HIGH DIMENSIONAL FEATURE SELECTION UNDER INTERACTIVE MODELS

HE YAWEI

NATIONAL UNIVERSITY OF SINGAPORE

2013

HIGH DIMENSIONAL FEATURE SELECTION UNDER INTERACTIVE MODELS

HE YAWEI

(B.Sc. Wuhan University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2013

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Professor Chen Zehua, for his invaluable guidance, encouragement, kindness and patience. I really appreciate that he led me into the field of statistical research. And I am grateful for all the efforts and time Prof Chen has spent in helping me overcome my problems in the past four years. I learned a lot from him and I am greatly honoured to be a student of him. Secondly, I would like to express my sincere gratitude to my senior and dear friend Luo Shan for all the help she provided. Thanks also to staff members in department of statistics and applied probability for their continuous supports. Finally, special thanks to my friends and my family for their concerns and encouragements.

CONTENTS

SUMMARY

In contemporary statistics, the need to extract useful information from large data boosts the popularity of high dimensional feature selection. High dimensional feature selection aims at selecting relevant features from the suspected high dimensional feature space by removing redundant features. Among high dimensional feature selection studies, a large number of them have considered the main effect features only, although the interactive effect features are also necessary for the explanation of the response variable. In this thesis, we propose feasible feature selection procedures under the high dimensional feature space by considering both the main effect features and the interactive effect features, in the context of linear models and generalized linear models. An efficient feature selection procedure usually comprises two important steps. The first step is designed to generate a sequence of candidate models and the second step is designed to identify the best model from these candidate models. In order to obtain an elaborate selection procedure under the high dimensional space with interactions, we are committed to improving both two steps.

In chapter 2 of this thesis, we expand current studies of the new model selection criterion EBIC (Chen and Chen, 2008) to interactive cases. The theoretical properties of EBIC for linear interactive models with a diverging number of relevant parameters, as well as for generalized linear interactive models, are investigated. The acceptable conditions under which EBIC is selection consistent are identified and some numerical studies are provided to show sample properties of EBIC. In chapter 3 of our study, we firstly propose a novel feature selection procedure, called sequential L_1 regularization algorithm (SLR), for generalized linear models with only main effects. In this SLR, EBIC is applied as the identification criterion of the optimal model, as well as the stopping rule. Subsequently, SLR is extended to interactive models by handling main effects and interactive effects differently. The theoretical property of SLR is explored and the corresponding conditions required for its selection consistency are identified. In chapter 4 of our thesis, extensive numerical studies are provided to show the effectiveness and the feasibility of SLR.

LIST Of NOTATIONS

n	the sample size or the number of independent observations
$\boldsymbol{y}$	the n-dimensional response variable
X	the design matrix with element x_{ij}
$X(s)$	the sub-matrix composed of the columns of X with indices in subset s
X_j	the j^{th} column vector of X
$\boldsymbol{x}_i$	the i^{th} row vector of X
$\boldsymbol{\beta}(s)$	the sub-vector of the coefficient vector $\boldsymbol{\beta}$ with indices in s
p	the number of the main effect features
$\nu(s)$	the number of components in sub-model s
s_{0n}	the true model

p_{0n}	i.e., $\nu(s_{0n})$, the number of the causal (relevant, true) features
I_n	the identity matrix with order n
$\lambda_{\min}(.)$	the smallest eigenvalue of a square matrix
$\lambda_{\max}(.)$	the largest eigenvalue of a square matrix
$O(.)$	$h(n) = O(f(n))$ indicates there exists positive integer K and some constant $C > 0$ such that $\lvert\frac{h(n)}{f(n)}\rvert < C$ for all $n > K$
$o(.)$	$h(n) = o(f(n))$ indicates $\lvert\frac{h(n)}{f(n)}\rvert \to 0$ when $n \to +\infty$
$\\|\boldsymbol{x}\\|_2$	$\sqrt{\sum_{i=1}^{n} x_i^2}$ for $\boldsymbol{x} = (x_1, x_2, ...x_n)$
$\\|\boldsymbol{x}\\|_1$	$\sum_{i=1}^{n} \lvert x_i\rvert$ for $\boldsymbol{x} = (x_1, x_2, ...x_n)$

List of Tables

CHAPTER 1 Introduction

With the rapid development of electrical industry and information technology, contemporary data from various fields like biotechnology and finance tends to be extremely large. Technologies related to large data are required in order to extract knowledge and insights from large and complex collections of digital data. In statistics, one of the most popular technology to deal with large data is high dimensional feature selection. High dimension means that the number of features p in the feature space is of polynomial order or exponential order of the sample size n, which is also known as small n large p situation. The small n large p situation, which is now commonly used, has experienced great changes if compared with the past, when few fields of statistics explored more than 40 features (Blum and Langley, 1997; Kohavi and John, 1997). Feature selection, referred to as variable selection, is a basic project which aims to select causal or relevant features from suspected space by removing the most irrelevant and redundant features. It is widely applied in many areas, including, for instance, quantitative trait loci (QTL) mapping and genome wide association studies (GWAS), e.g. Storey et.al (2005), Zou and Zeng (2009).

When the number of features p is fixed whereas the number of observations n is sufficiently large, two main objectives of feature selection, selection consistency and prediction accuracy, could be achieved simultaneously and effectively through some traditional criteria like Akaikes information criterion (AIC) (Akaike, 1973), Bayes information criterion (BIC) (Schwartz, 1978), cross-validation (CV) (Stone, 1974) and generalized cross-validation (GCV) (Craven and Wahba, 1979). Furthermore, in this fixed p large n situation, the optimal model is often decided directly from finite candidate models by applying one of these traditional model selection criteria. Actually, feature selection could be regarded as a special case of model selection. They are different in that feature selection concentrates on detecting causal features while model selection concentrates on the accuracy of the model. However, model selection cannot be employed to identify the optimal model directly in high dimensional feature space, probably because there would be nearly

2^p sub-models with a quite large p and this huge number of candidate sub-models make the identification of the best model impracticable in terms of computational cost. Therefore, a popular way for variable selection in large p situation is to obtain a certain number of candidate models first through some feature selection approaches before deciding the final model on the basis of various model selection criteria.

It is noted that, in small n large p situation, it is unlikely to address selection consistency and prediction accuracy at the same time because of the occurrence of over-fitting, thus it is necessary to address these two goals from different aspects. The selection consistency deserves more attention than the prediction accuracy since it is essential to extract effective information considering noise accumulation and model interpretation. For instance, in QTL mapping and disease gene mapping, our primary interest is the markers which are either QTL or disease genes themselves but not others. On the other hand, the occurrence of over-fitting also suggests the requirement for reappraising the feasibility of those traditional criteria under the new situation. In fact, it has been observed by many researchers that all four criteria AIC, BIC, CV and GCV tend to be liberal in selecting a model with many spurious covariants. This implies that they may not be suitable for small n large p situation. As a result, some works have been done on adjusting the priors on the basis of these criteria. Among these works, the most significant is the extended BIC information criterion (EBIC) developed by Chen and Chen (2008).

In high dimensional studies, the *sparsity* assumption, which indicates the true number of relevant or causal features is small, is commonly used. This assumption is reasonable for small n large p problems because it arises from many scientific endeavors. For instance, in disease classification, it is generally agreed that only a small fraction of total genes are responsible for a disease. However, it is a challenging task to select a few causal features that could explain the response variable from a large amount of candidates, with a relatively small sample size. And various difficulties in high dimensional space arise, such as high spurious correlation, mix of causal and non-causal features and complicated computation. Statisticians have made great efforts to develop new techniques to overcome these difficulties. Some of them proposed dimension reduction, a straightforward and effective strategy, to deal with the feature selection problem in high or ultra-high space. Strategies for dimension reduction, such as sure independence screening (SIS), iterative SIS (ISIS) (Fan and Lv, 2008), tournament screening (TS) (Chen and Chen, 2009) and maximum marginal likelihood estimator (MMLE) (Fan and Song, 2010), can ease the computation burden efficiently without losing important information, because they possess sure screening properties which assure the probability that the reduced lower-dimensional model contains the true model converges to 1 under certain conditions. Nevertheless, the reduced lower-dimensional space still requires further selection because it has a much larger dimension than expected.

In general, an efficient procedure for high dimensional feature selection often consists of two stages: a screening stage and a selection stage. The screening stage, that is, the dimension reduction stage, may not be necessary if the number of features p is large but not large enough. However, this stage becomes imperative when interactions of features are considered since the dimension increases significantly. The second stage, i.e. the further selection stage, is the core of feature selection in high dimensional space. This selection stage usually comprises two important steps. The first step aims at generating some candidate models and the second step aims at selecting a final model among the candidate models. The first step can be carried out through a suitable feature selection procedure. Feature selection procedures can be classified into two major categories: sequential procedures including classical methods like stepwise selection, backward elimination; penalized likelihood methods including Lasso (Tibshirani, 1996). Among these categories, the more popular one is penalized likelihood methods. The second step is realized by using an appropriate model selection criterion. Traditionally, the AIC, BIC or CV are used. In the case of high-dimensional data, a more suitable criterion is the EBIC.

In the following sections, a detailed review of literatures related to the selection stage are presented. In section 1.1, literatures about feature selection methods, especially penalized likelihood methods, are reviewed. In section 1.2, various model selection criteria, especially the EBIC, are introduced. In section 1.3, the aim and the organization of this thesis are given.

1.1 Feature Selection Methods

Many researchers have concentrated on developing efficient methods for feature selection recently, especially in small n large p situation. Most of these selection methods were initially proposed through observations in linear models (LMs). Under LMs, the well-known ordinary least squares (OLS) estimates, which are obtained by minimizing residual squared error, suffer from two main drawbacks (Tibshirani, 1996). The first drawback is prediction accuracy since OLS estimates usually have low bias but large variance. The second drawback is interpretation because a large number of OLS estimates are non-zero whereas only a small subset of predictors exhibiting the strongest effects are required. Best subset selection improves OLS by selecting or deleting an independent variable through hypothesis testing, thus it provides interpretable models. Many traditional criteria, such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), follow stepwise subset selection. However, the discrete process of subset selection may result in variability, that is, small changes in data might lead to very different models.

An alterative way to improve OLS is to add the penalty function coupled with the tuning parameter λ to the log-likelihood function, which is referred to as the penalized likelihood method. Penalized likelihood methods perform variable selection and estimate unknown parameters by jointly minimizing empirical errors and penalty functions. In light of penalty functions, penalized methods often shrink estimates to make tradeoff between variance and bias overcoming the drawbacks of OLS estimates and best subset selection. These penalized likelihood methods include, for instance, least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), least angle regression (LARS) (Efron et.al, 2004).

In the following paragraphs, literatures about penalized likelihood methods are reviewed in details. It is generally known that both linear models (LMs) and generalized linear models (GLMs) play an important role in feature selection whereas many penalized methods were initially developed through LMs, a special case of GLMs. Thus, we first introduce penalized likelihood methods in the context of LMs, that is, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ denotes the $n \times 1$ response vector, X is an $n \times r$ matrix and $\boldsymbol{\epsilon}$ represents the $n \times 1$ error term. Penalized likelihood estimates can be summarized in the following form

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{r} p_\lambda(\beta_j)\}.$$

The penalty function p_λ has a direct impact on the performance of various penalized approaches. It is regarded as a good penalty if it results in an estimator with three properties: unbiasedness, sparsity and continuity (Fan and Li, 2001).

Unbiasedness: The resulting estimator is unbiased for large true unknown parameters.

Sparsity: The resulting estimator can automatically set estimated coefficients with small values to zero.

Continuity: The resulting estimator is continuous in data to avoid instability in model prediction.

In 1993, Frank and Friedman proposed bridge regression with the L_q penalty, that is, $p_\lambda(\beta) = \lambda|\beta|^q$. When $q > 1$, penalized estimates shrink the solutions to reduce variability whereas do not enjoy sparsity. In particular, when $q = 2$, the corresponding process, referred to as ridge regression (Draper and Smith, 1998), shrinks coefficients continuously and thus obtains a better prediction result. Nevertheless, ridge regression fails to provide an easy interpretable model since it does not set any coefficients to zero.

When $q \leq 1$, the L_q penalty results in sparse solutions but relatively large biases. Among L_q families, the most famous one is the Lasso (Tibshirani, 1996) with L_1 penalty, which is also referred to as *basis pursuit* in signal processing (Chen, Donoho, and Saunders, 2001). Lasso's estimates approach OLS estimates if the value of λ is small whereas most of them are exactly zero when λ is sufficiently large. This nature of Lasso leads to a continuous shrinking operation and sparse estimates, which makes it catch researchers' attentions increasingly due to the fact that sparse models are more interpretable and preferred in sciences.

It was pointed out by Osborne et.al (2000) that Lasso provided a computationally feasible way for feature selection since its entire regularization path is computed in the complexity of one linear regression. Subsequently, asymptotic behaviors of Lasso estimates, i.e. consistency and limiting distributions, were investigated by Knight and Fu (2000). In order to apply Lasso for feature selection, it is essential to assess how well the sparse model given by Lasso relates to the true model. This assessment is made by some researchers through investigating the model selection consistency of Lasso, and they then proposed some conditions, for instance, Irrepresentable Condition (Zhao and Yu, 2006), Mutual Incohorence Condition (Wainwright, 2009), Neighborhood Stability Condition (Meinshausen and Buhlmann, 2006). These conditions require non-causal features to weakly correlate with the relevant features, which seems too strong to be satisfied.

Lasso can be fitted efficiently by Least Angle Regression (LARS) (Efron et.al, 2004), the version of stagewise via the L_1 penalty. LARS has a similar result with Lasso and it is useful in enhancing the understanding of Lasso. In addition, although Lasso yields almost the same solution path with LARS, it might have a slower speed in tracing the entire solution path. In general, Lasso is a valuable tool for model fitting and feature selection. Nevertheless, it has several fundamental limitations. Firstly, Lasso lacks the oracle property (Fan and Li, 2001): estimates perform as well as if the true model is given in advance, because of its biased estimates for large coefficients. Secondly, Lasso cannot handle the collinearity, which reflects in its poor performance when high correlations exist. Actually, for a group of features among which two-way correlations are high, Lasso tends to select one feature from this group but does not care which one it is (Zou and Hastie, 2005).

Motivated by Lasso, numerous alternatives or extensions arose quickly. Zou and Hastie (2005) proposed a new shrinkage and selection method, referred to as elastic net, by combining Lasso and ridge regression, that is, $p_\lambda(\beta) = \lambda_1|\beta| + \lambda_2|\beta|^2$. The elastic net produces a sparse model with better prediction accuracy than Lasso, especially for microarray data analysis, although it encourages a grouping effect unfortunately. This grouping effect suggests that strongly correlated predictors tend to be in or out of the model together. Zou (2006) advocated a new version of Lasso, adaptive Lasso, by utilizing penalty for penalizing different coefficients, i.e. $p_\lambda(\beta_j) = \lambda w_j|\beta_j|$ for $w_j = 1/|\widehat{\beta}_j|$ with an initial estimator $\widehat{\beta}_j$. If a reasonable initial estimator is available, adaptive Lasso enjoys the oracle property in the sense of Fan and Li (2001) under either fixed p (Zou, 2006) or sparse high feature space (Huang, Ma and Zhang, 2008) whereas Lasso does not. In summary, elastic net and adaptive Lasso improve Lasso in two different ways: elastic net handles collinearity whereas lacks the oracle property; adaptive Lasso owns the oracle property but does not handle collinearity. To improve Lasso in both ways, Zou and Zhang (2009) combined the strength of elastic net and adaptive Lasso and developed a better method called the adaptive elastic-net.

Another significant extension of Lasso, sequential Lasso (SLasso), was proposed by Luo and Chen (2013b) through solving a sequence of partial L_1 penalized problems. By letting the earlier selected features not be penalized in later stages, SLasso ensures $s_k \subset s_{k+1}$, where s_k represents the set of features selected until step k. This differs from Lasso in which a feature included in previous stages may be left out in a later step. Under reasonable assumptions, SLasso enjoys the oracle property in the scenario that the number of features $p = exp(n^k)$ and the number of relevant features p_{0n} diverges. It bears a similarity with OMP (Cai and Wang, 2011) but is advantageous in revealing properties of OMP under much weaker conditions. In addition, SLasso is computationally appealing due to the intrinsic nature of sequential methods and L_1 penalty, which makes it more powerful for high dimensional linear regression than other approaches like the elastic-net.

In comparison with L_q families, SCAD (Fan and Li, 2001) is a successful alternative because of its desirable properties including unbiasedness, sparsity and continuity. The SCAD has a nonconcave penalty, which is given by

$$p'_\lambda(\beta) = \lambda I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{a-1} I(\beta > \lambda) \; for\ some\ a > 2\ and\ \beta > 0.$$

A penalty similar to SCAD is the minimax concave penalty (MCP) (Zhang, 2010), whose derivative is expressed by $p'_\lambda(\beta) = (a\lambda - \beta)_+/a$. SCAD clearly takes off at the origin as the L_1 penalty and then gradually levels off, and MCP translates the flat part of $p'_\lambda(\beta)$ of SCAD to the origin (Fan and Lv, 2010). The SCAD estimator enjoys the asymptotically oracle property when the dimension of covariates is either fixed (Fan and Li, 2001) or diverging slowly (Fan and Peng, 2004) or much larger than the sample size, i.e. small n large p (Kim et.al, 2008). Nevertheless, it is more difficult to compute SCAD estimates than other concave approaches, for example, the L_1 approach, although there has been effort to develop efficient algorithms for these non-convex penalized problems.

Besides LMs, feature selection in other GLMs is also prevalent because of their wide range of applications. However, GLMs are relatively little studied in high feature space in comparison with LMs, probably because GLMs have more complex data structures, complicated solution paths and implicit estimates and thus feature selection in GLMs is more challenging. In fact, GLMs and LMs are only different in that the former accepts different links between $E(\boldsymbol{y})$ and $X\boldsymbol{\beta}$, for example, *identity, log, logit*, whereas the later only allows *identity.* In light of the similarity of LMs and GLMs, it is noteworthy to extend feature selection methods from LMs to GLMs. As mentioned by previous literatures, feature selection methods such as Lasso (Tibshirani, 1996), adaptive Lasso (Zou, 2006) and SLasso (Luo and Chen, 2013b) are efficient and powerful for high dimensional linear regression. Among these methods, some like the adaptive Lasso (Zou, 2006) were extended to GLMs only through a brief discussion while some were systematically investigated. For instance, Lasso was systematically explored under GLMs and Park and Hastie (2007) then developed the path-following algorithm. Nevertheless, SLasso, the significant method which is highly advantageous in the oracle property and the computation complexity, is not included in these extensions.

1.2 Model Selection Criteria

In high dimensional feature space, penalized methods can generate a sequence of candidate models in light of different values of the tuning parameter λ. The identification of the optimal model from these candidate models depends on the appropriate choice of the tuning parameter, a choice which can be made through some suitable model selection criteria. The selection criteria are determined by the aim of a study. For instance, in a GLM, when a study focuses on the prediction performance of candidate models, it would be better to apply deviance or CV. But if this study concentrates on singling out causal features, EBIC (Chen and Chen, 2008) may become a good selection criterion.

Over the past four decades, many traditional model selection criteria, including the C_p criterion (Mallows, 1973), AIC (Akaike, 1973), BIC (Schwarz, 1978), CV (Stone, 1974) and GCV (Craven and Wahba, 1979), have been proposed. The C_p criterion mainly relies on some forms of the mean squared error (MSE) that is frequently used for measuring the performance of a prediction. AIC and BIC have similar forms, which are defined as minus twice log-likelihood for model s combining with a penalized part, although they are developed from different philosophy. The penalized part is given by $2\nu(s)$ in AIC and $\nu(s)\log n$ in BIC, where $\nu(s)$ represents the cardinality of s. In CV, the dataset is divided into training set and testing set alternatively. CV fits a model on the training set but validates the performance of the model on the testing set. GCV is a generalization of CV by averaging diagonal elements of the hat matrix. All these traditional criteria performed well when the total number of features was small.

Recently high dimensional datasets frequently appear and pose great challenges to model selection. In high feature space, AIC and BIC, which focus more on selection consistency, have a strong tendency to overestimate the number of regressors. Furthermore, AIC seems to select the model with more features than BIC because of AIC's relative smaller penalized part. Other classic criteria like CV and GCV, which aim to minimize prediction errors, are also overly liberal by selecting a lot of spurious features. This liberal phenomenon implies all these traditional criteria may not be suitable for high dimensional feature selection and this implication has been observed by many authors, e.g. Siegmund (2004), Bogdan et.al (2004), Chen and Chen (2008).

Many authors attempted to improve traditional model selection criteria in high dimensional space. Some of them concentrated on adjusting priors for BIC, including modified BIC (mBIC) (Bogdan et.al, 2004) and extended BIC (EBIC) (Chen and Chen, 2008). The mBIC supplements the original BIC with an additional term $\nu(s)\log(l-1)$ for the study of QTL mapping with interactions. However, its viability and effectiveness were reflected only through some simulations. In contrast, EBIC, which was firstly developed by Chen and Chen (2008) through examining both the number of unknown parameters and the complexity of the model space, is shown to be selection consistent through strict demonstration under different types of models, e.g. Chen and Chen (2008), Chen and Chen (2012), Luo and Chen (2013a).

