the multiple tests approach show that the generalized tournament method cum EBIC may perform better than the multiple testing method (Marchini *et al*., 2005) in genome-wide association studies. Consequently, the generalized tournament screening cum EBIC could become a promising way to detect genetic variants associated with many diseases.

As shown in Table 5.1, the total number of candidate SNPs (P) impacts on the positive selection rates (PSRs) of all methods. The larger number of candidate SNPs is corresponding to the lower positive selection rate (PSR). The possible reason is that it is more difficult to select the causal SNPs from more candidate SNPs. This result is most prominent in the multiple testing method. For instance, the PSR of the multiple testing method in the second case is 0.710, while its PSR has declined to 0.480 in the fifth case where the number of candidate SNPs (P) increases to 5000. It would be a result of the Bonferroni adjustment. For instance, if the number of SNPs is 1000, any

interaction whose P-value is less than $0.05/(500 \times 999)$ is declared to be significant. If the number of SNPs increases to 5000, only interactions whose P-value are less than $0.05/(2500 \times 4999)$ are declared the significance. From Table 5.1, it can be seen that the average positive selection rate is also affected by the allele frequency. Generally speaking, the small allele frequency corresponds to the low positive selection rate. In the following simulations, we fix the value 0.1 for the allele frequency.

Table 5.3: The average PSR for "Two-locus interaction threshold effects" model

(n, P)	λ	disease allele frequency	PSR			
			GT1	GT2	GT3	MT
(800,1000)	0.8	0.1	0.840	0.655	0.530	0.455
	0.9	0.1	0.920	0.835	0.730	0.695
	1.0	0.1	0.930	0.895	0.810	0.840
(800,5000)	0.8	0.1	0.865	0.530	0.350	0.270
	0.9	0.1	0.910	0.720	0.620	0.490
	1.0	0.1	0.960	0.813	0.712	0.657

Table 5.4: The average FDR for "Two-locus interaction threshold effects" model

(n, P)	λ	disease allele frequency	FDR			
			GT1	GT2	GT3	MT
(800,1000)	0.8	0.1	0.954	0.229	0.086	0.884
	0.9	0.1	0.954	0.204	0.052	0.965
	1.0	0.1	0.951	0.179	0.047	0.970
(800,5000)	0.8	0.1	0.961	0.159	0.028	0.800
	0.9	0.1	0.971	0.459	0.101	0.999
	1.0	0.1	0.955	0.195	0.060	0.982

Table 5.3 and 5.4 describe the average positive selection rate (PSR) and false discovery rate (FDR) of two-locus interaction threshold effects model. The result is similar to

that of two-locus interactions multiplicative effects model. The generalized tournament screening cum EBIC enjoys high positive selection rate (PSR) and low false discovery rate (FDR) in comparison with the multiple testing method (MT). The difference is that the positive selection rate in two-locus interaction multiplicative effects model is higher than that in two-locus interaction threshold effects model. It is account for different odds in these two models.

Table 5.5: The average PSR for "Multiplicative within and between loci" model

(n, P)	λ	disease allele frequency	PSR			
			GT1	GT2	GT3	MT
(800,1000)	0.8	0.1	0.980	0.860	0.610	0.780
	0.9	0.1	0.990	0.940	0.850	0.900
	1.0	0.1	0.990	0.980	0.960	1.000
(800,5000)	0.8	0.1	0.960	0.740	0.470	0.660
	0.9	0.1	0.980	0.890	0.750	0.870
	1.0	0.1	0.960	0.940	0.890	0.930

Table 5.6: The average FDR for "Multiplicative within and between loci" model

(n, P)	λ	disease allele frequency	FDR			
			GT1	GT2	GT3	MT
(800,1000)	0.8	0.1	0.964	0.423	0.358	0.996
	0.9	0.1	0.959	0.343	0.320	0.998
	1.0	0.1	0.953	0.242	0.219	0.999
(800,5000)	0.8	0.1	0.923	0.460	0.405	0.999
	0.9	0.1	0.922	0.414	0.380	0.999
	1.0	0.1	0.922	0.273	0.233	0.999

Table 5.5 and 5.6 summarize the average positive selection rate (PSR) and false discovery rate (FDR) of multiplicative within and between loci model. As shown in Table