The definition and derivation of EBIC could be described in detail below. Assume $\{(y_i, x_{i1}, x_{i2}, ..., x_{ip}) : i = 1, 2, ..., n\}$ are the response variable and predictors while $f(y_i|x_{ij}, \boldsymbol{\beta})$ is the conditional density of y_i. The log likelihood function of y_i

is defined as

$$l_n(\boldsymbol{\beta}) = \log \Pi_{i=1}^{n} f(y_i|x_{ij}, \boldsymbol{\beta}).$$

Let $\boldsymbol{\beta}(s)$ be the sub-vector of the coefficient vector $\boldsymbol{\beta}$ with those components outside s being 0 and $\widehat{\boldsymbol{\beta}}(s)$ be its corresponding maximum likelihood estimator (without penalty). For $s \subset \{1, 2, ..., p\}$, the EBIC selects the optimal model which minimizes $EBIC_\gamma(s)$, where

$$EBIC_\gamma(s) = -2l_n(\widehat{\boldsymbol{\beta}}(s)) + \nu(s)\log n + 2\gamma \log \binom{p}{\nu(s)}.$$

Various prior probabilities on models in different sub-models, which are indexed by a parameter γ in the range greater than zero, are what make the difference between EBIC and BIC. The original BIC is actually a special case of EBIC with $\gamma = 0$. The mBIC could also be considered a special situation of EBIC in an asymptotic sense; that is, it is asymptotically equivalent to EBIC with $\gamma = 1$.

The most important property of EBIC, selection consistency, is defined as

$$P(EBIC_\gamma(s_{0n}) < \min_{s \neq s_{0n}} EBIC_\gamma(s)) \to 1 \; when \; n \to \infty.$$

It indicates that the selected model with the smallest EBIC converges to the true model s_{0n} at the probability 1. Under the constraint $\gamma > 1 - \frac{1}{2k}$, EBIC (Chen and Chen, 2008) was shown to be selection consistent in LMs for $p = O(n^k)$ and a fixed p_{0n}, where p_{0n} denotes the number of true features. This finding also implies that BIC is not selection consistent because its corresponding γ is out of range. Generally, in comparison with BIC, EBIC controls the entry of spurious features efficiently while keeping most of the true features, which may be its biggest improvement. Luo and Chen (2013a) extended the selection consistency of EBIC to the ultra-high feature space which allowed $p = exp(O(n^k))$ but a diverging p_{0n}, for instance, $O(n^c)$ with a small c. This diverging setting for p_{0n} is more promising than a fixed setting for the purpose of reflecting the estimability of feature effects. That's because causal features in high dimensional space are still relatively large and their effects often taper off to zero, although the true model is assumed to be sparse. Besides LMs, EBIC is still selection consistent under the more complicated and helpful GLMs with either canonical link (Chen and Chen, 2012) or non-canonical link (Luo and Chen, 2013c). This significant work has constituted an integral part for EBIC in ultra-high feature space. It is worth noting that EBIC is not restricted to LMs and GLMs. In fact, it also performs well in other types of models like gaussian graphical models (Foygel and Drton, 2010) and Cox Proportional Hazards models (CPH) (Luo and Chen, 2013d).

The vast majority of previous studies for EBIC are limited to main effects. The interactive effects are not considered in these studies although interactions are prominent in explaining the response variable in some practical fields. For example, empirical studies in QTL mapping have shown that interactions among loci might conduce to most common diseases. The lack of interactive cases in high dimensional space may result in an inaccurate choice. In particular, for some significant two-covariate interactions, there may be little main effects at a single covariate, thus we cannot detect them when only main effects are considered. As mentioned by many authors, such as Storey et.al (2005), Zou and Zeng (2009), Zhao and Chen (2012), it is necessary to consider both main effects and interactive effects for high dimensional feature selection. Therefore, in our thesis, for a wider application of the EBIC, we would examine the properties of the EBIC under LMs and GLMs, taking into consideration of interactions.

1.3 Aims and Organizations

For feature selection, both LMs and other GLMs play an important role in high or ultra-high feature space. Among studies in high dimensional space, only a relatively small number have been written on sparse models involving interactive terms or non-linearity. As mentioned in section 1.1, the most popular feature selection method under LMs and GLMs is the penalized likelihood method. Among penalized methods, the more significant one is SLasso (Luo and Chen, 2013b) proposed for LMs with only main effects. Therefore, in our thesis, we first provide its extension, called sequential L_1 regularization algorithm (SLR), to improve the feature selection process for GLMs; and secondly we promote SLR to interactive models.

It was mentioned in section 1.2 that EBIC (Chen and Chen, 2008) is suitable for high dimensional feature selection, because it can efficiently restrict the false discovery rate while maintaining the positive discovery rate whereas classic model selection criteria cannot. Nevertheless, the selection consistency of EBIC has been demonstrated in models with main effect features only and it has not been explored in either LMs or GLMs when interactions are taken into consideration. Denote LMs and GLMs containing both main effects and interactive effects by linear interactive models (LIMs) and generalized linear interactive models (GLIMs) respectively. Under LIMs and GLIMs, the selection consistency of EBIC are also established in our study.

In summary, our main purpose in this thesis was to propose feature selection procedures for high dimensional space with interactions. Only two-way interactions are considered in our interactive models since high order interactive effects are rare and complicated. The results of our study may contribute to a more effective and accurate way of selecting relevant features in QTL mapping and GWAS. At the same time, the correct extraction of useful information in these fields of biology, that is, the selection of relevant features, may offer a clear explanation for some diseases like cancer, thus having a great potential impact upon our everyday life.

The thesis is arranged as follows: In chapter 2, we will concentrate on examining the selection consistency of EBIC in LIMs and GLIMs under a general scenario where the number of relevant features is allowed to vary with sample size. In chapter 3, with the application of EBIC, we will provide an efficient procedure SLR to conduct feature selection in GLMs. SLR will be explored under models with only main effects and interactive models respectively through section 3.1 and section 3.2. In section 3.3, we will establish the selection consistency of SLR. In chapter 4, extensive numerical studies will be provided to verify finite sample properties of SLR. In the final chapter, chapter 5, some overall conclusions will be presented and suggestions for future research will be given.

CHAPTER 2

EBIC Under Interactive Models

EBIC is a new model selection criterion firstly developed by Chen and Chen (2008) for feature selection in high dimensional space. It was motivated from the classic BIC (Schwarz, 1978) by examining the complexity of the model space through a parameter γ in the range $[0, 1]$. Under high or ultra-high space, EBIC had been shown to be selection consistent under either LMs (Luo and Chen, 2013a) or GLMs (Chen and Chen, 2012; Luo and Chen, 2013c). Nevertheless, in all these studies, only the main effect features are taken into account whereas the interactive effect features are not.

In this chapter, properties of EBIC under interactive models are explored. Only two-way interactive effect features are considered in this study and the data is generally assumed to be centered. In section 2.1, we give a brief description for EBIC under models with pairwise interactions. The selection consistency of EBIC under linear interactive models (LIMs) and generalized linear interactive models (GLIMs) is explored and discussed in section 2.2 and section 2.3 respectively.

2.1 Description for EBIC

In model selection, either main effect features or interactive effect features may be related to the response variable $\boldsymbol{y}$. As mentioned in section 1.2, for the study of only main effects, EBIC is equivalent to an additional penalty term $2\gamma \log \binom{p}{\nu(s)}$ in the original BIC. When pairwise interactions are considered, this additional penalty term should be $2\gamma \log \binom{p(p+1)}{\nu(s)}$. Nevertheless, this approach is not credible because the effect of selecting a main effect feature differs from that of selecting an interactive effect feature. For example, a pairwise interaction involves two covariates whereas a main effect feature only includes one corresponding covariate. Thus, under either LIMs or GLIMs, EBIC should be modified by penalizing model s with two parts of penalized functions in order to emphasize different roles of main effect features and interactive effect features. We prepare one penalty part $2\gamma_m \log \binom{p}{\nu_m(s)}$ for main effects and the other part $2\gamma_I \log \binom{p(p-1)}{\nu_I(s)}$ for interactive effects, where $\nu_m(s)$ and $\nu_I(s)$ represent the number of main effect features and the number of interactive effect features in the model s. As a result, EBIC under models with interactions can then be expressed by

$$EBIC_\gamma(s) = -2l_n(\widehat{\boldsymbol{\beta}}(s)) + \nu(s)\log n + 2\gamma_m \log\binom{p}{\nu_m(s)} + 2\gamma_I \log\binom{p(p-1)/2}{\nu_I(s)}. \tag{2.1}$$

2.2 Selection Consistency Under Linear Interactive Model

Let $\{(y_i, x_{i1}, ..., x_{ip}) : i = 1, 2, ..., n\}$ be independent observations. We consider the following linear interactive model (LIM)

$$y_i = \boldsymbol{x}_i^\tau \boldsymbol{\beta} + \epsilon_i = \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_h x_{ij} x_{ik} + \epsilon_i,\ i = 1, 2, ..., n. \tag{2.2}$$

This model is equivalent to $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ if it is expressed in matrix notation, where $\boldsymbol{y} = (y_1, y_2, ..., y_n)^\tau$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p(p+1)/2})^\tau$, $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\tau$, $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)^\tau$. The first p components of $\boldsymbol{x}_i$ are $x_{ij} = x_{ij}$ while other $p(p-1)/2$ components x_{ih} satisfy $x_{ih} = x_{ij}x_{ik}$, where $h = (2p - j + 1)j/2 + k - j$ for $1 \le j < k \le p$. There are two assumptions for this LIM. Firstly, the error term $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\mathbf{I}_n$ represents the identity matrix. Secondly, the model is *sparse*, which suggests most components of $\boldsymbol{\beta}$ should be 0.

Some notations required are introduced here first. We use $s_{0n} = \{j : \beta_j \neq 0, j \in \{1, ..., p(p+1)/2\}\}$ to denote the true model. Refer to s as a submodel and let $\nu(s)$ be the number of components in s. Let $p_{0n} = \nu(s_{0n})$ and thus it represents the number of relevant (or causal, true) features. In addition, we assume $X(s)$ is the matrix composed of the columns of X with indices in s and $X^\tau(s)$ is the transpose of $X(s)$. Define $\Delta_n(s) = \|\boldsymbol{\mu} - H_n(s)\boldsymbol{\mu}\|_2^2$, where $\boldsymbol{\mu} = X(s_{0n})\boldsymbol{\beta}(s_{0n})$ and $H_n(s) = X(s)(X^\tau(s)X(s))^{-1}X^\tau(s)$. Then we state the main result on the selection consistency of EBIC under high dimensional space with a diverging p_{0n}.

Theorem 2.1. *Assume model (2.2) and* $\min(\frac{\triangle_n(s)}{p_{0n}\ln p} : s_{0n} \not\subseteq s, \nu(s) \leq k_n) \to \infty$ *for* $k_n = rp_{0n}$ *with any fixed* $r > 1$. *Besides, assume that* $p_{0n}\ln p = o(n)$, $\ln p_{0n}/\ln p \to 0$. *Then when* n *goes to* $+\infty$,

$$P(\min_{s:\nu(s)\leq k_n} EBIC_\gamma(s) > EBIC_\gamma(s_{0n})) \to 1 \tag{2.3}$$

if $\gamma_m > 1 - \frac{\ln n}{2\ln p}$, $\gamma_I > 1 - \frac{\ln n}{4\ln p}$.

Theorem 2.1 indicates that the selected model with the smallest EBIC among models having a cardinality less than rp_{0n} $(r > 1)$, with a probability converging to 1, will be the true model. The restriction for the cardinality of selected models is reasonable since only models with the size comparable with the true model will be considered in practice. This consistency theorem allows $p = O(n^k)$ $(k > 0)$ or $\ln p = O(n^k)$ $(0 < k < 1)$ and a diverging p_{0n} satisfying $\ln p_{0n} = o(\ln p)$. Certainly, it is still valid for a fixed p_{0n} under either high or ultra-high feature space.

The assumption $\lim_{n\to\infty}\min(\frac{\triangle_n(s)}{p_{0n}\ln p} : s_{0n} \not\subseteq s, \nu(s) \leq k_n) = \infty$ is called *consistency condition* in Luo and Chen (2013a), which is shown to be weaker and greater than *asymptotic identifiability condition* (Chen and Chen, 2008). This assumption implicitly requires

$$\sqrt{\frac{n}{p_{0n}\ln p}}\min\{|\beta_j| : j \in s_{0n}\} \to \infty, \tag{2.4}$$

and thus it determines a constraint on the pattern $(n, p_{0n}, p, \boldsymbol{\beta})$. For example, if $p = O(exp(n^k))$ and $p_{0n} = O(n^c)$, (2.4) reduces to $n^{(1-c-k)/2}\min\{|\beta_j| : j \in s_{0n}\} \to \infty$, which implies $\min\{|\beta_j| : j \in s_{0n}\}$ should have a magnitude larger than $O(n^{(c+k-1)/2})$. In this way, we obtain a consistency pattern $(n, p_{0n}, p) = (n, O(n^c), O(exp(n^k)))$, $\min\{|\beta_j| : j \in s_{0n}\} = O(n^{(b-1)/2})$, $0 < c, k < 1$, $k + c < b < 1$. Similarly, when $p = O(n^k)$ and $p_{0n} = O(\ln n)$, the following pattern is still consistent: $(n, p_{0n}, p) = (n, O(\ln n), O(n^k))$, $\min\{|\beta_j| : j \in s_{0n}\} = O(n^{(b-1)/2})$, $k > 0$, $0 < b < 1$.

Proof of Theorem 2.1: Let S_j^m be the class of submodels including j main effects but no interactive effects; Let S_j^I be the class of submodels containing j interactions but no main effects. Thus, the size of S_j^m, $\tau(S_j^m)$, should be C_p^j; the size of S_j^I, $\tau(S_j^I)$, should be $C_{p(p-1)/2}^j$. Under LIMs, for any s, $EBIC_\gamma(s) - EBIC_\gamma(s_{0n})$ can be decomposed into $T_1 + T_2$, where

$$\begin{aligned} T_1 &= n\ln\frac{\boldsymbol{y}^\tau\{I_n - H_n(s)\}\boldsymbol{y}}{\boldsymbol{y}^\tau\{I_n - H_n(s_{0n})\}\boldsymbol{y}} = n\ln\frac{\boldsymbol{y}^\tau\{I_n - H_n(s)\}\boldsymbol{y}}{\boldsymbol{\epsilon}^\tau\{I_n - H_n(s_{0n})\}\boldsymbol{\epsilon}} \\ &= n\ln\{1 + \frac{\boldsymbol{y}^\tau\{I_n - H_n(s)\}\boldsymbol{y} - \boldsymbol{\epsilon}^\tau\{I_n - H_n(s_{0n})\}\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}^\tau\{I_n - H_n(s_{0n})\}\boldsymbol{\epsilon}}\} \end{aligned}$$

and

$$T_2 = (\nu(s)-\nu(s_{0n}))\ln n+2\gamma_m(\ln\tau(S^m_{\nu_m(s)})-\ln\tau(S^m_{\nu_m(s_{0n})}))+2\gamma_I(\ln\tau(S^I_{\nu_I(s)})-\ln\tau(S^I_{\nu_I(s_{0n})})).$$

Based on T_1 and T_2, the selection consistency is then explored under two cases: $s_{0n} \nsubseteq s$ and $s_{0n} \subset s$, in which two lemmas given by Luo and Chen (2013a) are required, that is,

$$P(\chi^2_j \geq m) = \frac{1}{\Gamma(j/2)}(m/2)^{j/2-1}e^{-m/2}(1+o(1)) \ if \ m\to\infty \ and \ j/m\to 0 \quad (2.5)$$

and

$$\ln(\frac{p!}{j!(p-j)!}) = j\ln p(1-\delta)(1+o(1)) \ if \ p\to\infty \ and \ \frac{\ln j}{\ln p}\to\delta. \quad (2.6)$$

Case 1: $s_{0n} \nsubseteq s$

Without loss of generality, we assume $\sigma^2 = 1$ and can write

$$\boldsymbol{\epsilon}^T\{I_n - H_n(s_{0n})\}\boldsymbol{\epsilon} = \sum_{i=1}^{n-p_{0n}} Z_i^2 = (n-p_{0n})(1+o_p(1)) = n(1+o_p(1)),$$

where Z_i are i.i.d. stand normal variable. We then have

$$\begin{aligned} & \boldsymbol{y}^\tau[I_n - H_n(s)]\boldsymbol{y} - \boldsymbol{\epsilon}^\tau[I_n - H_n(s_{0n})]\boldsymbol{\epsilon} \\ = \quad & \Delta_n(s) + 2\boldsymbol{\mu}^\tau[I_n - H_n(s)]\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\tau H_n(s_{0n})\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\tau H_n(s)\boldsymbol{\epsilon}. \end{aligned}$$

For this equation, the following statements will be established uniformly for all s with $\nu(s) \leq k_n$, that is:

$$\boldsymbol{\epsilon}^\tau H_n(s_{0n})\boldsymbol{\epsilon} = p_{0n}(1+o_p(1)); \quad (2.7)$$

$$\max\{\boldsymbol{\epsilon}^\tau H_n(s)\boldsymbol{\epsilon}, \nu(s) \le k_n\} = O_p(k_n \ln p); \tag{2.8}$$

$$|\boldsymbol{\mu}^\tau[I_n - H_n(s)]\boldsymbol{\epsilon}| = \sqrt{\Delta_n(s)O_p(k_n \ln p)}. \tag{2.9}$$

Under our assumptions in the theorem, (2.7)-(2.9) then imply that

$$\boldsymbol{y}^\tau[I_n - H_n(s)]\boldsymbol{y} - \boldsymbol{\epsilon}^\tau[I_n - H_n(s_{0n})]\boldsymbol{\epsilon} = \Delta_n(s)(1 + o_p(1)).$$

Thus

$$T_1 = n\ln(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1)))$$

uniformly for all s with $\nu(s) \le k_n$.

It is trivial that (2.7) is satisfied. We then prove (2.8) and (2.9). Let $m = 2k_n[\ln a_n + \ln(k_n \ln a_n)]$, where $a_n = p(p+1)/2$. Obviously, $\frac{k_n}{m} \to 0$. Let S_j be the class of submodels consisting of j features. Note that $\boldsymbol{\epsilon}^\tau H_n(s)\boldsymbol{\epsilon} = \chi_j^2(s)$ for $j = \nu(s)$. By the Bonferroni inequality, we get

$$\begin{aligned} & P(\max\{\boldsymbol{\epsilon}^\tau H_n(s)\boldsymbol{\epsilon} : \nu(s) \le k_n\} \ge m) \\ = \;& P(\max\{\boldsymbol{\epsilon}^\tau H_n(s)\boldsymbol{\epsilon} : s \in S_j, j \le k_n\} \ge m) \le \sum_{j=1}^{k_n} \tau(S_j)P(\chi_j^2 \ge m). \end{aligned}$$

The fact $\tau(S_j) = C_{a_n}^j \le a_n^j$, combined with the equation (2.5), suggests that there is some c closing to 1 but not depending on j, such that

$$\begin{aligned} \tau(S_j)P(\chi_j^2 \ge m) &\approx c\frac{1}{2^{j/2-1}\Gamma(j/2)}\frac{\tau(S_j)}{a_n^{k_n}}(k_n \ln a_n)^{-k_n} m^{j/2-1} \\ &\le \frac{c}{m}(k_n \ln a_n)^{-j} m^{j/2} = \frac{c}{m}[\sqrt{\frac{m}{(k_n \ln a_n)^2}}]^j = \frac{c}{m}q_n^j, \end{aligned}$$

where

$$q_n = \sqrt{\frac{m}{(k_n \ln a_n)^2}} = \sqrt{\frac{2k_n[\ln a_n + \ln(k_n \ln a_n)]}{(k_n \ln a_n)^2}}(1 + o(1)) \leq q$$

for some $0 < q < 1$ when n is sufficiently large. Therefore

$$P(\max\{\boldsymbol{\epsilon}^\tau H_n(s)\boldsymbol{\epsilon} : s \in S_j, j \leq k_n\} \geq m) \leq \frac{c}{m}\sum_{j=1}^{k_n} q_n^j \leq \frac{c}{m}\frac{q}{1-q} \to 0.$$

Thus

$$\max\{\boldsymbol{\epsilon}^\tau H_n(s)\boldsymbol{\epsilon}, \nu(s) \leq k_n\} = m(1 + o_p(1)) = O_p(k_n \ln p),$$

which establishes (2.8).

To verify (2.9), note that

$$\boldsymbol{\mu}^\tau[I_n - H_n(s)]\boldsymbol{\epsilon} = \sqrt{\Delta_n(s)}Z(s)$$

for $Z(s) \sim N(0, 1)$. Then we have

$$|\boldsymbol{\mu}^\tau[I_n - H_n(s)]\boldsymbol{\epsilon}| = \sqrt{\Delta_n(s)} \max\{|Z(s)| : \nu(s) \leq k_n\}.$$

For the same m, we have

$$\begin{aligned}
P(\max\{|Z(s)| : \nu(s) \leq k_n\} \geq \sqrt{m}) &= P(\max\{|Z(s)| : s \in S_j, j \leq k_n\} \geq \sqrt{m}) \\
&\leq \sum_{j=1}^{k_n} \tau(S_j)P(|Z(s)| \geq \sqrt{m}) \\
&= \sum_{j=1}^{k_n} \tau(S_j)P(\chi_1^2 \geq m) \\
&\leq \sum_{j=1}^{k_n} \tau(S_j)P(\chi_j^2 \geq m)
\end{aligned}$$

because $P(\chi_1^2 \geq m) < P(\chi_j^2 \geq m)$ by (2.5). Similarly, the last sum converges to zero. This establishes (2.9).

Let $\gamma = \max(\gamma_m, \gamma_I)$, then we turn to $EBIC_\gamma(s) - EBIC_\gamma(s_{0n})$. When n is sufficiently large, if $\frac{\Delta_n(s)}{n} \to 0$, $T_1 = n \ln(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1)))$ is nearly $\Delta_n(s)(1 + o_p(1))$. Thus

$$\begin{aligned} & EBIC_\gamma(s) - EBIC_\gamma(s_{0n}) \\ \geq\ & \Delta_n(s)(1 + o_p(1)) - p_{0n} \ln n - 2\gamma p_{0n} \ln a_n \geq \frac{\Delta_n(s)}{p_{0n} \ln p}(p_{0n} \ln p - \frac{\ln n}{\ln p} - 4\gamma) \to \infty \end{aligned}$$

uniformly for all s with $\nu(s) \leq k_n$ and any bounded γ. If $\frac{\Delta_n(s)}{n} > 0$, then for some positive c, we have

$$\begin{aligned} & EBIC_\gamma(s) - EBIC_\gamma(s_{0n}) \\ \geq\ & n \ln(1 + c) - p_{0n} \ln n - 2\gamma p_{0n} \ln a_n \geq n \ln(1 + c) - p_{0n} \ln n - 4\gamma p_{0n} \ln p \to \infty \end{aligned}$$

uniformly for all s with $\nu(s) \leq k_n$ and any bounded γ.