5.6, the false discovery rate of the multiple testing is intolerably high. The lowest has already reached 0.996 and the highest is 0.999. The high false discovery rate makes the multiple testing inappropriate. Although the false discovery rate of GT3 is not controlled below 0.1 like in model 1 and 2, its false discovery rate is much lower than that of the multiple testing. The main reason for higher false discovery rate in this model is that the causal loci can be detectable independent of other loci and cause many interactions highly correlated with the response variable. From Table 5.5, it can be seen that the positive selection rate of GT3 is lower than that of the multiple testing. It is likely that some gene-gene interactions between causal loci and other non-causal loci sometimes enter the model before the main effect. However, if we fix a specific false discovery rate, the positive selection rate of the generalized tournament approach must be higher than that of the multiple testing.

Table 5.7: The average PSR for "Interactions with negligible marginal effects" model

(n, P)	ξ_{12}	disease allele frequency	PSR			
			GT1	GT2	GT3	MT
(800,1000)	1.9	0.1	0.990	0.955	0.828	0.702
	2.0	0.1	1.000	0.985	0.945	0.860
	2.1	0.1	0.995	0.975	0.965	0.915
(800,5000)	1.9	0.1	0.990	0.810	0.555	0.460
	2.0	0.1	0.995	0.930	0.730	0.710
	2.1	0.1	0.995	0.970	0.885	0.795

Table 5.7 and 5.8 summarize the average positive selection rate and the average false discovery rate for significant interaction effect with negligible marginal effects model.

Table 5.8: The average FDR for "Interactions with negligible marginal effects" model

(n, P)	ξ_{12}	disease allele frequency	FDR			
			GT1	GT2	GT3	MT
(800,1000)	1.9	0.1	0.789	0.031	0.012	≥0.550
	2.0	0.1	0.791	0.034	0.026	≥0.641
	2.1	0.1	0.795	0.025	0.015	≥0.915
(800,5000)	1.9	0.1	0.792	0.024	0.009	≥0.406
	2.0	0.1	0.783	0.026	0.014	≥0.427
	2.1	0.1	0.775	0.015	0.006	≥0.562

For the multiple testing, we only test all interactions involving causal SNPs, which guarantees the same positive selection rate but produces a less or equal false discovery rate. Hence, we use " ≥ " in the column representing the FDR of the multiple testing. Table 5.7 and 5.8 shows that positive selection rate of GT3 is higher than that of the multiple testing, and the false discovery rate is lower than that of the multiple testing. This is a similar result with model 1 and 2. However, it can be seen that the false discovery rates for these four methods are lower than corresponding FDRs in the first three models. This may be accounted by no or little marginal effects in this model.

5.4.2 Numerical study 2

In the second set of numerical studies, we present the performance of the generalized tournament screening cum EBIC in some simulated datasets with more complex structures. The datasets also contains 800 samples with 400 cases and 400 controls, but the number of SNPs under consideration is increased to 10000. Moreover, the number of causal SNPs is set to be 10. The allele at each causal locus is in LD with a par-

ticular single marker allele. The remaining SNPs are assumed to be independent and in Hardy-Weinberg equilibrium. When we generate the genotypes of causal loci and linked markers, linkage disequilibrium between a disease allele and one marker allele is specified by the square correlation coefficients $r^2 = 0.5$. The disease allele frequencies for the ten causal loci are fixed at:

$$\pi = (0.15, 0.21, 0.09, 0.12, 0.13, 0.18, 0.10, 0.14, 0.08, 0.16).$$

There are two structures for interaction effects in the datasets.

In the first structure, the ten causal SNPs affect the disease by five independent interactions: 1 multiplicative within and between loci effect (Model 3), 2 two-locus interaction multiplicative effects (Model 1) and 2 two-locus interaction threshold effects (Model 2). The marginal effect parameters of the ten causal loci are specified as

$$\lambda = (1.02, 0.88, 1.02, 0.76, 0.93, 0.77, 1.17, 0.80, 1.52, 0.69)$$

and

$$\lambda = (0.82, 0.72, 0.86, 0.66, 0.84, 0.69, 1.04, 0.71, 1.26, 0.58)$$

in two settings. The prevalence p is set to be 0.01. Let $g_i,\ i = 1, \ldots, 10$ denote the genotypes of the ten disease loci. Let $N_i,\ i = 1, \ldots, 10$ be the variables representing the number of disease alleles in the genotype of the i-th locus. Let $I_j,\ j = 1, \ldots, 10$ denote the variables indicating whether the genotype of the j-th locus involves the disease allele. The log conditional odds given the genotypes of ten disease loci can be expressed

by

$$\log(\eta(g_1,\ldots,g_{10})) = \log(\alpha) + \log(\theta_1)N_1 + \log(\theta_2)N_2 + \log(\theta_3)N_3N_4 + \log(\theta_4)N_5N_6 + \log(\theta_5)I_7I_8 + \log(\theta_6)I_9I_{10}$$