Case 2: $s_{0n} \subset s$

When $s_{0n} \subset s$, $\{I_n - H_n(s)\}X(s_{0n}) = 0$. As a result,

$$\boldsymbol{y}^\tau [I_n - H_n(s)]\boldsymbol{y} = \boldsymbol{\epsilon}^\tau [I_n - H_n(s)]\boldsymbol{\epsilon}$$

and

$$\boldsymbol{\epsilon}^\tau [I_n - H_n(s_{0n})]\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\tau [I_n - H_n(s)]\boldsymbol{\epsilon} = \boldsymbol{\epsilon}^\tau [H_n(s) - H_n(s_{0n})]\boldsymbol{\epsilon} = \chi_j^2(s),$$

where $j = \nu(s) - \nu(s_{0n})$. Denote $s = (s^m, s^I)$, where s^m represents the submodel with only $\nu_m(s)$ main effects while s^I denotes the submodel with only $\nu_I(s)$ interactive effects. Then $j = j_m + j_I$ for $j_m = \nu_m(s) - \nu_m(s_{0n})$ and $j_I = \nu_I(s) - \nu_I(s_{0n})$. We have

$$\begin{aligned}
& n \log \frac{\boldsymbol{\epsilon}^\tau [I_n - H_n(s_{0n})]\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}^\tau [I_n - H_n(s)]\boldsymbol{\epsilon}} \\
= \; & n \log(1 + \frac{\chi_j^2(s)}{\boldsymbol{\epsilon}^\tau [I_n - H_n(s_{0n})]\boldsymbol{\epsilon} - \chi_j^2(s)}) \\
\leq \; & \frac{n\chi_j^2(s)}{\boldsymbol{\epsilon}^\tau [I_n - H_n(s_{0n})]\boldsymbol{\epsilon} - \chi_j^2(s)}.
\end{aligned}$$

When $n \to \infty$, $n^{-1}\boldsymbol{\epsilon}^\tau [I_n - H_n(s_{0n})]\boldsymbol{\epsilon} \to \sigma^2 = 1$, that is, $\boldsymbol{\epsilon}^\tau [I_n - H_n(s_{0n})]\boldsymbol{\epsilon} = n(1 + o(1))$.

Let $\widetilde{S_{j_m, j_I}} = \{(s^m, s^I) : (s^m, s^I) \subset (S^m_{j_m + \nu_m(s_{0n})}, S^I_{j_I + \nu_I(s_{0n})}); s_{0n} \subset s\}$.

Note that $\tau(\widetilde{S_{j_m, j_I}}) = C^{j_m}_{p - \nu_m(s_{0n})} C^{j_I}_{b_n - \nu_I(s_{0n})} \leq p^{j_m} b_n^{j_I}$, where $b_n = p(p-1)/2$.

Let $m_{j_m, j_I} = 2j_m(\ln p + \ln(j_m \ln p)) + 2j_I(\ln b_n + \ln(j_I \ln b_n))$.

Then we have

$$\begin{aligned}
& P(\max_{j_m, j_I} \frac{\max\{\chi^2_{j_m + j_I}(s) : s \subset \widetilde{S_{j_m, j_I}}\}}{m_{j_m, j_I}} \geq 1) \\
\leq \; & \sum_{j_m, j_I} P(\max\{\chi^2_{j_m + j_I}(s) : s \subset \widetilde{S_{j_m, j_I}}\} \geq m_{j_m, j_I}) \\
\leq \; & \sum_{j_m, j_I} \tau(\widetilde{S_{j_m, j_I}}) P(\chi^2_{j_m + j_I}(s) \geq m_{j_m, j_I}).
\end{aligned}$$

Follow a similar way with case (1), we can get

$$\sum_{j_m, j_I} \tau(\widetilde{S_{j_m, j_I}}) P(\chi^2_{j_m + j_I}(s) \geq m_{j_m, j_I}) \leq \frac{c}{m_{j_m, j_I}} (\sqrt{\frac{m_{j_m, j_I}}{j_m \ln p}})^2 (\sqrt{\frac{m_{j_m, j_I}}{j_I \ln b_n}})^2,$$

where

$$q_{j_m} = \sqrt{\frac{m_{j_m,j_I}}{j_m \ln p}} \le \sqrt{\frac{c}{\ln p}(1+o(1))} \to 0$$

and

$$q_{j_I} = \sqrt{\frac{m_{j_m,j_I}}{j_I \ln b_n}} \le \sqrt{\frac{c}{\ln p}(1+o(1))} \to 0$$

for some finite c.

Thus,

$$\max\{\chi^2_{j_m+j_I}(s) : s \subset \widetilde{S_{j_m,j_I}}, s_{0n} \subset s\} = m_{j_m,j_I}(1+o_p(1)).$$

Noting that $\ln b_n = 2\ln p(1+o(1))$, we have

$$\begin{aligned} m_{j_m,j_I} &\le 2j_m(\ln p + \ln((k_n - p_{0n})\ln p)) + 2j_I(\ln b_n + \ln((k_n - p_{0n})\ln b_n)) \\ &\le (2j_m + 4j_I)\ln p(1+o_p(1)) \end{aligned}$$

since $\frac{\ln((k_n-p_{0n})\ln p)}{\ln p} \to 0$. In addition,

$$\begin{aligned} &\frac{n\chi^2_j(s)}{\boldsymbol{\epsilon}^\tau[I_n - H_n(s_{0n})]\boldsymbol{\epsilon} - \chi^2_j(s)} \\ \le\ & \frac{nm_{j_m,j_I}}{n - m_{j_m,j_I}(1+o_p(1))} \le m_{j_m,j_I}(1+o_p(1)) \\ \le\ & [2j_m \ln p + 4j_I \ln p](1+o_p(1)). \end{aligned}$$

Thus

$$T_1 \ge -[2j_m \ln p + 4j_I \ln p](1+o_p(1))$$

and

$$T_2 = j\ln n + [2\gamma_m j_m \ln p + 4\gamma_I j_I \ln p](1+o(1)).$$

Finally, we have

$$\begin{aligned}
& EBIC_{\gamma}(s) - EBIC_{\gamma}(s_{0n}) \\
\geq\ & (j_m + j_I)\ln n + [2\gamma_m j_m \ln p + 4\gamma_I j_I \ln p](1 + o(1)) \\
& -[2j_m \ln p + 4j_I \ln p](1 + o_p(1)) \\
=\ & j_m \ln n + 2\gamma_m j_m \ln p(1 + o(1)) - 2j_m \ln p(1 + o_p(1)) \\
& +j_I \ln n + 4\gamma_I j_I \ln p(1 + o(1)) - 4j_I \ln p(1 + o_p(1)).
\end{aligned}$$

When $\gamma_m > 1 - \frac{\ln n}{2\ln p}$, $\gamma_I > 1 - \frac{\ln n}{4\ln p}$, it can be deduced that $EBIC_{\gamma}(s) - EBIC_{\gamma}(s_{0n}) > 0$ uniformly for all s with $\nu(s) \leq k_n$ and $s_{0n} \subset s$, if n is sufficiently large.

2.3 Selection Consistency Under Generalized Linear Interactive Model

In the generalized linear interactive model (GLIM), $\boldsymbol{y} = (y_1, ..., y_n)^{\tau}$ follow a particular exponential distribution with the density function

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} exp\{y_i\theta_i - b(\theta_i)\}. \tag{2.10}$$

The parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^\tau$ is referred to as the natural parameter and its corresponding space Θ is convex. Based on the properties of the exponential family,

$$b'(\theta_i) = E(y_i) = \mu_i, b''(\theta_i) = Var(y_i) = \sigma_i^2. \tag{2.11}$$

The mean $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^\tau$ is related to the design matrix $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^\tau$ through the linear predictor $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\tau$ and a one-to-one continuous differentiable transformation g, that is,

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\tau \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} x_{ij}x_{ik}\beta_h, i = 1, 2, ..., n, \tag{2.12}$$

where the last $p(p-1)/2$ components of $\boldsymbol{x}_i$ satisfy $x_{ih} = x_{ij}x_{ik}$ $(1 \le j < k \le p)$ if $h > p$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p(p+1)/2})^\tau$. This GLIM has a canonical link if $\boldsymbol{\eta} = \boldsymbol{\theta}$. An advantage of the canonical link is the existence of a minimal sufficient statistic for $\boldsymbol{\beta}$, that is, all information about $\boldsymbol{\beta}$ is contained in a function of the data with the same dimension as $\boldsymbol{\beta}$. The commonly used distributions for $\boldsymbol{y}$, like normal, poisson, bernoulli, all satisfy the canonical link. Thus, we only consider the GLIM with the canonical link in our study.

Assume $\boldsymbol{\beta}(s)$ is the sub-vector composed of the elements of $\boldsymbol{\beta}$ with indices in subset s and $\nu(s)$ is the number of features in s. We denote the set of true features by s_{0n} and denote the true coefficient vector by $\boldsymbol{\beta}_0$. By the assumption of sparsity, most components of $\boldsymbol{\beta}_0$ are zero except for those in s_{0n}, which implies $p_{0n} = \nu(s_{0n})$ is relatively small in comparison with p. The log likelihood function of $\boldsymbol{y}$ is given by $l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i|\theta_i) = \sum_{i=1}^{n}[y_i \boldsymbol{x}_i^{\tau}\boldsymbol{\beta} - b(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta})]$. Besides, let $s_n(\boldsymbol{\beta}) = \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ and $H_n(\boldsymbol{\beta}) = \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\tau}}$.

For $C = rp_{0n}$ $(r > 1)$, define $A_0 = \{s : s_{0n} \subset s; \nu(s) \leq C\}$; $A_1 = \{s : s_{0n} \not\subseteq s; \nu(s) \leq C\}$. The following conditions are imposed for EBIC under GLIMs.

C1: $p = O(exp(n^k))$, $0 < k < 1/3$; $p_{0n} = o(n^k)$ uniformly for all $k > 0$.

C2: Suppose $B(s) = \{\boldsymbol{\beta} : \boldsymbol{x}_i^{\tau}(s)\boldsymbol{\beta}(s) \in \Theta, i = 1, ..., n\}$, then the interior of $B(s)$ is not empty, and $\boldsymbol{\beta}_0(s) \in B(s)$ for $s \in A_0 \cup A_1$.

C3: $\inf \min\{|\beta_{0j}| : j \in s_{0n}\} > n^{-1/4}$.

C4: For all j, there exists a positive constant K such that $|x_{ij}| \leq K$ and

$$\max_{1 \leq i \leq n} \frac{x_{ij}^2}{\sum_{i=1}^{n} x_{ij}^2 \sigma_i^2} \leq \frac{Kn^{-1/6}}{\log n}$$

when n is sufficiently large.

C5: When n is sufficiently large, for $s \in A_1$, there exists positive constants k_1 and k_2, such that

$$k_1 \leq \lambda_{\min}(n^{-1}H_n(\boldsymbol{\beta}_0(s \cup s_{0n})) \leq \lambda_{\max}(n^{-1}H_n(\boldsymbol{\beta}_0(s \cup s_{0n})) \leq k_2.$$

C6: There exists a constant $\delta > 0$, for all $s \in A_1$, such that for any $\epsilon > 0$,

$$|\frac{H_n(\boldsymbol{\beta}(s \cup s_{0n}))}{H_n(\boldsymbol{\beta}_0(s \cup s_{0n}))} - 1| \leq \epsilon$$

when $\|\boldsymbol{\beta}(s \cup s_{0n}) - \boldsymbol{\beta}_0(s \cup s_{0n})\|_2 \leq \delta$.

These conditions are almost similar to those in Chen and Chen (2012). Chen and Chen (2012) investigated properties of EBIC under GLMs with only main effects and our study can be regarded as the extension or improvement of their work. C1 points out the application range of EBIC in the GLIM, i.e. ultra-high feature space and C3 determines a constraint on the coefficients. C4 is a weak condition since it won't be violated if the square of a feature is not severely skewed. C6 extends C5 to a small neighborhood while both of them are only provided for $s \in A_1$.

Lemma 2.1: *Under conditions C1-C6, for all $s \in A_0$,*

$$\|\widehat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2^2 = O_p(n^{-1/3}) \tag{2.13}$$

uniformly when $n \to +\infty$.

Theorem 2.2. *Under conditions C1-C6, when $n \to \infty$,*

$$P\{\min_{s \in A_1} EBIC_\gamma(s) \leq EBIC_\gamma(s_{0n})\} \to 0 \tag{2.14}$$

for any $\gamma_m > 0$, $\gamma_I > 0$;

$$P\{\min_{s \in A_0, s \neq s_{0n}} EBIC_\gamma(s) \leq EBIC_\gamma(s_{0n})\} \to 0 \tag{2.15}$$

for any $\gamma_m > 1 - \frac{\log n}{2 \log p}$; $\gamma_I > 1 - \frac{\log n}{4 \log p}$.

Lemma 2.1 gives the convergence rate of the L_2-consistency of the MLE $\widehat{\boldsymbol{\beta}}(s)$ when $s_{0n} \subset s$. Theorem 2.2 rigorously establishes the selection consistency of EBIC under models with a fixed p_{0n}. The EBIC remains selection consistent if p_{0n} diverges slowly with n at a low rate like $p_{0n} = O(\log n)$.

Proof for Lemma 1: This proof is quite similar to the theorem 1 of Chen and Chen (2012), except for the changes for the class of models in A_0. Firstly, we review an inequality (Chen and Chen, 2012) required, that is,

$$P(\sum_{i=1}^{n} a_{ni}(y_i - \mu_i) > \sqrt{2m}) \leq exp(-m(1-\epsilon)) \tag{2.16}$$

for any $m = O(n^{1/3})$ and a_{ni} satisfies $\sum_{i=1}^{n} a_{ni}^2 \sigma_i^2 = 1$ and $\max_i |a_{ni}| = o(n^{-1/6})$. Then we state this proof in details.

Let $\boldsymbol{\beta}(s) = \boldsymbol{\beta}_0(s) + n^{-1/3}\boldsymbol{r}$, where $\boldsymbol{r}$ is a unit vector. It is clear that $\boldsymbol{\beta}(s)$ falls into the neighborhood of $\boldsymbol{\beta}_0(s)$ when n is sufficiently large and thus C5 and C6 are applicable. For all $s \in A_0$,

$$\begin{aligned} l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) &= n^{-1/3}\boldsymbol{r}^\tau s_n(\boldsymbol{\beta}_0(s)) - 1/2n^{1/3}\boldsymbol{r}^\tau\{n^{-1}H_n(\widetilde{\boldsymbol{\beta}}(s))\}\boldsymbol{r} \\ &\leq n^{-1/3}\boldsymbol{r}^\tau s_n(\boldsymbol{\beta}_0(s)) - k_1(1-\epsilon)n^{1/3}. \end{aligned}$$

As a result,

$$\begin{aligned} & P(l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) > 0 : for\ some\ \boldsymbol{r}) \\ \leq\ & P(\boldsymbol{r}^\tau s_n(\boldsymbol{\beta}_0(s)) \geq cn^{2/3} : for\ some\ \boldsymbol{r}) \end{aligned}$$

$$\leq \sum_{j\in s} P(s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}) + \sum_{j\in s} P(-s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}).$$

Let $a_{ni}^2 = x_{ij}^2 / \sum_{i=1}^n x_{ij}^2 \sigma_i^2$ and we get $\max |a_{ni}| = o(n^{-1/6})$ by C4. Note that $s_{nj}(\boldsymbol{\beta}_0(s)) = \sum_{i=1}^n (y_i - \mu_i) x_{ij}$ and $\sum_{i=1}^n x_{ij}^2 \sigma_i^2 = O(n)$, thus by (2.16),

$$P(s_n(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}) \leq P(\sum_{i=1}^n a_{ni}(y_i - \mu_i) \geq \sqrt{2cn^{1/3}}) \leq exp(-cn^{-1/3}).$$

The total number of models in A_0 is less than

$$C^1_{p(p+1)/2} + ... + C^C_{p(p+1)/2} \leq (\frac{p(p+1)}{2})^C \leq p^{2C} = exp\{2Cn^k\} = exp\{o(n^{1/3})\}$$

because $p_{0n} n^k = o(n^{1/3})$ for all $0 < k < 1/3$. Thus

$$\sum_{s\in A_0} \sum_{j\in s} P(s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}) = o(1).$$

In a similar way, we can have

$$\sum_{s\in A_0} \sum_{j\in s} P(-s_{nj}(\boldsymbol{\beta}_0(s)) \geq cn^{2/3}) = o(1).$$

It should be noted that $l_n(\boldsymbol{\beta}(s))$ is a concave function, which suggests $\widehat{\boldsymbol{\beta}}(s)$ exists and falls into the $n^{-1/3}$-neighborhood of $\boldsymbol{\beta}_0(s)$ uniformly for all $s \in A_0$ with a probability tending to 1. This lemma is then proved.

Proof for Theorem 2.2:

Case 1: proof for (2.14)

For any s in A_1, $EBIC_\gamma(s) \leq EBIC_\gamma(s_{0n})$ if and only if

$$l_n(\widehat{\boldsymbol{\beta}}(s)) - l_n(\widehat{\boldsymbol{\beta}}(s_{0n}))$$

$$\begin{aligned}
&\geq \; 0.5[\nu(s)-\nu(s_{0n})]\log n+\gamma_m[\log\binom{p}{\nu_m(s)}-\log\binom{p}{\nu_m(s_{0n})}] \\
&\quad +\gamma_I[\log\binom{\frac{p(p-1)}{2}}{\nu_I(s)}-\log\binom{\frac{p(p-1)}{2}}{\nu_I(s_{0n})}] \\
&\geq \; -0.5p_{0n}(\log n+2\gamma_m\log p+4\gamma_I\log p)>-cn^{1/3}\; \textit{for some positive c.}
\end{aligned}$$

We then need to show that the probability that this inequality occurs goes to zero. Let $\widetilde{s}=s\cup s_{0n}$ for any $s\in A_1$. For those $\boldsymbol{\beta}(\widetilde{s})$ near $\boldsymbol{\beta}_0(\widetilde{s})$, we get

$$\begin{aligned}
&l_n(\boldsymbol{\beta}(\widetilde{s}))-l_n(\boldsymbol{\beta}_0(\widetilde{s})) \\
= \; &\{\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\}^\tau s_n(\boldsymbol{\beta}_0(\widetilde{s}))-\frac{1}{2}\{\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\}^\tau H_n(\boldsymbol{\beta}^*(\widetilde{s}))\{\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\},
\end{aligned}$$

where $\boldsymbol{\beta}^*(\widetilde{s})$ is between $\boldsymbol{\beta}(\widetilde{s})$ and $\boldsymbol{\beta}_0(\widetilde{s})$. Clearly, by C5 and C6,

$$\{\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\}^\tau H_n(\boldsymbol{\beta}^*(\widetilde{s}))\{\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\}\geq k_1 n(1-\epsilon)\|\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\|_2^2.$$

Thus

$$l_n(\boldsymbol{\beta}(\widetilde{s}))-l_n(\boldsymbol{\beta}_0(\widetilde{s}))\leq\{\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\}^\tau s_n(\boldsymbol{\beta}_0(\widetilde{s}))-\frac{k_1}{2}n(1-\epsilon)\|\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\|_2^2.$$

For any $\boldsymbol{\beta}(\widetilde{s})$ satisfies $\|\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\|_2=n^{-1/4}$, we obtain

$$l_n(\boldsymbol{\beta}(\widetilde{s}))-l_n(\boldsymbol{\beta}_0(\widetilde{s}))\leq n^{-1/4}\|s_n(\boldsymbol{\beta}_0(\widetilde{s}))\|_2-\frac{k_1}{2}(1-\epsilon)n^{1/2}.$$

By (2.16), it can be deduced that $\max_{s\in A_1}\|s_n(\boldsymbol{\beta}_0(\widetilde{s})\|_2=O_p((nk)^{1/2})$ for $k=o(n^{1/3})$. Thus,

$$\begin{aligned}
&\max\{l_n(\boldsymbol{\beta}(\widetilde{s}))-l_n(\boldsymbol{\beta}_0(\widetilde{s})):\|\boldsymbol{\beta}(\widetilde{s})-\boldsymbol{\beta}_0(\widetilde{s})\|_2=n^{-1/4},s\in A_1\} \\
\leq \; &c\{n^{-1/4}(nk)^{1/2}-n^{1/2}\}\leq c(n^{5/12}-n^{1/2})\leq -cn^{1/2}
\end{aligned} \tag{2.17}$$

for a generic constant c. This (2.17) indicates that $l_n(\boldsymbol{\beta}(\widetilde{s}))$ obtains its maximum value inside $\|\boldsymbol{\beta}(\widetilde{s}) - \boldsymbol{\beta}_0(\widetilde{s})\|_2 \le n^{-1/4}$, because of its concavity. It also suggests

$$\begin{aligned} & \max\{l_n(\boldsymbol{\beta}(\widetilde{s})) - l_n(\boldsymbol{\beta}_0(\widetilde{s})) : \|\boldsymbol{\beta}(\widetilde{s}) - \boldsymbol{\beta}_0(\widetilde{s})\|_2 \ge n^{-1/4}, s \in A_1\} \\ \le \ & \max\{l_n(\boldsymbol{\beta}(\widetilde{s})) - l_n(\boldsymbol{\beta}_0(\widetilde{s})) : \|\boldsymbol{\beta}(\widetilde{s}) - \boldsymbol{\beta}_0(\widetilde{s})\|_2 = n^{-1/4}, s \in A_1\} \le -cn^{1/2}. \end{aligned}$$

After that, let $\overline{\boldsymbol{\beta}}(\widetilde{s}) = \binom{\widehat{\boldsymbol{\beta}}(s)}{\mathbf{0}}$. By C3, we can conclude

$$\|\overline{\boldsymbol{\beta}}(\widetilde{s}) - \boldsymbol{\beta}_0(\widetilde{s})\|_2 \ge \|\boldsymbol{\beta}_0(s_{0n} - s)\|_2 > n^{-1/4}.$$

As a consequence, with a probability tending to 1,

$$l_n(\widehat{\boldsymbol{\beta}}(s)) - l_n(\boldsymbol{\beta}_0(s_{0n})) = l_n(\overline{\boldsymbol{\beta}}(\widetilde{s})) - l_n(\boldsymbol{\beta}_0(\widetilde{s})) \le -cn^{1/2}$$

uniformly for all $s \in A_1$. Therefore, (2.14) is proved.