Given the value of the marginal effect parameter vector λ, the disease allele frequency vector π and the prevalence p, we can calculate the value of the parameters $(\alpha, \theta_1, \ldots, \theta_6)$ in the conditional odds. These two simulation results are summarized in Table 5.9 and 5.10 respectively. The column labeled "Correct" presents the average restricted to the truly associated SNPs, and the column "Incorrect" depicts the average of wrongly selected associated SNPs.

Table 5.9: Simulation results for the first structure

Method	Correct(SD)	Incorrect(SD)	PSR	FDR
Setting $1-1$				
GT1	9.53(0.67)	27.66(6.57)	0.953	0.744
GT2	7.52(1.16)	1.28(1.82)	0.752	0.145
GT3	7.09(1.18)	0.46(0.77)	0.709	0.061
Setting $1-2$				
GT1	9.06(0.93)	31.94(8.55)	0.906	0.779
GT2	6.26(1.25)	1.09(1.71)	0.626	0.148
GT3	5.97(1.28)	0.38(0.83)	0.597	0.060

In the second structure, the ten causal SNPs affect the disease by five independent interactions: 1 multiplicative within and between loci effect (Model 1) and 4 significant

interactions with no or little marginal effects (Model 4). In this structure, there are 4 interactions with no or little marginal effects. Thus, instead of the marginal effect parameters, the coefficients in the logistic regression model are used to specify the effects of disease loci. The response variable follows the Bernoulli distribution with parameter $\eta(g_1, \ldots, g_{10})/(1 + \eta(g_1, \ldots, g_{10}))$, where the log conditional odds is given by

$$\log(\eta(g_1, \ldots, g_{10})) = \alpha + \beta_1 N_1 + \ldots + \beta_{10} N_{10} + \xi_{12} N_1 N_2 + \ldots + \xi_{9,10} N_9 N_{10}.$$

In this two settings, the corresponding coefficients for main effects and interaction effects are respectively specified as:

$$\boldsymbol{\beta} = (0.78, 0.69, -0.40, -0.54, -0.58, -0.80, -0.45, -0.62, -0.71, -0.36),$$

$$\boldsymbol{\xi} = (0.00, 2.23, 1.93, 2.06, 2.37).$$

and

$$\boldsymbol{\beta} = (0.53, 0.45, -0.31, -0.42, -0.45, -0.62, -0.35, -0.48, -0.55, -0.28),$$

$$\boldsymbol{\xi} = (0.00, 1.73, 1.45, 1.76, 1.67)$$

The intercept α is set to be -5.30 in both two settings.

From Table 5.9 and 5.10, it can be seen that the positive selection rate of major SNPs (Setting 1-1 and Setting 2-1) is higher than that of minor SNPs (Setting 1-2 and Setting 2-2). In summary, the numerical study 2 demonstrates that the generalized tournament screening cum EBIC has a good performance in the situation where the structure of interaction effects is complex. In addtion, the generalized tournament approach has

Table 5.10: Simulation results for the second structure

Method	Correct(SD)	Incorrect(SD)	PSR	FDR
Setting $2-1$				
GT1	9.91(0.38)	15.54(3.93)	0.991	0.601
GT2	8.99(0.92)	1.27(0.87)	0.899	0.124
GT3	8.65(0.98)	0.92(0.86)	0.865	0.096
Setting $2-2$				
GT1	9.28(0.94)	18.77(3.76)	0.928	0.669
GT2	5.55(2.11)	0.54(0.72)	0.555	0.089
GT3	4.03(2.19)	0.21(0.48)	0.403	0.050

high power and low false discovery rate in detecting major SNPs. Even in the case of minor SNPs, the generalized tournament approach has reasonable positive selection rate and false discovery rate.

5.5 Summary

We have applied the generalized tournament screening cum EBIC in genome-wide association studies for detecting SNPs associated with some common diseases. Not only main effects but also gene-gene interactions were considered as possible factors in logistic regression model. When one SNP has a significant marginal effect, it is likely that its interactions with other SNPs are highly correlated with the disease. In this situation, the multiple testing may declare the significance for many interactions, which

causes a high false discovery rate. In contrast, the generalized tournament approach selects dummy variables jointly not individually. This, combining with the extended Bayesian information criterion, effectively controls the false discovery rate. In addition, the generalized tournament approach is never affected by the separation phenomenon. However, the log-likelihood ratio test in the multiple testing cannot work normally in case of separation.

The generalized tournament screening cum EBIC approach enjoys high positive selection rate and low false discovery rate in comparison with the multiple testing of all possible pairwise gene-gene interactions. Hence, the generalized tournament approach may be a promising way to detect genetic variants responsible for many common diseases.

Chapter 6

Conclusion and Further Research

In this chapter, we summarized the results of the thesis and discuss some further research directions related to the thesis. The main purpose of this thesis is to develop a high dimensional feature selection method for generalized linear models with main effects and interaction effects and then apply it in genome-wide association studies to detect multiple loci associated with diseases.

6.1 Conclusion

The separation phenomenon in a logistic regression model makes the original SCAD method (Fan and Li, 2001) unable to work normally. The reason is that separation results in at least one infinite estimates when maximizing the SCAD penalized log-likelihood function. In Chapter 2, the modified SCAD method is proposed to solve

the problem raised by the separation phenomenon. Compared to the original SCAD method, the modified SCAD function adds the logarithm of the Jeffreys penalty function (Jeffreys, 1948) in the SCAD penalized log-likelihood function. The simulation results show that the modified SCAD method maintains the selection performance of the original SCAD method in case of no separation. It could be explained by the influence of the Jeffreys penalty function is asymptotically negligible. Moreover, the modified SCAD method always guarantees finite parameter estimates in case of separation unlike the SCAD method. The main reason is that the effect of Jeffreys penalty function is equivalent to split each original observation of the response variable into a response and a non-response. Although the original SCAD method was proposed in seven years ago, it has not provided a solution to the problem raised by separation. Hence, this work develops a necessary and reasonable modification for the original SCAD method since separation is a non-negligible problem for logistic regression model.

In Chapter 3, the extended Bayesian information criteria (EBIC; Chen and Chen, 2007) are discussed in generalized linear regression models with both main effects and two-covariate interactions. When both main effects and interaction effects are considered as possible factors in a generalized linear model, the extended Bayesian information criteria put different emphases on main effects and interactions. In addition, the performance of EBIC in generalized linear models is evaluated in comparison with the ordinary Baysian information criterion (BIC). The results in Chapter 3 and 5 demonstrate that the EBIC method has much lower false discovery rate (FDR) than the BIC

method in generalized linear models when the dimension of model space is high. The intolerantly high FDR of BIC would explained by the unreasonable prior probabilities assigned to candidate models. In contrast, the EBIC method uses a possibly more appropriate prior probability, which would account for the low FDR in EBIC. This work has provided clear evidence that the EBIC method is more appropriate in generalized linear models when the dimension of model space is high. Moreover, this work would make the EBIC method more popular.

The generalized tournament screening cum EBIC is proposed in Chapter 4 to deal with high dimensional feature selection in the context of generalized linear models. The generalized tournament approach is suitable to the generalized linear models with not only main effects but also interaction effects. In addition, this method is computationally feasible however high the dimension of feature space is. It is attributed to the principle of the generalized tournament approach that it can transfer a high dimensional model selection problem to some relatively low dimensional model selection problems. Hence, one key advantage of the generalized tournament method is that the dimension of feature space is no longer considered as a great challenge.