Case 2: proof for (2.15)

For $s \in A_0$, let $c = \nu(s) - \nu(s_{0n})$, $c_1 = \nu_m(s) - \nu_m(s_{0n})$ and $c_2 = \nu_I(s) - \nu_I(s_{0n})$. Clearly, $c = c_1 + c_2$. By (2.6), $EBIC_\gamma(s) \le EBIC_\gamma(s_{0n})$ is equivalent to

$$\begin{aligned} & l_n(\widehat{\boldsymbol{\beta}}(s)) - l_n(\widehat{\boldsymbol{\beta}}(s_{0n})) \\ \ge \ & 0.5[(\nu(s) - \nu(s_{0n})] \log n + \gamma_m[\log \binom{p}{\nu_m(s)} - \log \binom{p}{\nu_m(s_{0n})}] \\ & + \gamma_I[\log \binom{\frac{p(p-1)}{2}}{\nu_I(s)} - \log \binom{\frac{p(p-1)}{2}}{\nu_I(s_{0n})}] \\ \ge \ & (0.5c \log n + \gamma_m c_1 \log p + \gamma_I c_2 \log \frac{p(p-1)}{2})(1 + o(1)). \end{aligned}$$

We then show that, the probability that this inequality occur goes to zero uniformly for $s \in A_0$ with $\nu_m(s) = c_1$ and $\nu_I(s) = c_2$.

When n is sufficiently large,

$$\begin{aligned}
& l_n(\widehat{\boldsymbol{\beta}}(s)) - l_n(\widehat{\boldsymbol{\beta}}(s_{0n})) \\
\leq\ & l_n(\widehat{\boldsymbol{\beta}}(s)) - l_n(\boldsymbol{\beta}_0(s)) \\
\leq\ & \{\widehat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\}^\tau s_n(\boldsymbol{\beta}_0(s)) - \frac{1-\epsilon}{2}\{\widehat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\}^\tau H_n(\boldsymbol{\beta}_0(s))\{\widehat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\} \\
\leq\ & \frac{1}{2(1-\epsilon)} s_n^\tau(\boldsymbol{\beta}_0(s))\{H_n(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s)).
\end{aligned}$$

The next, we show that the following

$$\frac{s_n^\tau(\boldsymbol{\beta}_0(s))\{H_n(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s))}{2(1-\epsilon)} \geq 0.5c\log n + \gamma_m c_1 \log p + \gamma_I c_2 \log\frac{p(p-1)}{2}$$

does not occur. Due to the fact that $\{H_n(\boldsymbol{\beta}_0(s))\}^{-1/2} s_n(\boldsymbol{\beta}_0(s))$ is a linear combination of $y_i - \mu_i$, follow (2.16), then for every $s \in A_0$,

$$\begin{aligned}
& P[s_n^\tau(\boldsymbol{\beta}_0(s))\{H_n(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s)) \geq 2(1-\epsilon)(0.5c\log n + \gamma_m c_1 \log p + \gamma_I c_2 \log\frac{p(p-1)}{2})] \\
\leq\ & exp\{-(1-\epsilon)(0.5c\log n + \gamma_m c_1 \log p + \gamma_I c_2 \log\frac{p(p-1)}{2})\}
\end{aligned}$$

with an arbitrarily small but generic $\epsilon > 0$.

Let

$$0.5c_1 \log n + \gamma_m c_1 \log p \geq c_1 \log p \tag{2.18}$$

and

$$0.5c_2 \log n + \gamma_I c_2 \log\frac{p(p-1)}{2} \geq c_2 \log\frac{p(p-1)}{2}. \tag{2.19}$$

We have $r_m \geq 1 - \frac{\log n}{2\log p}$ and $r_I \geq 1 - \frac{0.5\log n}{\log\frac{p(p-1)}{2}}$. Plus (2.18)-(2.19), thus

$$exp\{-(1-\epsilon)(0.5c\log n + \gamma_m c_1 \log p + \gamma_I c_2 \log\frac{p(p-1)}{2})\} \leq p^{-c_1(1-\epsilon)}(\frac{p(p-1)}{2})^{-c_2(1-\epsilon)}$$

when $r_m \geq 1 - \frac{\log n}{2 \log p}$ and $r_I \geq 1 - \frac{\log n}{4 \log p}$.

For any fixed c_1, c_2, the number of models in A_0 is

$$C_p^{c_1} C_{p(p-1)/2}^{c_2} \leq p^{c_1} (\frac{p(p-1)}{2})^{c_2}.$$

Therefore, uniformly for $s \in A_0$,

$$P(\frac{s_n^\tau(\boldsymbol{\beta}_0(s))\{H_n(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s))}{2(1-\epsilon)} \geq 0.5c \log n + \gamma_m c_1 \log p + \gamma_I c_2 \log \frac{p(p-1)}{2}) \to 0.$$

This completes the proof for (2.15).

CHAPTER 3

Feature Selection Procedures

In high dimensional space, EBIC can identify the optimal model from candidate models with cardinality up to rp_{0n} $(r > 1)$. With the application of EBIC, we develop a novel feature selection procedure, referred to as sequential L_1 regularization algorithm (SLR), in this chapter. This chapter comprises three sections. The first two sections separately explore SLR under models with only main effects and interactive models while the third section aims at investigating theoretical properties of SLR. In section 3.1, SLasso (Luo and Chen, 2013b), a powerful procedure for high dimensional linear regression, is reviewed first. Analogous to SLasso, we select the next feature (features) maximizing the profile marginal score function, and propose SLR for feature selection in GLMs. In section 3.2, SLR is extended from models with only main effects to interactive models. The core idea for this extension is to group features into main effects and interactive effects and handle them differently. In section 3.3, the selection consistency of SLR under a GLM with the canonical link is established and the corresponding conditions required are provided.

3.1 Models with Only Main Effects

3.1.1 Linear Model: SLasso

A linear model with only main effects is given by

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{y} = (y_1, ..., y_n)^\tau$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\tau$, $X = (x_{ij})_{n\times p}$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\tau$ with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. For feature selection under this LM, SLasso (Luo and Chen, 2013b) is superior to other selection procedures. It selects features sequentially by letting earlier selected features not be penalized in later steps, which can be described as follows.

- SLasso starts with the L_1 penalized sum of squares:

  $$l_1 = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

  l_1 is minimized by tuning λ to the largest value to allow some β_j nonzero. Denote the set of indices of all nonzero β_j by s_1 and it is referred to as the active set.

- Assume the active set s_k is obtained after k steps have been carried out. In the $(k+1)^{th}$ step, the partial penalized function

  $$l_{k+1} = \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j \in s_k^c} |\beta_j|$$

  is then minimized by letting λ to be the largest value to allow at least one β_j with $j \in s_k^c$ nonzero. All features with nonzero estimated coefficients then form the active set s_{k+1}.

This process continues until some stopping rule is satisfied. EBIC (Chen and Chen, 2008) serves as an appropriate and workable stopping rule because of its selection consistency and the tendency of $\min_{\nu(s)=k} EBIC_\gamma(s) > \min_{\nu(s)=k+1} EBIC_\gamma(s)$.

Assume X_j is the j^{th} column vector of X and $X(s)$ is the matrix composed of the columns of X with indices in s. For $j \in s^c$, define

$$\gamma(X_j|s) = X_j^\tau (I - H_n(s))\boldsymbol{y} \tag{3.2}$$

for $H_n(s) = X(s)(X^\tau(s)X(s))^{-1}X^\tau(s)$. From the point of calculation, the process of SLasso can be restated as follows.

- Initial Step: Standardize $\boldsymbol{y}$, X_j, $j = 1, 2, ..., p$ such that $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n y_i^2 = n$, $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = n$. SLasso selects the feature (features) given by

$$s_1 = \{l : |X_l^\tau \boldsymbol{y}| = \max_{j=1,\dots,p} |X_j^\tau \boldsymbol{y}|\}.$$

- General Step: For $k \geq 1$, let

$$s_{temp} = \{l : |\gamma(X_l|s_k)| = \max_{j \in s_k^c} |\gamma(X_j|s_k)|\}.$$

  Update $s_{k+1} = s_k \cup s_{temp}$. If $EBIC_\gamma(s_{k+1}) > EBIC_\gamma(s_k)$, stop and take s_k as the optimal model; otherwise, continue.

Clearly, the feature in s_{temp} corresponds to the estimated nonzero β_j with $j \in s_k^c$. After s_k is obtained, SLasso selects the next feature or features maximizing $|\gamma(X_j|s_k)|$, which shares the same way of identification with OMP (Cai and Wang, 2011). These two procedures choose the same feature when only one feature maximizes $|\gamma(X_j|s_k)|$. However, SLasso differs from OMP when there are more than one features in s_{temp}. This difference is embodied in that OMP selects all these features whereas SLasso may not select all of them due to the restriction of *partial cone condition* (Luo and Chen, 2013b). Actually, there are very few cases when there are more than one features that maximize $|\gamma(X_j|s_k)|$. Thus, the difference of SLasso and OMP can be passed over. In general, SLasso is essentially equivalent to OMP.

3.1.2 Generalized Linear Model: SLR

SLasso (Luo and Chen, 2013b) and OMP (Cai and Wang, 2011) are powerful for high dimensional linear regression because of selection consistency and fast implementation. These two sequential procedures select the next feature that maximizes $|\gamma(X_j|s)|$ ($j \in s^c$). The identification criterion $\gamma(X_j|s)$ deserves to be extended to GLMs in view of the similarity of LMs and GLMs. It is interpreted from the perspective of residuals by Cai and Wang (2011). Nevertheless, there are several kinds of residuals for GLMs, for instance, raw residuals, pearson residuals and deviance residuals, which makes it quite difficult to be promoted. Fortunately, it can also be interpreted from the perspective of score function. Compared with residuals, the score function is advantageous in that it is unique in both LMs and GLMs while it is also frequently applied in statistics.

In this subsection, we propose a novel sequential L_1 regularization algorithm (SLR) to conduct feature selection in GLMs. SLR is implemented by promoting the identification criterion $\gamma(X_j|s)$ from the perspective of score function. There are three parts for this subsection. The concept of profile marginal score function is introduced in the first part and we show that $\gamma(X_j|s)$ can be described as a profile marginal score function in the context of LMs. In the second part, we provide a detailed description for SLR by applying the profile marginal score function under a GLM with only main effects. Finally, i.e. the third part, we modify SLR in the logistic model, an integral part of GLMs, when separation (Albert and Anderson, 1984) occurs.

3.1.2.1 Profile Marginal Score: $\gamma(X_j|s)$

Most practical problems of parameter inference aim at inferring part of the parameter vector of interest in the presence of nuisance parameters, thus motivate the emergence of profile evaluation. For the parameter vector $(\boldsymbol{\psi}, \boldsymbol{\omega})$ with a nuisance $\boldsymbol{\omega}$, the profile evaluation firstly supposes $\boldsymbol{\psi}$ is known and then rewrites the log-likelihood function as $l_n(\boldsymbol{\psi}, \boldsymbol{\omega}) = l_{\boldsymbol{\psi}}(\boldsymbol{\omega})$ to show that $\boldsymbol{\omega}$ varies whereas $\boldsymbol{\psi}$ is fixed. To estimate $\boldsymbol{\omega}$, it maximizes $l_{\boldsymbol{\psi}}(\boldsymbol{\omega})$, i.e. $\widetilde{\boldsymbol{\omega}}_{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\omega}}\ l_{\boldsymbol{\psi}}(\boldsymbol{\omega})$, and succeeds in evaluating $\widetilde{\boldsymbol{\omega}}_{\boldsymbol{\psi}}$ for each $\boldsymbol{\psi}$. The interest $\boldsymbol{\psi}$ can then be estimated by

$$\widetilde{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}}\ l_{\boldsymbol{\psi}}(\widetilde{\boldsymbol{\omega}}_{\boldsymbol{\psi}}) = \arg\max_{\boldsymbol{\psi}}\ l_n(\boldsymbol{\psi}, \widetilde{\boldsymbol{\omega}}_{\boldsymbol{\psi}}). \tag{3.3}$$

In this way, the nuisance $\boldsymbol{\omega}$ is profiled out. A bit of logical deduction illustrates that $\widetilde{\boldsymbol{\psi}}$ and $\widetilde{\boldsymbol{\omega}}_{\widetilde{\boldsymbol{\psi}}}$ are maximum likelihood estimators $(\widetilde{\boldsymbol{\psi}}, \widetilde{\boldsymbol{\omega}}) = \arg\max_{\boldsymbol{\psi},\boldsymbol{\omega}}\ l_n(\boldsymbol{\psi}, \boldsymbol{\omega})$. The log-likelihood function $l_{\boldsymbol{\psi}}(\widetilde{\boldsymbol{\omega}}_{\boldsymbol{\psi}}) = l_n(\boldsymbol{\psi}, \widetilde{\boldsymbol{\omega}}_{\boldsymbol{\psi}})$ is completely in terms of $\boldsymbol{\psi}$ and is referred to as the profile log-likelihood function. Take the derivative of $l_n(\boldsymbol{\psi}, \widetilde{\boldsymbol{\omega}}_{\boldsymbol{\psi}})$ with respect to ψ_j, and the corresponding function can be called the profile marginal score function.

We then show that $\gamma(X_j|s) = X_j^\tau(I - H_n(s))\boldsymbol{y}$ can be interpreted as the profile marginal score function of β_j with $j \in s^c$. Decompose the parameter vector $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}(s), \boldsymbol{\beta}(s^c))$, where $\boldsymbol{\beta}(s)$ denotes the sub-vector consisting of the components of $\boldsymbol{\beta}$ with indices in s. We firstly assume $\boldsymbol{\beta}(s^c)$ is fixed, and obtain a $\widetilde{\boldsymbol{\beta}}(s)$ which varies with $\boldsymbol{\beta}(s^c)$ through

$$\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)}|_{\boldsymbol{\beta}(s^c)} = \frac{\partial l_{\boldsymbol{\beta}(s^c)}(\boldsymbol{\beta}(s))}{\partial \boldsymbol{\beta}(s)} = \mathbf{0}. \tag{3.4}$$

Under LMs, the equation (3.4) indicates

$$\widetilde{\boldsymbol{\beta}}(s) = (X^\tau(s)X(s))^{-1}X^\tau(s)(\boldsymbol{y} - X(s^c)\boldsymbol{\beta}(s^c)) \tag{3.5}$$

and

$$\sum_{i=1}^{n}(y_i - \widetilde{\mu}_i)x_{ij} = 0 \; for \; all \; j \in s, \tag{3.6}$$

where $\widetilde{\mu}_i = \boldsymbol{x}_i^\tau(s)\widetilde{\boldsymbol{\beta}}(s) + \boldsymbol{x}_i^\tau(s^c)\boldsymbol{\beta}(s^c)$ while $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^\tau$.

For $j \in s^c$, take the first derivative of $l_n(\widetilde{\boldsymbol{\beta}}(s), \boldsymbol{\beta}(s^c))$, we have the profile marginal score function

$$\frac{\partial l_n(\widetilde{\boldsymbol{\beta}}(s), \boldsymbol{\beta}(s^c))}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - \widetilde{\mu}_i)(\boldsymbol{x}_i(s)\frac{\partial \widetilde{\boldsymbol{\beta}}(s)}{\partial \beta_j} + x_{ij}). \tag{3.7}$$

Plug (3.6)-(3.7), it becomes

$$\frac{\partial l_n(\widetilde{\boldsymbol{\beta}}(s), \boldsymbol{\beta}(s^c))}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - \widetilde{\mu}_i)x_{ij} = X_j^\tau(\boldsymbol{y} - \widetilde{\boldsymbol{\mu}}), \; \forall j \in s^c. \tag{3.8}$$

Specially, when $\boldsymbol{\beta}(s^c) = \mathbf{0}$, features outside the current active set s are treated equally. Impacts of these features on the variation of $l_n(\widetilde{\boldsymbol{\beta}}(s), \boldsymbol{\beta}(s^c))$ are measured and the feature with the greatest impact, i.e. the feature corresponds to the largest absolute profile marginal score value, is selected by SLasso and OMP. Let $\widehat{\boldsymbol{\beta}}(s) = \widetilde{\boldsymbol{\beta}}(s)|_{\boldsymbol{\beta}(s^c)=\mathbf{0}}$ and $\widehat{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}}|_{\boldsymbol{\beta}(s^c)=\mathbf{0}}$. Combining with the equation (3.5), for $j \in s^c$, we have

$$\frac{\partial l_n(\widetilde{\boldsymbol{\beta}}(s), \boldsymbol{\beta}(s^c))}{\partial \beta_j}|_{\boldsymbol{\beta}(s^c)=\mathbf{0}} = X_j^\tau(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = X_j^\tau(\boldsymbol{y} - X(s)\widehat{\boldsymbol{\beta}}(s)) = X_j^\tau(I - H_n(s))\boldsymbol{y}. \tag{3.9}$$

Thus $\gamma(X_j|s)$ is indeed a profile marginal score function with respect to β_j.

3.1.2.2 SLR in GLM

A GLM including only main effects is composed of three components. Firstly, the response variable $\boldsymbol{y} = (y_1, ..., y_n)^\tau \sim \prod_{i=1}^n f(y_i|\theta_i) = \prod_{i=1}^n exp\{\theta_i y_i - b(\theta_i)\}$. At the same time, the mean $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^\tau$ and variance $\boldsymbol{\sigma}^2 = (\sigma_1^2, ..., \sigma_n^2)^\tau$ of $\boldsymbol{y}$ satisfy $\mu_i = b'(\theta_i)$ and $\sigma_i^2 = b''(\theta_i)$. The second component is a linear predictor $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\tau$ which is expressed by $\eta_i = \boldsymbol{x}_i^\tau \boldsymbol{\beta}$, where $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{ip})^\tau$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\tau$. The third component is the link function g between $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$, i.e. $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. The coefficient vector $\boldsymbol{\beta}$ of this GLM is slightly different with that of previous sections because of the existence of the intercept β_0, as well as the design matrix $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^\tau$. Thus for a current active set s, we keep $\boldsymbol{\beta}(s^c)$ and $X(s^c)$ unchanged whereas redefine $\boldsymbol{\beta}(s)$ and $X(s)$ by automatically including the corresponding β_0 and $X_0 = (1, ..., 1)^\tau$. For instance, when $s = \{j\}$, let $\boldsymbol{\beta}(s) = (\beta_0, \beta_j)^\tau$ while $X(s) = (X_0 \ X_j)$ with $X_j = (x_{1j}, ..., x_{nj})^\tau$.

In this GLM, the log-likelihood function is given by

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \theta(\boldsymbol{\beta})_i - b(\theta(\boldsymbol{\beta})_i)\}. \tag{3.10}$$

Take the first derivative of $l_n(\boldsymbol{\beta})$ with respect to β_j for any j, we have

$$\frac{\partial l_n(\boldsymbol{\beta})}{\partial \beta_j} = X_j^\tau W(\boldsymbol{y} - \boldsymbol{\mu}) g'(\boldsymbol{\mu}), \tag{3.11}$$

where W is a diagonal matrix with n diagonal elements $W_{ii} = 1/b''(\theta_i)(g'(\mu_i))^2$ and $(\boldsymbol{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu})$ is a vector with n elements $(y_i - \mu_i)g'(\mu_i)$. SLR selects the next feature among features outside the current active set s that maximize $|\gamma_g(X_j|s)|$, where

$$\gamma_g(X_j|s) = X_j^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) g'(\widehat{\boldsymbol{\mu}}), \ j \in s^c. \tag{3.12}$$

This identification criterion $\gamma_g(X_j|s)$ is obtained from the point of profile marginal score function, which is analogous to $\gamma(X_j|s)$. The corresponding process can be described in detail as follows.

Assume $\boldsymbol{\beta}(s^c)$ is fixed and take the derivative of $\boldsymbol{\beta}(s)$, we have $\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)}|_{\boldsymbol{\beta}(s^c)} = \mathbf{0}$. It indicates

$$X^\tau(s)\widetilde{W}(\boldsymbol{y} - \widetilde{\boldsymbol{\mu}}) g'(\widetilde{\boldsymbol{\mu}}) = \mathbf{0}. \tag{3.13}$$

Both $\widetilde{W}$ and $\widetilde{\boldsymbol{\mu}}$ depend only on $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}(s), \boldsymbol{\beta}(s^c))$ with $\widetilde{\boldsymbol{\beta}}(s) = \arg\max_{\boldsymbol{\beta}(s)} l_n(\boldsymbol{\beta}(s), \boldsymbol{\beta}(s^c))$.

Specially, when $\boldsymbol{\beta}(s^c) = \mathbf{0}$, we have the profile marginal score function

$$\frac{\partial l_n(\widetilde{\boldsymbol{\beta}}(s), \boldsymbol{\beta}(s^c))}{\partial \beta_j}|_{\boldsymbol{\beta}(s^c)=\mathbf{0}} = X_j^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) g'(\widehat{\boldsymbol{\mu}}), j \in s^c. \tag{3.14}$$

The $\widehat{W}$ and $\widehat{\boldsymbol{\mu}}$ in (3.14) vary with $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}|_{\boldsymbol{\beta}(s^c)=\mathbf{0}} = (\widehat{\boldsymbol{\beta}}(s), \mathbf{0})$, thus can be written as $\widehat{W} = W(\widehat{\boldsymbol{\beta}})$ and $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$.

Unlike LMs, it is unable to get $\widehat{\boldsymbol{\beta}}$ directly, thus we apply the popular iterated weighted least squares (IWLS) procedure to solve

$$\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)}|_{\boldsymbol{\beta}(s^c)=\mathbf{0}} = X^\tau(s) \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) g'(\widehat{\boldsymbol{\mu}}) = \mathbf{0}. \tag{3.15}$$

The IWLS procedure starts with an initial value $\boldsymbol{\beta}^{(0)} = \begin{pmatrix} \boldsymbol{\beta}^{(0)}(s) \\ \mathbf{0} \end{pmatrix}$. For the positive integer $h \geq 1$, let $\boldsymbol{\beta}^{(h)} = (\boldsymbol{\beta}^{(h)}(s), \boldsymbol{\beta}^{(h)}(s^c))$. The $\boldsymbol{\beta}^{(h)}(s^c)$ is always fixed as $\mathbf{0}$ while $\boldsymbol{\beta}^{(h)}(s)$ can be obtained through

$$\boldsymbol{\beta}^{(h)}(s) = \boldsymbol{\beta}^{(h-1)}(s) + (X^\tau(s) W(\boldsymbol{\beta}^{(h-1)}) X(s))^{-1} X^\tau(s) W(\boldsymbol{\beta}^{(h-1)}) \boldsymbol{z}(\boldsymbol{\beta}^{(h-1)}), \tag{3.16}$$

where $\boldsymbol{z}(\boldsymbol{\beta}^{(h-1)}) = (\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(h-1)})) g'(\boldsymbol{\mu}(\boldsymbol{\beta}^{(h-1)}))$. That's because

$$\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{new}} \approx \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{old}} + \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s) \partial \boldsymbol{\beta}^\tau(s)}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{old}} (\boldsymbol{\beta}^{new}(s) - \boldsymbol{\beta}^{old}(s)) \approx \mathbf{0} \tag{3.17}$$

and

$$E(\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s) \partial \boldsymbol{\beta}^\tau(s)}) = -X^\tau(s) W X(s). \tag{3.18}$$

This fitting process would be stopped if $\|\boldsymbol{\beta}^{(h)} - \boldsymbol{\beta}^{(h-1)}\|_2^2$ is close to 0. In this way, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(h)}$ and $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$, $\widehat{W} = W(\widehat{\boldsymbol{\beta}})$ can be got subsequently.

In this IWLS process, an appropriate $\boldsymbol{\beta}^{(0)}$ is required to be given first. For the sequential procedure SLR, it is reasonable to let the initial $\boldsymbol{\beta}^{(0)}$ of the later step be the final $\widehat{\boldsymbol{\beta}}$ of the previous step, which suggests that we only need to decide an initial $\boldsymbol{\beta}^{(0)}$ at the very start of SLR. Let the starting model $s_0 = \{\emptyset\}$. Due to the fact that $\frac{\partial l_n(\boldsymbol{\beta})}{\partial \beta(s_0)} = \mathbf{1}^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}}) = 0$ when $\beta_j = 0$ for $j = 1, 2, ...p$, the initial estimator can be $(\widehat{\beta}_0, 0, ..., 0)^\tau$ with $\widehat{\beta}_0 = g(\overline{y})$.

Then we describe SLR in details on the basis of the identification criterion $\gamma_g(X_j|s)$ give in (3.12).

- Initial step: Standardize X_j, $j = 1, 2, .., p$; Initialize $\widehat{\boldsymbol{\beta}} = (g(\overline{y}), 0, ..., 0)^\tau$, $\widehat{\boldsymbol{\mu}} = \overline{y}\mathbf{1}$, $\widehat{\boldsymbol{\theta}} = b'^{-1}(\widehat{\boldsymbol{\mu}})$, $\widehat{W} = diag\{1/b''(\widehat{\theta}_i)g'(\widehat{\mu}_i)^2\}$. SLR selects the feature (features) given by

$$s_1 = \{l : |X_l^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})| = \max_{j=1,2,..,p} |X_j^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})|\}.$$

- General Step: For $k \geq 1$,
  - IWLS: new $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\mu}}$, $\widehat{W}$:

    Initialize $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}$, $\boldsymbol{\mu}(\boldsymbol{\beta}^{(0)}) = \widehat{\boldsymbol{\mu}}$, $W(\boldsymbol{\beta}^{(0)}) = \widehat{W}$;

    For $h = 1, 2, 3, ...$

    $\boldsymbol{\beta}^{(h)}(s_k) = \boldsymbol{\beta}^{(h-1)}(s_k) + (X^\tau(s_k)W(\boldsymbol{\beta}^{(h-1)})X(s_k))^{-1}X^\tau(s_k)W(\boldsymbol{\beta}^{(h-1)})(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(h-1)}))g'(\boldsymbol{\mu}(\boldsymbol{\beta}^{(h-1)}))$ while $\boldsymbol{\beta}^{(h)}(s_k^c) = \mathbf{0}$;

    $\boldsymbol{\mu}(\boldsymbol{\beta}^{(h)}) = g^{-1}(X\boldsymbol{\beta}^{(h)})$; $\boldsymbol{\theta}(\boldsymbol{\beta}^{(h)}) = b'^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}^{(h)}))$;

$W(\boldsymbol{\beta}^{(h)}) = diag\{1/b''(\theta_i(\boldsymbol{\beta}^{(h)}))g'(\mu_i(\boldsymbol{\beta}^{(h)}))^2\}$;

Stop if $\|\boldsymbol{\beta}^{(h-1)} - \boldsymbol{\beta}^{(h)}\|_2^2 \leq 1e - 10$;

New $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(h)}$, $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\boldsymbol{\beta}^{(h)})$, $\widehat{W} = W(\boldsymbol{\beta}^{(h)})$.

– Active set s_{k+1}:

$$\gamma_g(X_j|s_k) = X_j^{\tau}\widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}});$$

$$s_{temp} = \{l : |\gamma_g(X_l|s_k)| = \max_{j \in s_k^c} |\gamma_g(X_j|s_k)|\}.$$

Update $s_{k+1} = s_k \cup s_{temp}$. If $EBIC_\gamma(s_{k+1}) > EBIC_\gamma(s_k)$, stop and take s_k as the optimal model; otherwise, continue.

3.1.2.3 Special Situation: Separation In Logistic Model

The logistic model has become increasingly popular in many areas like medical or genome-wide association studies. In those logistic studies, datasets are usually small or sparse, which is likely to cause the phenomenon that is called separation (Albert and Anderson, 1984). Separation is a non-negligible problem in models where the response variable of interest is dichotomous, and it occurs when covariates perfectly predicts some binary outcomes (Heinze and Schemper, 2002). More specifically, in binary logistic model, separation occurs if there exists a $\boldsymbol{\beta}$ such that: $\boldsymbol{x}_i^{\tau}\boldsymbol{\beta} \geq 0$ for $i \in E_1$; $\boldsymbol{x}_i^{\tau}\boldsymbol{\beta} \leq 0$ for $i \in E_2$, where E_i represents the set of row identifiers of X for observations from the same value of response variable, i.e.

$E_1 = \{i : y_i = 1\}$, $E_2 = \{i : y_i = 0\}$.

When separation occurs, E_1 and E_2 are separated by one feature or a linear combination of some variables, which results in the monotonicity of log-likelihood function on at least one parameter. From the perspective of estimation, the separation phenomenon tends to lead to some infinite maximum likelihood estimates in the fitting process, thus it poses a challenge to MLE method. As a result, our SLR, which also includes a MLE procedure, might not work normally when separation exists.

To solve the problem caused by separation, Firth (1993) proposed a modified score procedure to remove $O(n^{-1})$ bias of MLE by adding the Jeffreys invariant prior $|I(\boldsymbol{\beta})|^{1/2}$ (Jeffreys, 1946), where $I(\boldsymbol{\beta})$ represents the fisher information matrix. This prior is shown to be an effective tool to produce finite MLE (Heinze and Schemper, 2002). Thus, under logistic models, we modify SLR by adding a penalty part $\log|I(\boldsymbol{\beta})|^{1/2}$ to $l_n(\boldsymbol{\beta})$ in the fitting process. This modification does not influence the performance of SLR much because the Jeffreys invariant prior is asymptotic negligible.

Consider the logistic model for a binary dependent variable y_i $(i \in 1, 2, ..n)$ which satisfies

$$P(y_i = 1|X) = 1 - P(y_i = 0|X) = \pi_i$$

and

$$\pi_i = \frac{exp\{\boldsymbol{x}_i^\tau \boldsymbol{\beta}\}}{1 + exp\{\boldsymbol{x}_i^\tau \boldsymbol{\beta}\}} \ or \ \boldsymbol{x}_i^\tau \boldsymbol{\beta} = \log \frac{\pi_i}{1 - \pi_i}.$$

Under this logistic regression, the new SLR follows the same way of the original SLR except for using the modified likelihood function in the IWLS process. The corresponding modified function can be expressed as

$$l_M(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \boldsymbol{x}_i^\tau \boldsymbol{\beta} - \log(1 + exp\{\boldsymbol{x}_i^\tau \boldsymbol{\beta}\})\} + 1/2 \log |I(\boldsymbol{\beta})|. \tag{3.19}$$

The matrix $I(\boldsymbol{\beta})$ depends on the previous selected indices s only and is given by

$$I(\boldsymbol{\beta}) = X^\tau(s) W X(s),$$

where $W = diag\{\pi_i(1 - \pi_i)\}$ with $\boldsymbol{\pi} = g^{-1}(X\boldsymbol{\beta})$.

Since

$$\frac{\partial \log |I(\boldsymbol{\beta})|}{\partial \beta_j} = tr[I(\boldsymbol{\beta})^{-1} \frac{\partial I(\boldsymbol{\beta})}{\partial \beta_j}].$$

Take the derivative of $l_M(\boldsymbol{\beta})$, we have

$$\frac{\partial l_M(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)} |_{\boldsymbol{\beta}(s^c)} = \sum_{i=1}^{n} \{y_i - \pi_i + H_{ii}(\frac{1}{2} - \pi_i)\} \boldsymbol{x}_i(s). \tag{3.20}$$

The H_{ii} in (3.20) represents the i_{th} diagonal element of the matrix $H_n(s)$, where

$$H_n(s) = W^{\frac{1}{2}} X(s) (X^\tau(s) W X(s))^{-1} X^\tau(s) W^{\frac{1}{2}}.$$

Redefine a new diagonal matrix $H_d = diag\{H_{ii}\}$, then (3.20) can be simplified as

$$\frac{\partial l_M(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)} |_{\boldsymbol{\beta}(s^c)} = X^\tau(s)[\boldsymbol{y} - \boldsymbol{\pi} + H_d(\frac{1}{2}\mathbf{1} - \boldsymbol{\pi})].$$

Subsequently, for $h \geq 1$, a finite $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}(s), \mathbf{0})$ can be obtained through

$$\boldsymbol{\beta}^{(h)}(s) = \boldsymbol{\beta}^{(h-1)}(s) + I^{-1}(\boldsymbol{\beta}^{(h-1)})X^{\tau}(s)[\boldsymbol{y} - \boldsymbol{\pi}(\boldsymbol{\beta}^{(h-1)}) + H_d(\boldsymbol{\beta}^{(h-1)})(\frac{1}{2}\mathbf{1} - \boldsymbol{\pi}(\boldsymbol{\beta}^{(h-1)}))]. \tag{3.21}$$

Thus, the computing algorithm for the modified SLR under the binary logistic model can be described as follows:

- Initial Step: Standardize X_j, $j = 1, 2, ..., p$; Initialize $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, 0, ..., 0)'$ with $\widehat{\beta}_0 = \log \overline{y}/(1 - \overline{y})$; $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\pi}} = \overline{y}\mathbf{1}$; $\widehat{W} = diag\{\widehat{\pi}_i(1 - \widehat{\pi}_i)\}$; $\widehat{H}_d = diag\{[\widehat{W}^{\frac{1}{2}}\mathbf{1}(\mathbf{1}^{\tau}\widehat{W}\mathbf{1})^{-1}\mathbf{1}^{\tau}\widehat{W}^{\frac{1}{2}}]_{ii}\}$. SLR chooses the feature given by

$$s_1 = \{l : |X_l^{\tau}\widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})| = \max_{j=1,2,..,p} |X_j^{\tau}\widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})|\},$$

  where $g'(\widehat{\boldsymbol{\mu}}) = 1/\widehat{\boldsymbol{\mu}} + 1/(1 - \widehat{\boldsymbol{\mu}})$.

- General Step: For $k \geq 1$,

  - IWLS: new $\widehat{\boldsymbol{\beta}}$, $\widehat{\mu}$, $\widehat{W}$, $\widehat{H_d}$:

    Initialize $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}$, $\boldsymbol{\pi}(\boldsymbol{\beta}^{(0)}) = \widehat{\boldsymbol{\pi}}$, $W(\boldsymbol{\beta}^{(0)}) = \widehat{W}$, $H_d(\boldsymbol{\beta}^{(0)}) = \widehat{H}_d$;

    For $h = 1, 2, 3, ...$

    $I(\boldsymbol{\beta}^{(h-1)}) = X^{\tau}(s_k)W(\boldsymbol{\beta}^{(h-1)})X(s_k)$;

    $\boldsymbol{\beta}^{(h)}(s_k) = \boldsymbol{\beta}^{(h-1)}(s_k) + I^{-1}(\boldsymbol{\beta}^{(h-1)})X^{\tau}(s_k)[\boldsymbol{y} - \boldsymbol{\pi}(\boldsymbol{\beta}^{(h-1)}) + H_d(\boldsymbol{\beta}^{(h-1)})(\frac{1}{2}\mathbf{1} - \boldsymbol{\pi}(\boldsymbol{\beta}^{(h-1)})]$ while $\boldsymbol{\beta}^{(h)}(s_k^c) = \mathbf{0}$;

    $\boldsymbol{\pi}(\boldsymbol{\beta}^{(h)}) = exp\{X\boldsymbol{\beta}^{(h)}\}/(1 + exp\{X\boldsymbol{\beta}^{(h)}\})$;

    $W(\boldsymbol{\beta}^{(h)}) = diag\{\boldsymbol{\pi}(\boldsymbol{\beta}^{(h)})_i(1 - \boldsymbol{\pi}(\boldsymbol{\beta}^{(h)}))_i\}$;

    $H_d(\boldsymbol{\beta}^{(h)}) = diag\{[W^{\frac{1}{2}}(\boldsymbol{\beta}^{(h)})X(s_k)(X^{\tau}(s_k)W(\boldsymbol{\beta}^{(h)})X(s_k))^{-1}X^{\tau}(s_k)W^{\frac{1}{2}}(\boldsymbol{\beta}^{(h)})]_{ii}\}$;

Stop if $\|\boldsymbol{\beta}^{(h-1)} - \boldsymbol{\beta}^{(h)}\|_2^2 < 1e - 10$;

New $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(h)}$, $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\boldsymbol{\beta}^{(h)})$, $\widehat{W} = W(\boldsymbol{\beta}^{(h)})$, $\widehat{H}_d = H_d(\boldsymbol{\beta}^{(h)})$.

– Active set s_{k+1}:

$$\gamma_g(X_j|s_k) = X_j^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}});$$

$$s_{temp} = \{l : |\gamma_g(X_l|s_k)| = \max_{j \in s_k^c} |\gamma_g(X_j|s_k)|\}.$$

Update $s_{k+1} = s_k \cup s_{temp}$. When $EBIC_\gamma(s_{k+1}) > EBIC_\gamma(s_k)$, stop and regard s_k as the optimal model; otherwise, continue.

3.2 Interactive Models

In some practical fields like QTL mapping, it is no longer enough to consider only main effect features because interactive effects are also an indispensable part for explaining the response variable. Thus, it is essential to popularize feature selection procedures in interactive models. In this section, we focus on extending the procedure SLR from models with only main effects to interactive models which include both main effects and pairwise interactive effects. We first introduce techniques that we apply for this extension, then we give a detailed description of SLR under the interactive case.

3.2.1 Techniques For Extension

Interactive models differ from models with only main effects primarily in two aspects. Firstly, the total number of features in interactive models is much larger due to the existence of interactions, thus computation complexity increases. Secondly, interactions are usually highly correlated and assumption of model sparsity may not hold. Thus for high dimensional feature selection, it is inappropriate to employ SLR in interactive models directly. Under interactive cases, SLR needs to be improved, taking into consideration of differences of models with only main effects and interactive models.

For interactive models with an extremely large number of features, it is natural to reduce the dimension of feature space first before selecting features when practical costs are taken into consideration. The maximum marginal likelihood estimator (MMLE) (Fan and Song, 2010) is a popular dimension reduction method for GLMs. It selects a set of features in $M_r = \{j : |\widehat{\beta}_j| \geq r\}$ for a predefined threshold value r, where $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{j,0}, \widehat{\beta}_j) = \arg\max_{\beta_0,\beta_j} \sum_{i=1}^{n} n^{-1} l(\beta_0 + \beta_j x_{ij}, y_i)$ and $l(y, \theta) = \theta y - b(\theta)$. Through this screening process, MMLE reduces the dimension to a proper number without losing important features. Thus, under interactive models, we start with a screening step by subjecting the main effect features and the interactive effect features to screening respectively. Those main effect features that survived the screening step are denoted by s_m while interactive features that survived the screening step are denoted by s_I. Then we turn to the further feature selection stage under the reduced lower-dimensional space.

Suppose some steps have been carried out and the active set s has been obtained. Then the further selection stage for SLR under interactive models mainly comprises three steps. The first step aims at dividing the survived main effect features and the survived interactive effect features into two different groups: G1 and G2, where $G1 = s_m$ and $G2 = s_I$. In the second step, we select the next feature (features) that maximizes $|\gamma_g(X_j|s)|$ separately for $j \in s_m \setminus s$ and $j \in s_I \setminus s$. The corresponding set of the selected main effect feature (features) is denoted by a_{temp} while the set of the selected interactive effect feature (features) is denoted by b_{temp}. Finally, i.e. the third step, the feature (features) in a_{temp} and b_{temp} are compared, and the better one is taken as the final selected feature (features). Clearly, the comparison of a_{temp} and b_{temp} is equivalent to selecting a better model from $s \cup a_{temp}$ and $s \cup b_{temp}$, thus it is natural to apply the model selection criterion EBIC (Chen and Chen, 2008) to make this comparison.

In general, the selection stage under interactive models is mainly different from the selection stage under models with only main effects in that the former groups features into main effects and interactive effects before it conducts feature selection separately on these two groups. This group selection keeps the flexibility of selecting features within a group. It differs from the classical group selection which chooses features in an all-in-all-out fashion, that is, all features of a group would be selected/deleted as long as any feature in this group is selected/deleted. In summary, our techniques for extension from models with only main effects to interactive models are particularly promising. Because they contribute to a better and more stable performance under different interaction proportions if compared with the case without them, i.e., the case applying SLR introduced in subsection 3.1.2 to interactive models directly.

3.2.2 SLR in Generalized Linear Interactive Model

A generalized linear interactive model (GLIM) is quite similar to a GLM introduced in the subsection 3.1.2.2, except that they have distinct X and $\boldsymbol{\beta}$. In a GLIM, for $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^\tau$,

$$\eta_i = \boldsymbol{x}_i^\tau \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} x_{ij}x_{ik}\beta_h,\ i = 1, 2, ..., n. \tag{3.22}$$

Clearly, the dimension of both $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$ is $p(p+1)/2+1$ rather than $p+1$. In addition, the last $p(p-1)/2$ components of $\boldsymbol{x}_i$ satisfy $x_{ih} = x_{ij}x_{ik}$ $(1 \leq j < k \leq p)$.

For feature selection under a GLIM, we firstly employ MMLE (Fan and Song, 2010) for screening and denote the set of survived main/interactive effect features by s_m/s_I. It is clear that $s_m \subset \{1, 2, ..., p\}$ and $s_I \subset \{p+1, ..., p(p+1)/2\}$. Define

$\overline{s}^c = s_m \setminus s$ and $\underline{s}^c = s_I \setminus s$. In the further selection stage, SLR can be restated as follows.