In Chapter 5, the generalized tournament screening cum EBIC is applied in genome-wide association studies to detect SNPs associated with some common diseases. The performances of the multiple testing method (Marchini, 2005) and the generalized tournament approach are compared by some simulation studies. As shown in Chapter 5, the

multiple testing method suffers much higher false discovery rate (FDR) than the generalized tournament method cum EBIC. The possible reason is that the multiple testing method assesses gene-gene interactions individually, which may ignore joint effects among interactions. In addition, one significant SNP may cause that some other non-causative SNPs are wrongly detected. At the same time, although the multiple testing selects too many spurious SNPs, it does not enjoy high positive selection rate (PSR). It would explained by the Bonferroni adjustment, which is very conservative when the number of possible gene-gene interactions is huge. Hence, the generalized tournament method cum EBIC could be more appropriate than the multiple testing method since it enjoys higher PSR and lower FDR. Some studies suggest that interactions among loci contribute broadly to complex diseases. Thus, the generalized tournament method cum EBIC would be a promising way to detect SNPs associated with common diseases.

6.2 Topics for further research

There are several interesting directions for future work in the areas of research presented in this thesis. Some future works related to this thesis are as follows:

1. In Chapter 3 and 5, when we compare the performances of the extended Baysian information criteria and the ordinary Bayesian information criterion, the value of the parameter (γ_1, γ_2) was set to be some specific constants. However, it has been shown that the performance of the extended Bayesian information criteria depends on the value

of parameter (γ_1, γ_2). As the parameter is imposed with an increased value, the false discovery rate decreases, but the positive selection rate also decreases in the meantime. As a result, a larger value of (γ_1, γ_2) would cause the power of detecting the significant variables to be low. Therefore, we should develop a method for choosing an appropriate parameter value in a real dataset.

2. The penalized likelihood methodology was used to select features in the generalized tournament approach. However, many features may be highly correlated and should be put into clusters. Hence, if we combine the generalized tournament approach with the group selection methodology (Yuan and Lin, 2006) instead of the penalized likelihood, the power of identifying the significant variables is expected to be improved.

3. In the generalized tournament approach, we put the same penalty on the main effects and interaction effects in the semi-final stage and final stage. It is likely more appropriate that the main effects of two variables are contained in the model before the interaction between two variables. Hence, it is necessary to consider different penalties for main effects and interaction effects.

References

Abecasis, G. R. (2007). Turning a flood od data into a deluge: in silico genotyping for genome-wide association scans. *Genetic Epidemiology* **31**, 653.

Allen, A. S. and Satten, G. A. (2007). Statistical Methods for haplotype sharing in case-parent trio data. *Human Heredity* **64**, 35-44.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **64**, 125-127.

Albert, A. and Anderson, J. S. (1984). On the existence of maximum likelihood estimates in logistic regression model. *Biometrika* **71**, 1-10.

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

Balding, D.J., Bishop, M. and Cannings, C. (2001). *Handbook of Statistical Genetics*, John Wiley and Sons, New York.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57**, 289-300.

Breiman, L. (1995). Better subset regression using non-negative garrote. *Technometrics* **37**, 373-384.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24**, 2350-2383.

Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society. Series B* **64**, 641-656.

Bruce, S. Weir (1996). *Genetic Data Analysis II* Sinauer Associates, Canada.

Burnham, K. P., Anderson, D. R. and White, G. C. (1994). Evaluation of the Kullback-Leibler discrepancy for model selection in open population capture-recapture models. *Biometrical Journal* **36**, 299-315.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model space. *Biometrika* (to appear).

Chen, Z. and Chen, J. (2007). Tournament screening cum EBIC for feature selection with high dimensional feature spaces. *Annals of Statistics* (submitted).

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377-403.

Culverhouse, R., Suarez, B. K., Lin, J. and Reich, T. A. (2002). A perspective on epistasis: limits of models displaying no main effects. *American Journal of Human Genetics* **70**, 461-471.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*. **32**, 407-451.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol* **23**, 70-86.

Efroymson, M. A. (1960). *Multiple Regression Analysis*, Mathematical methods for digital computers (Ralston, A. and Wilf, H. S., ed.), vol. 1, Wiley: New York, 191-203.

Epstein, M. P., Allen, A. S. and Satten, G. A. (2007). Efficient and flexible testing of untyped variants in case-control studies [abstracts 30]. *Annual Meeting of The American Society of Human Genetics, October 25, 2007*, San Diego (CA), 40. http://www.ashg.org/genetics/ashg06s/index.shtm

Fan, J. (1997). Comments on "Wavelet in Statistics: a review" by A. Antoniadis. *J. Italian Statist. Assoc.*, **6**, 131-138.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society. Series B* (to appear).