- Initial step:
  - Standardize X_j, $j = 1, 2, .., p$;

    Initialize $\widehat{\boldsymbol{\beta}} = (g(\overline{y}), 0, ..., 0)^\tau$, $\widehat{\boldsymbol{\mu}} = \overline{y}\mathbf{1}$, $\widehat{\boldsymbol{\theta}} = b'^{-1}(\widehat{\boldsymbol{\mu}})$, $\widehat{W} = diag\{1/b''(\widehat{\theta}_i)g'(\widehat{\mu}_i)^2\}$;
  - Main effect feature:

    $$a_1 = \{l : |X_l^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})| = \max_{j \in s_m} |X_j^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})|\};$$

    Interactive effect feature:

    $$b_1 = \{l : |X_l^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})| = \max_{j \in s_I} |X_j^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}})|\};$$
  - SLR selects the feature (features) given by s_1, where

    $s_1 = a_1$ if $EBIC_\gamma(a_1) < EBIC_\gamma(b_1)$;

    $s_1 = b_1$ if $EBIC_\gamma(a_1) > EBIC_\gamma(b_1)$.
- General Step: For $k \geq 1$,
  - IWLS: new $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\mu}}$, $\widehat{W}$:

    Initialize $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}$, $\boldsymbol{\mu}(\boldsymbol{\beta}^{(0)}) = \widehat{\boldsymbol{\mu}}$, $W(\boldsymbol{\beta}^{(0)}) = \widehat{W}$;

    For $h = 1, 2, 3, ...$

    $\boldsymbol{\beta}^{(h)}(s_k) = \boldsymbol{\beta}^{(h-1)}(s_k) + (X^\tau(s_k)W(\boldsymbol{\beta}^{(h-1)})X(s_k))^{-1}X^\tau(s_k)W(\boldsymbol{\beta}^{(h-1)})(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}^{(h-1)}))g'(\boldsymbol{\mu}(\boldsymbol{\beta}^{(h-1)}))$ while $\boldsymbol{\beta}^{(h)}(s_k^c) = \mathbf{0}$;

    $\boldsymbol{\mu}(\boldsymbol{\beta}^{(h)}) = g^{-1}(X\boldsymbol{\beta}^{(h)})$; $\boldsymbol{\theta}(\boldsymbol{\beta}^{(h)}) = b'^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}^{(h)}))$;

$W(\boldsymbol{\beta}^{(h)}) = diag\{1/b''(\theta_i(\boldsymbol{\beta}^{(h)}))g'(\mu_i(\boldsymbol{\beta}^{(h)}))^2\}$;

Stop if $\|\boldsymbol{\beta}^{(h-1)} - \boldsymbol{\beta}^{(h)}\|_2^2 \leq 1e - 10$;

New $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(h)}$, $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\boldsymbol{\beta}^{(h)})$, $\widehat{W} = W(\boldsymbol{\beta}^{(h)})$.

– Identification Criterion:

$$\gamma_g(X_j|s_k) = X_j^\tau \widehat{W}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}})g'(\widehat{\boldsymbol{\mu}});$$

– Main effect feature:

$$a_{temp} = \{l : |\gamma_g(X_l|s_k)| = \max_{j \in \overline{s_k}^c} |\gamma_g(X_j|s_k)|\};$$

Interactive effect feature:

$$b_{temp} = \{l : |\gamma_g(X_l|s_k)| = \max_{j \in \underline{s_k}^c} |\gamma_g(X_j|s_k)|\};$$

– SLR selects the feature (features) in s_{temp}, where

$s_{temp} = a_{temp}$ if $EBIC_\gamma(s_k \cup a_{temp}) < EBIC_\gamma(s_k \cup b_{temp})$;

$s_{temp} = b_{temp}$ if $EBIC_\gamma(s_k \cup a_{temp}) > EBIC_\gamma(s_k \cup b_{temp})$;

– Update the active set $s_{k+1} = s_k \cup s_{temp}$.

If $EBIC_\gamma(s_{k+1}) > EBIC_\gamma(s_k)$, stop and take s_k as the optimal model; otherwise, continue.

Clearly, EBIC (Chen and Chen, 2008) paves the way for this selection procedure SLR. The application of EBIC can be described in three aspects: deciding the final selected feature of each step from the selected main effect feature and the selected interactive effect feature; being the stopping rule; and identifying the best model from candidate models generated. All three applications are regarded as reasonable because of the selection consistency of EBIC introduced in chapter 2.

3.3 Theoretical Property

The selection consistency of SLR under GLMs with the canonical link is explored in this section. Notations without special explanations are the same as those in subsection 3.1.2.2 and we do not restate them again. Under the canonical link, that is, $\boldsymbol{\theta} = \boldsymbol{\eta}$, the identification criterion of SLR becomes

$$\gamma_g(X_j|s) = X_j^\tau(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}),\ j \in s^c,$$

where $\widehat{\boldsymbol{\mu}} = b'(X(s)\widehat{\boldsymbol{\beta}}(s))$ and $\widehat{\boldsymbol{\beta}}(s)$ is obtained through $\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(s)}|_{\boldsymbol{\beta}(s^c)=\mathbf{0}} = \mathbf{0}$. This $\gamma_g(X_j|s)$ can be decomposed into three parts. The first part $\gamma_1(X_j|s) = X_j^\tau(\boldsymbol{y} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = E(\boldsymbol{y}) = b'(X\boldsymbol{\beta}_0)$ and $\boldsymbol{\beta}_0$ denotes the vector consisting of true coefficients. The second part $\gamma_2(X_j|s) = X_j^\tau(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{1}})$, where $\boldsymbol{\mu}_{\mathbf{1}} = b'(X(s)\boldsymbol{\beta}_1(s))$ for $\boldsymbol{\beta}_1(s) = E(\widehat{\boldsymbol{\beta}}(s))$. Actually, $\boldsymbol{\beta}_1(s)$ is related to $\boldsymbol{\beta}_0$, since $\boldsymbol{\beta}_1(s) = \boldsymbol{\beta}_0(s) + (X^\tau(s)W_1X(s))^{-1}X^\tau(s)W_1X(s^c)\boldsymbol{\beta}_0(s^c)$ with $W_1 = diag\{b''(\boldsymbol{x}_i^\tau(s)\boldsymbol{\beta}_1(s))\}$. The third part $\gamma_3(X_j|s) = X_j^\tau(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}})$. Clearly, $\gamma_1(X_j|s)$ and $\gamma_3(X_j|s)$ vary with $\boldsymbol{y}$ whereas $\gamma_2(X_j|s)$ not.

Let $\nu(s)$ be the cardinality of s. Denote the set of relevant (true) features by

s_{0n} and $p_{0n} = \nu(s_{0n})$. Besides, let $s^- = s_{0n} \cap s^c$. The selection consistency of SLR is established under the following assumptions.

A1.

$$\max_{j \in s_{0n}^c} |\gamma_2(X_j|s)| < r \max_{j \in s^-} |\gamma_2(X_j|s)|, \; 0 < r < 1.$$

A2. For any $s \subset s_{0n}$ and $i = 1, ..., n$, there exists positive m and M such that $m \leq b''(x) \leq M$ for all $x \in [x_i^\tau(s)\boldsymbol{\beta}_1(s), \boldsymbol{x}_i^\tau \boldsymbol{\beta}_0]$.

A3.

$$n^{1/2}(\ln p)^{-1/2}\lambda_{\min}[\frac{X^\tau(s_{0n})X(s_{0n})}{n}] \min\{|\beta_{0j}| : j \in s_{0n}\} \to \infty.$$

The $\lambda_{\min}[\frac{X^\tau(s_{0n})X(s_{0n})}{n}]$ is commonly assumed to be bounded away from zero in high feature space, which suggests that A3 is equivalent to $n^{1/2}(\ln p)^{-1/2} \min\{|\beta_{0j}| : j \in s_{0n}\} \to \infty$. Thus SLR allows the scenario that $\ln p = O(n^k)$ $(k < 1)$ and $min_{j \in s_{0n}} |\beta_{0j}| > Cn^{-\delta}$ for $\delta < (1-k)/2$.

Theorem 3.1. *Under assumptions A1-A3, SLR is selection consistent in the sense that*

$$P(s_k = s_{0n}) \to 1, \; as \; n \to \infty,$$

when $\nu(s_k) = p_{0n}$.

Proof for Theorem 3.1: Look at the general $(k+1)^{th}$ step, the identification criterion is given by

$$\gamma_g(X_j|s_k) = \sum_{i=1}^{n}(y_i - \widehat{\mu}_i)x_{ij}, \tag{3.23}$$

where $\widehat{\mu}_i = b'(\boldsymbol{x}_i^\tau(s_k)\widehat{\boldsymbol{\beta}}(s_k))$. Define

$$s_{temp} = \{l : |\gamma_g(X_l|s_k)| = \max_{j\in s_k^c} |\gamma_g(X_j|s_k)|\}.$$

We will show that, with probability tending to 1, $s_{temp} \subset s_{0n}$.

The following statements are established first:

$$\gamma_1(X_j|s_k) = O_p(n^{1/2}\sqrt{\ln p}); \tag{3.24}$$

$$\gamma_3(X_j|s_k) = O_p(n^{1/2}\sqrt{\ln p}); \tag{3.25}$$

$$\max_{j\in s_k^-} |\gamma_2(X_j|s_k)| \geq C_n n^{1/2}\sqrt{\ln p} \ for \ C_n \to \infty. \tag{3.26}$$

Note that $\gamma_1(X_j|s_k) = X_j^\tau(\boldsymbol{y}-\boldsymbol{\mu}) \sim (0, \sum_{i=1}^n x_{ij}^2\sigma_i^2)$. By Chebyshev's inequality, for any $j \in s_k^c$,

$$P(|\gamma_1(X_j|s_k)| > n^{1/2}\sqrt{\ln p}) \leq \frac{\sum_{i=1}^n x_{ij}^2\sigma_i^2}{n\ln p} \leq \frac{\max_i \sigma_i^2}{\ln p} \to 0,$$

and thus (3.24) is proved.

When n is sufficiently large, $\widehat{\boldsymbol{\beta}}(s_k) \sim N(\boldsymbol{\beta}_1(s_k), (X^\tau(s_k)W_1X(s_k))^{-1})$, where $W_1 = diag\{b''(\boldsymbol{x}_i^\tau(s_k)\boldsymbol{\beta}_1(s_k))\}$. This, combined with the fact that

$$\boldsymbol{\beta}_1(s_k) = \boldsymbol{\beta}_0(s_k) + (X^\tau(s_k)W_1X(s_k))^{-1}X^\tau(s_k)W_1X(s_k^c)\boldsymbol{\beta}_0(s_k^c),$$

shows that

$$\begin{aligned}\gamma_3(X_j|s_k) &= X_j^\tau(\boldsymbol{\mu}_1 - \widehat{\boldsymbol{\mu}}) \\ &= -X_j^\tau(b'(X(s_k)\widehat{\boldsymbol{\beta}}(s_k)) - b'(X(s_k)\boldsymbol{\beta}_1(s_k))) \\ &= -X_j^\tau W_1 X(s_k)(\widehat{\boldsymbol{\beta}}(s_k) - \boldsymbol{\beta}_1(s_k))(1+o(1)).\end{aligned}$$

It is clear that

$$X_j^\tau W_1 X(s_k)(\widehat{\boldsymbol{\beta}}(s_k) - \boldsymbol{\beta}_1(s_k)) \sim N(0, \widetilde{X_j}^\tau \widetilde{H}_n(s_k)\widetilde{X_j}),$$

where $\widetilde{X_j} = W_1^{1/2}X_j$, $\widetilde{X}(s_k) = W_1^{1/2}X(s_k)$ and $\widetilde{H}_n(s_k) = \widetilde{X}(s_k)(\widetilde{X}^\tau(s_k)\widetilde{X}(s_k))^{-1}\widetilde{X}^\tau(s_k)$. Therefore, for any $j \in s_k^c$,

$$\begin{aligned}P(|\gamma_3(X_j|s_k)| > n^{1/2}\sqrt{\ln p}) &\leq \frac{\widetilde{X_j}^\tau \widetilde{H}_n(s_k)\widetilde{X_j}}{n \ln p} \\ &\leq \frac{\lambda_{\max}(\widetilde{H}_n(s_k))\sum_{i=1}^n x_{ij}^2 b''(\boldsymbol{x}_i^\tau(s_k)\boldsymbol{\beta}_1(s_k))}{n\ln p} \\ &\leq \frac{M}{\ln p} \to 0,\end{aligned}$$

which establish (3.25).

By Taylor inequality with Lagrange remainder term,

$$\begin{aligned}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{1}})_i &= b'(\boldsymbol{x}_i(s_k)\boldsymbol{\beta}_0(s_k) + \boldsymbol{x}_i(s_k^c)\boldsymbol{\beta}_0(s_k^c)) - b'(\boldsymbol{x}_i(s_k)\boldsymbol{\beta}_1(s_k)) \\ &= b''(\xi_i)[\boldsymbol{x}_i(s_k)(\boldsymbol{\beta}_0(s_k) - \boldsymbol{\beta}_1(s_k)) + \boldsymbol{x}_i(s_k^c)\boldsymbol{\beta}_0(s_k^c)].\end{aligned}$$

Let $W(\boldsymbol{\xi}) = diag\{b''(\xi_i)\}$, we have

$$\gamma_2(X_j|s_k) = X_j^\tau(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{1}}) = X_j^\tau W(\boldsymbol{\xi})[X(s_k)(\boldsymbol{\beta}_0(s_k) - \boldsymbol{\beta}_1(s_k)) + X(s_k^c)\boldsymbol{\beta}_0(s_k^c)].$$

Define

$$\begin{aligned}\gamma_E(X_j|s_k) &= X_j^\tau W_1[X(s_k)(\boldsymbol{\beta}_0(s_k)-\boldsymbol{\beta}_1(s_k))+X(s_k^c)\boldsymbol{\beta}_0(s_k^c)] \\ &= X_j^\tau W_1^{1/2}[I-\widetilde{H}_n(s_k)]W_1^{1/2}X(s_k^c)\boldsymbol{\beta}_0(s_k^c) \\ &= \widetilde{X_j}[I-\widetilde{H}_n(s_k)]\widetilde{X}(s_k^c)\boldsymbol{\beta}_0(s_k^c),\end{aligned}$$

where $\widetilde{X}(s_k^c)=W_1^{1/2}X(s_k^c)$.

We have

$$\begin{aligned}\Delta &= \boldsymbol{\beta}_0^\tau(s_k^c)\widetilde{X}^\tau(s_k^c)(I-\widetilde{H}_n(s_k))\widetilde{X}(s_k^c)\boldsymbol{\beta}_0(s_k^c) \\ &= \boldsymbol{\beta}_0^\tau(s_k^-)\widetilde{X}^\tau(s_k^-)(I-\widetilde{H}_n(s_k))\widetilde{X}(s_k^-)\boldsymbol{\beta}_0(s_k^-) \\ &\geq \lambda_{\min}(\widetilde{X}^\tau(s_k^-)(I-\widetilde{H}_n(s_k))\widetilde{X}(s_k^-))\|\boldsymbol{\beta}_0(s_k^-)\|_2^2 \\ &\geq \lambda_{\min}(\widetilde{X}^\tau(s_{0n})\widetilde{X}(s_{0n}))\|\boldsymbol{\beta}_0(s_k^-)\|_2^2\end{aligned}$$

for $\widetilde{X}(s_{0n}) = W_1^{1/2}X(s_{0n})$. This inequality is obtained because $(\widetilde{X}^\tau(s_k^-)(I-\widetilde{H}_n(s_k))\widetilde{X}(s_k^-))^{-1}$ is a sub-matrix of $(\widetilde{X}^\tau(s_{0n})\widetilde{X}(s_{0n}))^{-1}$ through the formula of the inverse of blocked matrices.

On the other hand,

$$\begin{aligned}\Delta &= \sum_{j\in s_k^-}\beta_{0j}\widetilde{X_j}^\tau(I-\widetilde{H}_n(s_k))\widetilde{X}(s_k^-)\boldsymbol{\beta}_0(s_k^-) \\ &\leq \|\boldsymbol{\beta}_0(s_k^-)\|_1\max_{j\in s_k^-}|\gamma_E(X_j|s_k)|.\end{aligned}$$

Thus

$$\max_{j\in s_k^-}|\gamma_E(X_j|s_k)|$$

$$
\begin{aligned}
&\geq \; n\lambda_{\min}(\frac{\widetilde{X}^\tau(s_{0n})\widetilde{X}(s_{0n})}{n})\frac{\|\boldsymbol{\beta}_0(s_k^-)\|_2^2}{\|\boldsymbol{\beta}_0(s_k^-)\|_1} \\
&\geq \; n\lambda_{\min}(\frac{X^\tau(s_{0n})W_1X(s_{0n})}{n})\min_{j\in s_{0n}}|\beta_{0j}| \\
&= \; B_n n^{1/2}\sqrt{\ln p},
\end{aligned}
$$

where $B_n = n^{1/2}(\ln p)^{-1/2}\lambda_{\min}(\frac{X^\tau(s_{0n})W_1X(s_{0n})}{n})\min_{j\in s_{0n}}|\beta_j|$. By A2 and A3 , $B_n \to \infty$. In addition, due to the fact that $b''(x) > 0$ and $W(\xi) \geq \frac{m}{M}W_1$, we have $\max_{j\in s_k^-}|\gamma_2(X_j|s_k)| \geq \frac{m}{M}B_n n^{1/2}\sqrt{\ln p} = C_n n^{1/2}\ln p$, thus (3.26) is proved.

Finally, A1 indicates that

$$
\max_{j\in s_k^-}|\gamma_2(X_j|s_k)| - \max_{j\in s_{0n}^c}|\gamma_2(X_j|s_k)| \geq (1-r)C_n n^{1/2}\ln p.
$$

This fact, combined with (3.24), (3.25), (3.26), implies that, with probability converging to 1,

$$
\max_{j\in s_k^-}|\gamma_g(X_j|s_k)| > \max_{j\in s_{0n}^c}|\gamma_g(X_j|s_k)|.
$$

Thus, with probability tending to 1, $s_{temp} \subset s_k^- \subset s_{0n}$. The proof is then completed.

CHAPTER 4

Numerical Study

In this chapter, extensive numerical studies are provided to show the effectiveness of our SLR with EBIC (Chen and Chen, 2008). In section 6.1, we introduce some measures and correlation structures which will be used in simulations. In section 6.2, for the study of only main effects, we explore SLR under the poisson log linear model and the logistic model. Three other competing approaches are provided for comparison with SLR and this comparison can be described from two aspects: selection consistency and prediction accuracy. In section 6.3, we simulate SLR under two popular interactive models, linear interactive model and logistic interactive model, through subsection 6.3.1 and 6.3.2 respectively. The effectiveness of EBIC is also demonstrated in these two subsections by comparing SLR under different (γ_m, γ_I).

4.1 Introduction

4.1.1 Measures

Six measures are provided for the assessment of SLR: positive discovery rate (PDR), false discovery rate (FDR), positive selection rate (PSR), false selection rate (FSR), model size and deviance.

Discovery Rate: PDR and FDR are primary measures used to assess sample performances and they are defined as

$$PDR = \frac{\nu(s_r \cap s_{0n})}{\nu(s_{0n})}, FDR = \frac{\nu(s_r \setminus s_{0n})}{\nu(s_r)}, \tag{4.1}$$

where s_r represents the optimal model and s_{0n} denotes the true model. The higher PDR and the lower FDR a procedure has, the better it is. For two procedures with almost identical PDR and FDR, the one with a smaller model size, i.e. a smaller $\nu(s_r)$, is viewed as better. The asymptotic property of SLR, that is, selection consistency, indicates that PDR approaches 1 and FDR converges to 0 simultaneously when n goes to infinity.

Selection Rate: PSR (FSR) are identical with PDR (FDR) under models with only main effects. Under interactive models, PSR and FSR can also be given by (4.1) although they have slightly different s_r and s_{0n}. Specifically, for PSR and FSR, we assume there are total p features while these features are not classified by main effects and interactive effects. The X_i corresponds to feature i and X_iX_j corresponds to feature i and feature j. Nevertheless, for PDR and FDR, we suppose there are $p(p+1)/2$ features consisting of main effects and interactive effects. The X_i corresponds to the main effect feature i while X_iX_j $(i < j)$ corresponds to the interactive feature $(2p - i + 1)i/2 + j - i$.

Deviance: Deviance is an important criterion used to evaluate fitting performances. It is expressed by

$$Deviance = 2\{l_n(saturated\ model) - l_n(fitted\ model)\}$$

The saturated model allows the author to choose a predicted value μ_i for each observation and it fits perfectly to such that $l_n(saturated\ model) = 0$ in most cases. The deviance reduces as the model fit improves. Specially, it is zero if the model exactly fits the data.

4.1.2 Correlation Structure

Refer to p as the number of main effect features and refer to p_{0n} as the number of causal features. Denote X_j $(1 \leq j \leq p)$ by the j^{th} column vector of X. The following correlation structures are considered for X_j.

Structure 1: Power decay correlation: $X_j = \rho X_{j-1} + \sqrt{1-\rho^2}\boldsymbol{z}_j$ for $j = 1, ..., p$, where $\boldsymbol{z}_j$ is independently and identically distributed (i.i.d) as standard normal distribution.

Structure 2: Features have a constant pairwise correlation, that is, the covariance matrix of covariates satisfies $\rho_{ij} = \rho$ and $\rho_{ii} = 1$.

Structure 3: This correlation structure is taken from Luo and Chen (2013b):

$$X_j = \frac{\boldsymbol{z}_j + \boldsymbol{w}_j}{\sqrt{2}} \; for \; j \in s_{0n}, \quad X_j = \frac{\boldsymbol{z}_j + \sum_{k \in s_{0n}} \boldsymbol{z}_k}{\sqrt{1 + p_{0n}}} \; for \; j \in s_{0n}^c,$$

where all $\boldsymbol{z}_j$ and $\boldsymbol{w}_j$ are i.i.d $\sim N(\mathbf{0}, I_n)$.

Structure 4: It is adapted from Fan and Song (2010). Let X_j $(1 \leq j \leq p - 50)$ be i.i.d $\sim N(\mathbf{0}, I_n)$. Other X_j $(i = p - 49, ..., p)$ is given by

$$X_j = \frac{1}{5}[\sum_{t=1}^{p_{0n}} (-1)^{t+1} X_{10t} + \sqrt{25 - p_{0n}}\xi_i]$$

for independent ξ_i following $N(\mathbf{0}, I_n)$.

Structure 5: This structure is slightly different from structure 2 in that its covariance matrix is a diagonal block matrix. Each block matrix except the last one is of dimension 100 × 100 while $\rho_{ii} = 1$ and $\rho_{ij} = \rho$ in each block.

Structure 6: This structure is motivated by the phenomenon that gene markers from different chromosomes have little correlations whereas makers within the same chromosome are correlated. Let $\boldsymbol{z}_1$ and $\boldsymbol{r}_j$ $(j = 1, ..., p)$ are i.i.d $\sim N(\mathbf{0}, I_n)$.

$$X_j = \frac{1}{\sqrt{5}}\boldsymbol{z}_1 + \frac{\sqrt{2}}{5}\boldsymbol{r}_j, \ 1 \leq j \leq p_{0n};$$

$$X_j = \frac{1}{2}X_{j-k} + \frac{\sqrt{3}}{2}\boldsymbol{r}_j, \ p_{0n} + 1 \leq j \leq p.$$

Structure 7: Let $r = [p/3]$. Asssume $(X_j)_{j=1}^r$ and $(\boldsymbol{\epsilon}_j)_{j=1}^p$ i.i.d $\sim N(\mathbf{0}, I_n)$.

$$X_j = 0.6X_{j-r} + 0.8\boldsymbol{\epsilon}_j, \ j = r+1, ..., 2r;$$

$$X_j = \sum_{i=1}^{r} X_j/\sqrt{j} + \sqrt{j - r/j}\boldsymbol{\epsilon}_j, \ j = 2r+1, ..., p.$$

4.2 Models with Only Main Effects

4.2.1 Sample Properties

In this subsection, not only selection consistency but also prediction accuracy are considered, thus we provide two parts. The first part deals with selection consistency and applies EBIC with $\gamma_{EBIC} = 1 - \frac{\ln n}{4 \ln p}$ in SLR to identify the optimal model. The second part focuses on prediction accuracy and uses deviance in SLR for model comparison. Under the study of the second part, three independent data sets are generated in the same way in each simulation but serve different purposes. The first set is used for fitting, the second data is intended for testing and the third set is used for comparison.

Under GLMs, three competing regularization methods with the following penalties are considered: M1: $p_\lambda(\beta) = \lambda \sum |\beta_j|$ (Park and Hastie, 2007); M2: $p_\lambda(\beta) = \lambda \sum \omega_j |\beta_j|$ (Zhou, 2006; Huang, Ma and Zhang, 2008); M3: $p'_\lambda(|\beta|) = \lambda I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{a-1} I(|\beta| > \lambda)$ (Fan and Li, 2001; Zou and Li, 2008). The well-known names for these three penalties under linear expressions are Lasso, adaptive Lasso and SCAD respectively. They share the same stopping rule with SLR in this study. The R package *glmpath* can be used directly for the computation of M1. When $p > n$, the weight ω_j in M2 is chosen as $1/|\widehat{\beta}_j|$, where $\widehat{\beta}_j$ denotes the marginal regression estimator. By letting $X^*_j = X_j/\omega_j$ and $\beta^*_j = \omega_j \beta_j$, *glmpath* can also be adopted in M2. For M3, the package *SIS* is employed.

We take the diverging pattern as $(n, p_{0n}, p) = (n, [4n^{0.155}], [4e^{n^{0.275}}])$ for $n = 100, 200, 400$. The true coefficient β_j $(j \in s_{0n})$ is given by $(-1)^u(0.8 + 0.05u)$ for $u \sim binomial(2, 0.5)$. Four different correlation structures and two models are considered in this study. Structure 1 and 2 are prepared for the poisson log linear model while structure 3 and 4 are applied in the binary logistic model. We let $\rho = 0.5$ in structure 2. Two hundred datasets are generated and analyzed for each simulation setting.

We firstly use a simple example to show that the Jeffreys invariant prior does not affect performances of SLR much. This conclusion is achieved through the comparison between SLR with and without Jeffreys prior when there is no separation. We fix 200 observations under structure 3 and reduce p to 8. Then the following result is obtained.

	PDR	FDR	Msize
SLR (with prior)	0.940	0.042	2.98
SLR (without prior)	0.940	0.042	2.98

Clearly, the performance of SLR with and without Jeffreys prior are exactly the same, which may owe to the asymptotic negligible effect of Jeffreys invariant prior.

Simulation results of the first part are reported in Table 4.1 and the following conclusions can be made. Firstly, the performance of SLR closely matches its asymptotic property, that is, PDR approaches rapidly to 1 and FDR decreases to 0, under all four structures and both two models. Secondly, SLR is regarded as a better procedure than M1, M2 and M3 due to its quite higher PDR and a slightly lower FDR. Subsequently, we apply the deviance in the second part and the corresponding results are presented in Table 4.2. As shown in this table, M3 has the smallest deviance while M2 has the largest. It implies that M3 is the optimal approach, followed by SLR and M1, then by M2 in terms of prediction accuracy. SLR is comparable with M1 for two reasons. Firstly, SLR fits better under small n whereas M1 performs better under large n. Secondly, SLR has a better fitting performance under the poisson log linear model but M2 fits better under the logistic model. It is worth noting that SLR is no longer the optimal approach when we focus on prediction accuracy, which suggests that the best select procedure should be decided by the aim of the study.

4.2.2 Real Data Example 1

In this example, we apply SLR to a popular cancer data called leukemia data (Golub et.al, 1999). Leukemia data has been analyzed by many authors, for example, Lee et.al (2003) and Liao and Chin (2007), through different classification methods. This data consists of two parts: initial data and independent data. There are 38 bone marrow samples in the initial data, which comprises 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML). The independent data is an independent collection of 34 leukemia samples including 20 ALL and 14 AML. Expression levels of 7129 genes produced by Affymetrix high-density oligonucleotide microarrays are also included in this data. Code 1 for ALL and 0 for AML. We then use gene expression to classify between ALL and AML.

Summary of significant genes for the classification between ALL and AML are presented in Table 4.3 and Table 4.4, where EBIC is used in the former but deviance is applied in the later. When deviance is employed, we use the initial data for fitting while apply the independent data for testing. As shown in Table 4.3, SLR, M1 and M3 identify the same significant gene with frequency ID 3320. Actually, this gene is also selected by Golub et.al (1999) and Liao and Chin (2007). M2 chooses the gene with ID 6218, which is consistent with the finding of Lee et.al (2003) and Liao and Chin (2007), although this gene is not identified by other three approaches. Table 4.4 shows that SLR and M1 result in a almost identical deviance and model size. At the same time, M3 obtains a model with the largest deviance and the smallest model size while M2 achieves the smallest deviance. This result is slightly different with simulations in the previous subsection. In addition, all these four approaches separately overlap with part of significant genes given by Lee et.al (2003) or Liao and Chin (2007). However, the overlapping between M3 and Lee et.al (2003) is more than that of SLR and M1, although M3 has a smaller model size.

4.3 Interactive Model

4.3.1 Linear Interactive Model

This subsection consists of three parts. The finite sample property of SLR with the application of EBIC is what we are interested in, thus we provide the first part. In this part, the consistency of EBIC is also verified by exploring the impact of (γ_m, γ_I). As mentioned in subsection 3.2.1, we extend SLR from models with only main effects to interactive models mainly through grouping features into main effects and interactive effects and handling features with distinct effects separately. In the second part, we aim at showing the advantage of this grouping treatment through a simple comparison: grouping v.s. non-grouping. The third part investigates a special situation that marginal effects of some predictors are zero but their joint effects are not. Under this situation, the necessity of interactions is showed through the comparison of SLR between the case with interactive effects and the case with only main effects.

4.3.1.1 Finite Sample Properties

The true linear interactive model in this study is assumed to be

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \sum_{j=1}^{p_{0n}-k} x_{ij}x_{i(j+1)}\beta_{k+j}, \ i = 1, 2, ...n, \tag{4.2}$$

where ϵ_i i.i.d $\sim N(0, \sigma^2)$. The proportion of interaction terms is nearly 0.5 by letting $k = [p_{0n}/2] + 1$. Mimic heritability in broad sense and we define a ratio h which is expressed as $h = \frac{\boldsymbol{\beta}^T \Sigma^\star \boldsymbol{\beta}}{\boldsymbol{\beta}^T \Sigma^\star \boldsymbol{\beta} + \sigma^2}$. The variance σ^2 is determined by setting h to certain values when $n = 100$ and kept unchanged for other n. We let $\sigma = 1.5$, 1 in this study such that h is roughly 0.8 and 0.9. The covariates are generated according to structure 1, 5 and 6 with $\rho = 0.4$ while two hundred replicates are done.

The diverging pattern is taken as (n, p_{0n}, p)=$(n, 3[n^{0.345}], [exp(n^{0.325})])$ for $n = 100, 200, 500$. Let $\nu(s_m) = \nu(s_I) = n$. The true coefficient β_j is generated through two ways which are called the original case and the sequential case. In the original case, $\beta_j = n^{-0.150} + |z_j|/10$ for $z_j \sim N(0, 1)$, which ensures $\min\{|\beta_j| : j \in s_{0n}\} = O(n^{-0.150})$. For the sequential case, β_j is generated in the same way as the original case when $n = 100$. When $n = 200$, the first $p_{0n}|_{n=100}$ parameters β_j are kept unchanged whereas the remaining $p_{0n}|_{n=200} - p_{0n}|_{n=100}$ coefficients are generated as $200^{-0.150} + |z_j|/10$. Generate β_j in a similar way when $n = 500$ and we obtain sequential values.

Sample Properties: Results of SLR with $\gamma_{EBIC} = (1 - \frac{\ln n}{2\ln p}, 1 - \frac{\ln n}{4\ln p})$ are presented in Table 4.5. This table shows that: (i) PDR converges to 1 and FDR decreases to 0 rapidly as n increases from 100 to 500, under all three structures, two h levels and two cases. This finding demonstrates that the sample performance of SLR closely matches its asymptotic property. (ii) The sequential case performs better than the original case, which seems to provide clear evidence that larger coefficients contribute to the identification of true features.

Impact of (γ_m, γ_I): BIC is a traditional criterion while EBIC and mBIC are improvements for it. BIC and EBIC differ in the value of (γ_m, γ_I), i.e. $\gamma_{EBIC} = (1 - \frac{\ln n}{2\ln p}, 1 - \frac{\ln n}{4\ln p})$ but $\gamma_{BIC} = (0, 0)$. The mBIC is comparable with EBIC in an asymptotic sense with $\gamma_{as} = (1, 1)$. Under these three different (γ_m, γ_I), the following conclusions can be made from Table 4.5. (i) The PDR of SLR with γ_{BIC} is generally a bit higher since BIC selects much more features. The FDR of SLR with BIC does not reduce as n increases, which suggests that BIC is not selection consistent. (ii) Both EBIC and mBIC are selection consistent because sample properties of SLR with γ_{EBIC} and γ_{as} closely match their asymptotic properties. This finding verifies the effectiveness of Theorem 2.1 in the finite sample case. (iii) Compared with EBIC, mBIC loses certain power while overly controls FDR for small n.

Subsequently, we describe the impact of γ_m and γ_I respectively by assuming five different set of (γ_m, γ_I). The corresponding simulation results are provided in Table 4.6. We may conclude that: (i) γ_I appears to affect PDR and FDR more than γ_m. That's because: if γ_I is fixed, both PDR and FDR decrease when γ_m increases from 0 to $(1-\frac{\ln n}{2\ln p})$; if γ_m is fixed, FDR still decreases but PDR increases when γ_I increases from 0 to $(1-\frac{\ln n}{4\ln p})$. In addition, differences among patterns with the same γ_m and different γ_I seem to be larger than those with the same γ_I but distinct γ_m. (ii) $PDR \to 1$ and $FDR \to 0$ cannot be achieved simultaneously if either γ_m or γ_I is less than the threshold value. This finding demonstrates the effectiveness of the consistency theorem of EBIC again.

4.3.1.2 Comparison: Grouping v.s. Non-Grouping

We extend SLR from models with only main effects to interactive models primarily through grouping features into main effects and interactive effects and s-electing features separately on these two groups. Under interactive models, SLR with the application of grouping, i.e. SLR after the extension, is supposed to perform better than SLR without grouping, i.e. SLR before the extension. The contrast between the case with grouping (m1) and the case without grouping (m2) is investigated through a simple simulation under two frequently used structures, that is, $\rho_{ij} = 0.5^{|i-j|}$ (structure 1) and $\rho_{ij} = 0.5$ (structure 2). We take $(n, p, p_{0n}) = (n, n^{1.2}, 6)$ for n=100 and 200 in this simulation. Besides, we assume the same true model as (4.2) while let each $\epsilon_i \sim N(0,1)$. The true coefficient β_j in this model is given by $0.5 + |z_j|/10$ for $z_j \sim N(0,1)$. The symbol k in this model denotes the number of causal main effects features. PDR and FDR are used to assess this simulation and each PDR(FDR) is over 200 replications.

Under different k, the comparison between m1 and m2 is reported in Table 4.7. As shown in this table, m1 is selection consistent under various k, although m2 appears to perform a little better when k is conveniently close to p_{0n}. However, m2 is unable to achieve a good perform when k is small, that is, relevant interactions accounting for a large proportion, which is completely different with m1. In particular, when $k = 0$, m2 cannot identify any true features whereas m1 can select the vast majority of true features as the sample size increases. In summary, m1 has a more stable and better performance than m2. This finding further demonstrates that our techniques for extension is effective.

4.3.1.3 Special Situation: Main v.s. Main-interactive

The motivation for this study is a conjecture in QTL mapping studies. A QTL study can be regarded as a large-scale feature selection problem due to the existence of QTL with large effects or moderate effects or small effects, as well as interactive effects. In order to reduce the complexity of QTL studies, some researchers, for instance, Wang et.al (2011), focus on selecting markers with high LOD scores first before applying these preselected markers to identify main effects and interactive effects. Nevertheless, it may be difficult to identify these markers with high LOD scores if the marginal effects of them are small or even zero. To investigate properties of SLR under the special situation that marginal effects of some predictors are zero but their joint effects are not, we put forward this simulation and provide a comparison between the case considering only main effects (case A) and the case considering both main effects and interactive effects (case B). PSR and FSR are applied to assess the sample performance in this simulation rather than PDR and FDR.

Under the space with $p = [exp(n^{0.325})]$, we define the true model as

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k + x_{i1}x_{i2}\beta_{k+1} + x_{i3}x_{i4}\beta_{k+2} + \epsilon_i,\ \ i = 1, ..., n,$$

where each $\epsilon_i \sim N(0, 0.5^2)$ and $k = [n^{0.345}] + 1$ for $n = 100, 200, 400$. Let $\nu(s_m) = \nu(s_I) = n$. The covariates are generated in almost the same way as structure 1, 5, 6 except that the mean of X_j is $\mathbf{1}$ instead of $\mathbf{0}$. Subsequently, the true coefficient β_j is given by $\beta_j = -\beta_{k+1} = -\beta_{k+2} = n^{-0.135}$ for $1 \leq j \leq 4$ and $\beta_j = n^{-0.135} + |z_j|/10$ for other j, where $z_j \sim N(0, 1)$. Clearly, marginal effects of the first four true features are zero whereas their unique effects are not.

With the application of EBIC, results of SLR under case A and case B are presented in Table 4.8. The γ value in EBIC is chosen as $1-\frac{\ln n}{4\ln p}$ and $(1-\frac{\ln n}{2\ln p}, 1-\frac{\ln n}{4\ln p})$ separately for case A and B. From Table 4.8, we can conclude that SLR achieves a better performance under case B than under case A. That's because: (i) PSR of case B quickly becomes larger than that of case A when n increases, although this PSR is lower when $n = 100$. (ii) Under case B, FSR falls sharply and it will be less than the FSR of case A when n is sufficiently large. (iii) It is unable to identify any true features with zero marginal effects under case A. Because PSR_{1234}, the probability of selecting the first four true features, is always zero for all n. In contrast, we identify these four features with a satisfactory probability under case B. In summary, case B contributes to the identification of relevant features more than case A. The comparison between these two cases further illustrates that it is indeed imperative to consider interactive models in some practical fields.

4.3.1.4 Real Data Example 2

Under linear interactive models, we illustrate our SLR through a real QTL data set (Bailey et.al, 2008) containing 362 F2 mice and 211 gene markers. These markers imply there are 211 main effect features and 22155 interactive effect features. The corresponding QTL experiment of this data set is carried out to identify loci causing locomotor activation and anxiety. It tests 8 open field measures which may contribute to activation and anxiety disorders, that is, Percent time spent in center of arena, Total distance, Total rearing, Ambulatory episodes, Average velocity, Percent resting, Activity factor and Anxiety factor. We drop individuals that have large than 30 missing values and impute remaining missing values by R package *Imputation*.

Significant and suggestive QTL causing activation and anxiety are presented in Table 4.9. This table gives a model (with repetition) including 12 main effect features and 5 interactive effect features : 5 main effects on chromosome 8; 3 main effects on chromosome 17; 2 main effects and 2 interactive effects on chromosome 2; 1 main effect on chromosome 7; 1 main effect and 1 interactive effect on chromosome 13, 3 interactive effect on chromosome 6 and 12. In this model, a main effect locus on chromosome 8 is most significant because it associates with multiple measures like Total distance, Activity factor. This finding is the same as that of Bailey et.al (2008). The locus on chromosome 17 and the interaction between chromosomes 6 and 12 also play an important role in causing activation and anxiety. In addition, we find that loci on chromosome 2 and 13 are responsible for Percent time in center and Percent resting while the locus on chromosome 7 is suggestive. It is slightly different from Bailey et.al (2008) because they think the predominant interaction between chromosome 13 and 17 accounting for largest portion of behaviors.

4.3.2 Logistic Interactive Model

In this subsection, we simulate SLR with EBIC under the binary logistic interactive model. The diverging pattern is taken as $(n, p, p_{0n}) = (n, [6exp(n^{0.2575})], 2[n^{0.2125}])$. The y_i, $i = 1, 2, ..., n$ in this logistic model is generated from a Bernoulli distribution with probability $p(\theta_i) = exp(\theta_i)/(1 + exp(\theta_i))$, where θ_i is assumed to be $x_{i1}\beta_1 + ... + x_{ik}\beta_k + x_{i1}x_{i2}\beta_{k+1} + +x_{i3}x_{i4}\beta_{k+2} + ... + x_{i(2(p_{0n}-k)-1))}x_{i(2(p_{0n}-k))}\beta_{p_{0n}}$. Three different k, that is, $k_1 = p_{0n} - [0.25p_{0n}]$, $k_2 = [0.5p_{0n}]$ and $k_3 = [0.25p_{0n}]$, are considered in this simulation. The true β_j is given by $2\beta_1 = -\beta_{k+1} = 4n^{-0.175}$; $\beta_j = (-1)^{j+1}4n^{-0.175} + 0.025|r_j|$ for other j, where $r_j \sim binomial(10, 0.5)$. The covariates are firstly generated according to structure 6 and 7 and we then let $X_j = X_j + 0.5$. Thus, the marginal effect of the first true feature becomes zero but its joint effect is not. In the screening step of SLR, a relatively large $\nu(s_m) = \nu(s_I) = p$ is chosen to try to avoid losing any true features. In addition, we consider four different (γ_m, γ_I) in EBIC: $\gamma_{BIC} = (0, 0)$, $\gamma_{MID} = (\frac{1}{2}(1 - \frac{\ln n}{2\ln p}), \frac{1}{2}(1 - \frac{\ln n}{4\ln p}))$, $\gamma_{EBIC} = (1 - \frac{\ln n}{2\ln p}, 1 - \frac{\ln n}{4\ln p})$ and $\gamma_{as} = (1, 1)$.

Performances of SLR with various (γ_m, γ_I) are reported in Table 4.10. This table shows a similar trend as in the linear interactive model: (i) SLR with γ_{BIC} and γ_{MID} generally achieve a slightly higher PDR and a much larger FDR than that with γ_{EBIC}. (ii) SLR with γ_{EBIC} quickly achieves a comparable PDR when

n increases whereas its FDR is more satisfactory. In general, it closely matches its asymptotic property and this finding is consistent with Theorem 2.2. (iii) SLR with γ_{as} is also selection consistent although it appears to over control FDR and lose some power, especially for small n. In addition, Table 4.10 also implies that the proportion between main effects and interactive effects would influence performances of SLR. This implication is reflected in different PDR(FDR) under k_1, k_2 and k_3. Subsequently, we explore this influence in details by investigating discovery rate for main effects and interactive effects respectively. And the corresponding results are presented in Table 4.11. Denote the discovery rate of main effect features by $PDR_m(FDR_m)$. Denote the discovery rate of interactive effect features by $PDR_I(FDR_I)$. As shown in Table 4.11, PDR_m is generally higher than PDR_I under k_1 whereas PDR_I becomes larger than PDR_m under k_3. On the other hand, FDR_m is smaller than FDR_I under k_1 but FDR_I becomes lower than FDR_m under k_3. These findings appear to suggest that the more true main (interactive) effects in a fixed model the easier to identify the corresponding main (interactive) features.