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27-38.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* **41**, 190-195.

Hastie, T., Rosset, S. Tibshirani, R. and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* **5**, 1391-1415.

Heinze, G., Ploner, M. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine* **71**, 181-187.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409-2419.

Helgadottir, A., Throleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A. et al. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491-1493.

Hirchhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95-108.

Hoh, J. et al. (2000). Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Annals of Human Genetics* **64**, 413-417.

Hoh, J. and Ott, J. (2003). Mathmatical multi-locus approaches to localizing complex human trait genes. *Nature Review Genetics* **4**, 701-709.

Hurvich, C. M. and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Jeffreys, H. (1946). An invariant form for the proir probability in the estimation problem. *Proceedings of the Royal Society A* **186**, 453-461.

Klein, R. J., Zeiss, C. Chew, E. W., Tsai, J-Y et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* **15**, 958-975.

Lin, S., Chakravati, A. and Cutler, D. J. (2004). Exhaustive allelic transmission disequilibrium tests as a new approach to genomewide association studies. *Nature Genetics* **36** 1181-1188.

Lohmueller, K. E., Pearce, C. L. Pike, M., Lauder, E. S. and Hirchhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* **33**, 177-182.

Lowe, C. E., Cooper, J. D., Chapman, J. M., Barratt, B. J., Twells, R. C., Green, E. A. et al. (2004). Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun* **5**, 301-305.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics* **15**, 661-675.

Mallows, C. L. (1995). More comments on C_P. *Technometrics* **37**, 362-372.

Marchini, J., Donnelly, P. and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413-417.

Marchini, J. Howie, B., Myers, S., Mcvean, G. and Donnely, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906-913.

Moore, J. H. and Ritchie, M. D. (2004). The challenges of whole-genome approaches to common diseases. *JAMA* **291**, 1642-1643.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* **12**, 758-765.

Park, M. Y. and Hastie, T. (2006). Regularization path algorithms for detecting gene interactions. Tech. rep., Department of Statistics, Stanford University.

Park, M. Y. and Hastie, T. (2007). An L_1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B* **69**,659-677.

Rao, C. R. and Wu, Y. (1989). Strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369-374.

Rosset, S. and Zhu, J. (2004). Discussion of "Least Angle REgression" by Efron et al. *Annals of Statistics* **32**, 469-475.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Annals of Statistics* **35**, 1012-1030.

Prichard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**, 1-14.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). Akaike information criterion statistics. KTK Scientific Publishers, Tokyo

Sebat, J. Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H. et al. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-528.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics* **88**, 486-494.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3**, e114.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486-494.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221-264.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Sing, C. F. and Davignon, J. (1985). Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. *American Journal of Human Genetics* **37**, 168-285.

Stone, M. (1974). Cross-validatory choice and assessment statistical predictions. *Journal of the Royal Statistical Society. Series B* **36**, 111-147.

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society. Series B* **41**, 276-178.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *PNAS* **100**, 9440-9445.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *communications in Statistics, Theory and Methods* **A7**, 13-26.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**, 267-288.

Tibshirani, R., Saunders, M. Rosset, S. Zhu, L. and Knight, K. (2005). Sparsity and smoothness via the fused LASSO. *Journal of the Royal Statistical Society. Series B* **67**, 91-108.

Tiwari, H. K. (1997) Deriving components of genetic vaiance for multilocus models. *Genet. Epidemiol.* **14**, 1131-1136.

Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology* Oxford University, Oxford.

Thomas, D. C., Haile, R. W. and Duggan, D. (2005). Recent developments in genomewide association scans: a workshop summary and review. *American Journal of Human Genetics* **77**, 337-345.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G. and Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* **6**, 109-118.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49-67.

Zerba, K. E. and Sing, C. F. (2000). Complex adaptive systems and human health: the influence of common genotypes of the apolipoprotein E (ApoE) gene polymorphism and age on the relational order within a field of lipid metabolism traits. *Human Genetics* **107**, 466-475.

Ziegler, A., König, I. R. and Thompson J. R. (2008). Biostatistical aspects of genome-wide association studies. *Biometrical Journal* **50**, 8-28.

Zou, H. (2006). The adaptive Lasso and its oracle property. *Journal of the American Statistical Association*. **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B* **67**, 301-320.

Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **46**, 505-510.