Poisson Log Linear		PDR(FDR)			Model Size		
		n=100	n=200	n=400	n=100	n=200	nn=400
Structure 1	SLR	.928(.300)	.986(.285)	1.000(.194)	11.8	14.1	14.7
	M1	.394(.467)	.418(.472)	.473(.452)	6.3	8.2	9.9
	M2	.393(.470)	.412(.476)	.468(.443)	6.1	8.1	9.4
	M3	.404(.450)	.478(.442)	.537(.441)	7.2	9.5	11.2
Structure 2	SLR	.718(.298)	.778(.286)	.791(.272)	8.1	10.3	12.2
	M1	.658(.324)	.678(.298)	.724(.283)	7.7	8.8	11.4
	M2	.635(.379)	.672(.299)	.722(.284)	7.9	8.7	11.3
	M3	.674(.320)	.688(.296)	.755(.279)	7.9	9.0	11.8
Logistic		PDR(FDR)			Model Size		
		n=100	n=200	n=400	n=100	n=200	nn=400
Structure 3	SLR	.207(.186)	.526(.092)	.980(.044)	2.0	5.2	10.3
	M1	.166(.164)	.356(.076)	.790(.033)	1.6	3.5	8.1
	M2	.167(.164)	.351(.080)	.778(.032)	1.6	3.4	8.0
	M3	.168(.164)	.361(.076)	.791(.033)	1.6	3.5	8.1
Structure 4	SLR	.189(.170)	.527(.080)	.987(.035)	1.8	5.2	10.3
	M1	.148(.165)	.344(.057)	.865(.020)	1.4	3.3	8.8
	M2	.148(.167)	.339(.056)	.850(.018)	1.4	3.2	8.7
	M3	.148(.167)	.348(.058)	.868(.021)	1.4	3.3	8.9

Table 4.1 Models with Only Main Effects: Simulations under Poisson Log Linear Model and Logistic Model with the focus on Selection Consistency

Poisson Log Linear		Model Size			Deviance		
		n=100	n=200	n=400	n=100	n=200	nn=400
Structure 1	SLR	25.0	28.1	31.0	120.80	428.5	1012.6
	M1	23.8	26.8	29.3	197.76	1127.8	5223.8
	M2	25.6	29.2	32.5	257.6	1642.3	7289.6
	M3	31.6	35.6	39.7	57.4	146.2	326.9
Structure 2	SLR	25.0	28.0	31.1	65.18	153.51	332.52
	M1	26.5	32.3	36.7	62.73	148.59	322.68
	M2	25.7	32.4	37.2	108.15	198.45	497.53
	M3	31.6	35.7	39.9	56.06	142.48	317.71
Logistic		Model Size			Deviance		
		n=100	n=200	n=400	n=100	n=200	nn=400
Structure 3	SLR	38.6	35.6	38.0	45.26	130.04	268.53
	M1	28.1	34.5	39.5	62.35	122.39	265.01
	M2	26.7	33.5	39.4	73.92	125.53	264.89
	M3	39.8	45.1	50.2	45.01	101.03	251.28
Structure 4	SLR	38.5	34.6	37.4	39.19	142.50	267.45
	M1	28.1	34.2	39.4	61.31	124.73	264.66
	M2	26.8	33.0	39.2	61.65	128.53	264.90
	M3	39.9	45.1	50.2	58.60	111.01	250.86

Table 4.2 Models with Only Main Effects: Simulations under Poisson Log Linear Model and Logistic Model with the focus on Prediction Accuracy

	Frequency ID	Gene Description
SLR	3320	$U50136.rna1.at$
M1	1745	$M16038.at$
	3320	$U50136.rna1.at$
M2	6218	$M27783.s.at$
M3	3320	$U50136.rna1.at$
	4847	$X95735.at$

Table 4.3 Models with Only Main Effects: Real Data Example 1, Summary of Significant Genes for Classification by Applying EBIC

	Deviance(1e-10)	Model Size	Frequency ID
SLR	2.181	32	50 461 1250 1372 1753 1829 2065 2111 2242 2301 3320 3565 3847 3916 4137 4186 4190 4196 4245 4399 4499 4541 4855 5348 5376 5865 5970 6158 6169 6838 7066 7128
M1	2.183	29	129 230 461 894 1745 1862 2111 2242 2301 2697 3221 3320 3338 3847 3967 4137 4193 4196 4230 4499 5002 5039 5348 5772 5954 6021 6169 6539 6801
M2	2.167	31	312 461 1010 1144 1685 1779 1834 1882 2001 2015 2020 2267 2354 3320 3507 3967 4186 4399 4499 4847 5039 5171 5290 5772 6055 6167 6218 6281 6308 6539 6855
M3	5.607	14	461 1249 1779 1834 1846 2001 2020 3320 3847 4847 5039 5772 5954 6539

Table 4.4 Models with Only Main Effects: Real Data Example 1, Summary of Significant Genes for Classification by Applying Deviance

Original Case			$\sigma = 1$			$\sigma = 1.5$	
	n	γ_{BIC}	γ_{EBIC}	γ_{as}	γ_{BIC}	γ_{EBIC}	γ_{as}
Structure 1	100	.423(.949)	.414(.275)	.321(.099)	.324(.961)	.271(.256)	.238(.093)
	200	.655(.941)	.915(.182)	.825(.039)	.388(.965)	.615(.243)	.466(.061)
	500	.726(.940)	.952(.097)	.948(.021)	.671(.962)	.836(.139)	.577(.024)
Structure 5	100	.637(.923)	.405(.121)	.319(.051)	.485(.941)	.259(.155)	.238(.062)
	200	.880(.920)	.932(.113)	.829(.035)	.590(.947)	.642(.159)	.482(.046)
	500	.911(.918)	.958(.068)	.958(.011)	.764(.938)	.853(.099)	.589(.011)
Structure 6	100	.428(.948)	.411(.229)	.323(.052)	.351(.957)	.263(.244)	.237(.082)
	200	.695(.937)	.914(.155)	.831(.043)	.385(.965)	.612(.223)	.472(.056)
	500	.808(.938)	.958(.082)	.954(.015)	.684(.963)	.843(.140)	.581(.015)
Sequential Case			$\sigma = 1$			$\sigma = 1.5$	
	n	γ_{BIC}	γ_{EBIC}	γ_{as}	γ_{BIC}	γ_{EBIC}	γ_{as}
Structure 1	100	.423(.949)	.414(.275)	.321(.099)	.324(.961)	.271(.256)	.238(.093)
	200	.687(.940)	.926(.156)	.837(.042)	.402(.962)	.686(.231)	.493(.053)
	500	.931(.937)	.958(.081)	.958(.015)	.699(.961)	.842(.105)	.733(.018)
Structure 5	100	.637(.923)	.405(.121)	.319(.051)	.485(.941)	.259(.155)	.238(.062)
	200	.887(.918)	.934(.079)	.843(.028)	.622(.942)	.729(.142)	.505(.031)
	500	.954(.916)	.958(.047)	.958(.007)	.812(.936)	.855(.077)	.802(.005)
Structure 6	100	.428(.948)	.411(.229)	.323(.052)	.351(.957)	.263(.244)	.237(.082)
	200	.712(.936)	.934(.144)	.841(.032)	.411(.960)	.708(.205)	.501(.037)
	500	.934(.935)	.958(.073)	.958(.011)	.731(.961)	.845(.103)	.742(.013)

Table 4.5 Linear Interactive Model: Finite Sample Performance: PDR(FDR), $\gamma_{BIC} = (0,0)$, $\gamma_{EBIC} = (1 - \frac{\ln n}{2 \ln p}, 1 - \frac{\ln n}{4 \ln p})$, $\gamma_{as} = (1,1)$

				PRD(FDR)			
	n	γ_1	γ_2	γ_{EBIC}	γ_3	γ_4	γ_{EBIC}
Structure 1	100	.229(.972)	.336(.924)	.270(.255)	.393(.546)	.321(.403)	.270(.255)
	200	.222(.980)	.498(.945)	.616(.243)	.691(.581)	.676(.340)	.616(.243)
	500	.512(.974)	.723(.932)	.838(.138)	.877(.578)	.856(.289)	.838(.138)
Structure 5	100	.370(.951)	.313(.581)	.261(.156)	.374(.368)	.325(.262)	.261(.156)
	200	.364(.967)	.598(.601)	.642(.157)	.713(.376)	.689(.220)	.642(.157)
	500	.557(.952)	.842(.542)	.853(.090)	.895(.380)	.862(.201)	.853(.090)
Structure 6	100	.254(.969)	.343(.918)	.262(.246)	.376(.504)	.312(.405)	.262(.246)
	200	.231(.979)	.507(.942)	.611(.228)	.702(.536)	.681(.337)	.611(.228)
	500	.534(.967)	.801(.912)	.845(.136)	.882(.535)	.858(.273)	.845(.136)

Table 4.6 Linear Interactive Model: Impact of (γ_m, γ_I), $\sigma = 1.5$. $\gamma_1 = (1 - \frac{\ln n}{2 \ln p}, 0)$; $\gamma_2 = (1 - \frac{\ln n}{2 \ln p}, \frac{1}{2}(1 - \frac{\ln n}{4 \ln p}))$; $\gamma_3 = (0, 1 - \frac{\ln n}{4 \ln p})$; $\gamma_4 = (\frac{1}{2}(1 - \frac{\ln n}{2 \ln p}), 1 - \frac{\ln n}{4 \ln p})$; $\gamma_{EBIC} = (1 - \frac{\ln n}{2 \ln p}, 1 - \frac{\ln n}{4 \ln p})$

		Structure 1: PDR(FDR)		Structure 2: PDR(FDR)	
n	k	$m1$	$m2$	$m1$	$m2$
100	0	.083(.768)	.000(1.000)	.130(.390)	.000(1.000)
	1	.655(.457)	.167(.207)	.739(.146)	.167(.094)
	2	.333(.416)	.005(.970)	.326(.156)	.081(.515)
	3	.230(.364)	.168(.196)	.209(.152)	.168(.058)
	4	.748(.407)	.500(.145)	.782(.122)	.500(.081)
	5	.520(.285)	.512(.198)	.507(.080)	.501(.063)
	6	.988(.274)	.995(.142)	.997(.082)	.999(.064)
200	0	.963(.179)	.000(1.000)	.988(.043)	.000(1.000)
	1	.937(.259)	.127(.313)	.992(.055)	.158(.086)
	2	.973(.252)	.333(.164)	.983(.061)	.333(.056)
	3	1.000(.220)	.345(.124)	1.000(.034)	.338(.041)
	4	1.000(.221)	.658(.135)	1.000(.056)	.665(.044)
	5	.848(.187)	.833(.142)	.839(.050)	.835(.047)
	6	1.000(.185)	1.000(.124)	1.000(.046)	1.000(.045)

Table 4.7 Linear Interactive Model: Comparison: Grouping v.s. Non-Grouping

Structure 1			PSR(FSR)			PSR_{1234}	
ρ		n=100	n=200	n=400	n=100	n=200	n=400
.5	main	.200(.168)	.429(.136)	.500(.134)	.000	.000	.000
	main-interactive	.008(.991)	.565(.439)	1.000(.099)	.015	.450	1.000
.2	main	.200(.165)	.429(.151)	.500(.140)	.000	.000	.000
	main-interactive	.002(.997)	.508(.480)	1.000(.125)	.006	.388	1.000
Structure 5			PSR(FSR)			PSR_{1234}	
ρ		n=100	n=200	n=400	n=100	n=200	n=400
.5	main	.200(.097)	.429(.071)	.500(.062)	.000	.000	.000
	main-interactive	.041(.912)	.687(.288)	1.000(.053)	.055	.595	1.000
.2	main	.200(.141)	.429(.116)	.500(.113)	.000	.000	.000
	main-interactive	.002(.997)	.572(.416)	1.000(.095)	.009	.455	1.000
Structure 6			PSR(FSR)			PSR_{1234}	
ρ		n=100	n=200	n=400	n=100	n=200	n=400
.5	main	.200(.191)	.429(.139)	.500(.131)	.000	.000	.000
	main-interactive	.002(.999)	.494(.472)	1.000(.114)	.006	.353	1.000
.2	main	.200(.175)	.429(.139)	.500(.132)	.000	.000	.000
	main-interactive	.004(.993)	.491(.481)	.999(.132)	.010	.345	1.000

Table 4.8 Linear Interactive Model: Special Situation: Main v.s. Main-Interactive

	Feature ID	Chr	Location(Mb)	Effect	Interaction
Percent time in center	163	13	89.444	main	
	6559	2	178.315	interactive	Chr13:22.251
		13	22.251	interactive	Chr2:178.315
Total Distance	96	8	57.724	main	
	193	17	56.801	main	
	13116	6	102.455	interactive	Chr12:2.058
		12	2.058	interactive	Chr6:102.445
Total Rearing	30	2	153.094	main	
Ambulatory Episodes	98	8	68.129	main	
	193	17	56.801	main	
	13116	6	102.455	interactive	Chr12:2.058
		12	2.058	interactive	Chr6:102.455
Average Velocity	101	8	89.447	main	
Percent Resting	23	2	97.379	main	
	85	7	63.356	main	
	101	8	89.447	main	
Activity factor	96	8	57.724	main	
	193	17	56.801	main	
	13116	6	102.455	interactive	Chr12:2.058
		12	2.058	interactive	Chr6:102.455
Anxiety factor	6534	2	178.315	interactive	Chr11:68.383
		11	68.383	interactive	Chr2:178.315

Table 4.9 Linear Interactive Model: Real Data Example 2, Summary of Suggestive and Significant QTL

		Structure 6: PDR(FDR)			Structure 7: PDR(FDR)		
k_1	γ	n=100	n=200	n=500	n=100	n=200	nn=500
	γ_{BIC}	.395(.801)	.562(.746)	.737(.716)	.495(.747)	.630(.709)	.795(.706)
	γ_{MID}	.470(.622)	.637(.556)	.813(.548)	.527(.479)	.653(.454)	.817(.427)
	γ_{EBIC}	.460(.247)	.630(.236)	.787(.159)	.515(.158)	.637(.149)	.803(.136)
	γ_{as}	.435(.160)	.593(.143)	.777(.086)	.483(.127)	.620(.115)	.796(.098)
k_2	γ	n=100	n=200	n=500	n=100	n=200	nn=500
	γ_{BIC}	.340(.794)	.590(.716)	.776(.711)	.475(.713)	.655(.677)	.801(.672)
	γ_{MID}	.330(.491)	.607(.467)	.784(.442)	.473(.354)	.650(.426)	.820(.481)
	γ_{EBIC}	.280(.168)	.560(.153)	.782(.128)	.378(.134)	.610(.128)	.807(.113)
	γ_{as}	.255(.053)	.497(.036)	.769(.025)	.312(.109)	.585(.052)	.793(.047)
k_3	γ	n=100	n=200	n=500	n=100	n=200	nn=500
	γ_{BIC}	.468(.763)	.641(.712)	.774(.692)	.568(.714)	.783(.598)	.835(.755)
	γ_{MID}	.465(.662)	.636(.552)	.768(.503)	.580(.550)	.785(.400)	.835(.539)
	γ_{EBIC}	.385(.302)	.621(.228)	.767(.172)	.443(.260)	.733(.131)	.835(.125)
	γ_{as}	.313(.155)	.576(.133)	.762(.074)	.343(.119)	.585(.072)	.833(.065)

Table 4.10 Logistic Interactive Model: Performances under Different Interactions, $\gamma_{BIC} = (0,0)$, $\gamma_{MID} = (\frac{1}{2}(1-\frac{\ln n}{2\ln p}), \frac{1}{2}(1-\frac{\ln n}{4\ln p}))$, $\gamma_{EBIC} = (1-\frac{\ln n}{2\ln p}, 1-\frac{\ln n}{4\ln p})$, $\gamma_{as} = (1,1)$, $k_1 = p_{0n} - [0.25p_{0n}]$, $k_2 = [0.5p_{0n}]$, $k_3 = [0.25p_{0n}]$

Structure 6		$PDR_m(FDR_m)$			$PDR_I(FDR_I)$		
k	γ	n=100	n=200	n=500	n=100	n=200	nn=500
k_1	γ_{BIC}	.473(.518)	.574(.368)	.683(.358)	.160(.972)	.500(.935)	.928(.921)
	γ_{MID}	.600(.415)	.678(.412)	.792(.413)	.080(.877)	.430(.768)	.920(.713)
	γ_{EBIC}	.600(.189)	.690(.242)	.784(.177)	.040(.233)	.330(.073)	.800(.020)
	γ_{as}	.567(.144)	.667(.140)	.780(.101)	.040(.026)	.230(.005)	.760(.000)
k_3	γ_{BIC}	.090(.718)	.083(.729)	.078(.742)	.593(.732)	.788(.602)	.912(.501)
	γ_{MID}	.160(.676)	.084(.627)	.073(.639)	.567(.602)	.774(.432)	.908(.307)
	γ_{EBIC}	.160(.225)	.082(.246)	.070(.567)	.460(.246)	.767(.131)	.907(.048)
	γ_{as}	.090(.020)	.066(.143)	.053(.325)	.387(.148)	.727(.092)	.903(.009)
Structure 7		$PDR_m(FDR_m)$			$PDR_I(FDR_I)$		
k	γ	n=100	n=200	n=500	n=100	n=200	nn=500
k_1	γ_{BIC}	.570(.432)	.676(.386)	.778(.391)	.270(.938)	.400(.936)	.922(.911)
	γ_{MID}	.637(.312)	.714(.343)	.788(.368)	.200(.725)	.350(.650)	.960(.452)
	γ_{EBIC}	.650(.130)	.718(.159)	.784(.152)	.110(.105)	.230(.050)	.900(.012)
	γ_{as}	.623(.118)	.710(.103)	.784(.115)	.060(.065)	.170(.030)	.860(.002)
k_3	γ_{BIC}	.060(.678)	.010(.798)	.010(.910)	.737(.674)	.944(.538)	1.000(.728)
	γ_{MID}	.140(.558)	.040(.643)	.010(.848)	.727(.467)	.934(.290)	1.000(.444)
	γ_{EBIC}	.190(.252)	.060(.375)	.010(.615)	.527(.173)	.868(.058)	1.000(.037)
	γ_{as}	.150(.030)	.060(.180)	.000(.368)	.407(.113)	.690(.029)	1.000(.013)

Table 4.11 Logistic Interactive Model: Discovery Rate: Main v.s. Interactive, $\gamma_{BIC} = (0,0)$, $\gamma_{MID} = (\frac{1}{2}(1-\frac{\ln n}{2\ln p}), \frac{1}{2}(1-\frac{\ln n}{4\ln p}))$, $\gamma_{EBIC} = (1-\frac{\ln n}{2\ln p}, 1-\frac{\ln n}{4\ln p})$, $\gamma_{as} = (1,1)$, $k_1 = p_{0n} - [0.25p_{0n}]$, $k_3 = [0.25p_{0n}]$

CHAPTER 5

Conclusion and Future Research

5.1 Conclusion

In contemporary statistics, one of the most popular topics is high dimensional feature selection, in which both LMs and other GLMs play a major role. Among high dimensional feature selection studies, a large number considered the main effect features only while only a few considered the interactive effects, although interactions were also prominent in explaining the response variable. In our thesis, we aimed at proposing feasible feature selection procedures in the space including both main effects and interactive effects, with the emphasis on achieving selection consistency. These selection procedures may result in a great improvement in high dimensional feature selection process for both LMs and GLMs.

As mentioned in chapter 1, an efficient feature selection procedure usually consists of two important steps: a suitable feature selection method and an appropriate model selection criterion, where the former is designed to generate candidate models and the later aims at identifying the best model from these candidate models. Among model selection criteria, EBIC (Chen and Chen, 2008) is a desirable choice for high dimensional feature selection because it can effectively limit the false discovery rate while it suffers slightly lower positive discovery rate than the classic BIC (Schwarz, 1978). Nevertheless, the selection consistency of EBIC is not demonstrated when interactive effects are taken into consideration.

In chapter 2, we established the selection consistency of EBIC under high feature space through acceptable conditions by considering both main effects and pairwise interactive effects in LMs and GLMs. One advantage of our study is that we allow a diverging number of relevant features rather than a fixed number. Our subsequent simulations in chapter 4 showed that EBIC with a proper (γ_m, γ_I) is effective in high dimension model selection. One possible limitation of our study is that we did not consider the high order interaction due to its rarity and complexity.

In chapter 3, with the application of EBIC, we developed feature selection procedures under two kind of models: models with only main effects and interactive models. Selection procedures can be roughly classified into two categories: sequential procedures and penalized likelihood methods. Among these categories, penalized methods are more popular. Thus, under models with only main effects, we firstly reviewed SLasso (Luo and Chen, 2013b), a powerful partial penalized procedure for high dimensional linear regression. Analogous to SLasso that selected the feature maximizing the profile marginal score function, we proposed a novel procedure SLR for high dimension feature selection in GLMs. In this SLR, the application of EBIC was mainly reflected in two aspects: being the stopping rule and being the criterion to identify the optimal model from candidate models. Under reasonable conditions, SLR was shown to be selection consistent under the canonical link. Subsequently, we extended SLR to interactive models by grouping features into main effects and interactive effects and selecting features separately on these two groups. This extension had a key advantage in that it achieved a relatively stable performance under different number of interactions.

In chapter 4, we conducted extensive numerical studies to verify finite sample properties of SLR under different types of models. The sample performance of SLR was mainly assessed by positive discovery rate (PDR) and false discovery rate (FDR). With a proper (γ_m, γ_I), SLR was shown to closely match its asymptotic property, that is, PDR and FDR converged to 1 and 0 respectively when the number of observations n was sufficiently large. In contrast, SLR with $(\gamma_m, \gamma_I) = (0, 0)$, i.e. the traditional BIC (Schwarz, 1978), did not appear to be selection consistent due to the existence of many spurious features.

5.2 Future Research

In this section, we would like to state several interesting directions for future works related to this thesis.

In chapter 2, we established the selection consistency of EBIC under GLIMs with the canonical link. However, the canonical link does not always provide the best fit and a non-canonical link is more preferable in some situations (McCullagh and Nelder, 1989). In addition, in SLR of chapter 3, the computing algorithm we proposed was applicable for both the canonical link and the non-canonical link whereas SLR was only shown to be selection consistent under the former. Thus, a direct extension of our study is to conduct feature selection under the non-canonical link.

Our main purpose in this thesis was to develop a powerful feature selection procedure, especially for QTL mapping, under the small n large p situation. As mentioned in chapter 2, the model selection criterion EBIC relies on the value of (γ_m, γ_I). A larger (γ_m, γ_I) results in a lower PDR and FDR although the corresponding EBIC is still selection consistent. However, these distinct consistent (γ_m, γ_I) may produce completely different outcomes in some real datesets. Thus, future work should involve a method for choosing an appropriate (γ_m, γ_I) in QTL datasets.

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory*, *Akademiai Kiado*, 267-281.

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

[3] Bailey et.al (2008). Identification of QTL for locomotor activation and anxiety using closely-related inbred strains. *Genes Brain Behav*, **7**(7), 761-9.

[4] Blum, A. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**(1-2), 245-271.

[5] Bogdan et.al (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, **167**, 989-99.

[6] Cai, T. and Wang, L. (2011). Orthogonal Matching Pursuit for Sparse Signal Recovery with Noise. *IEEE. Trans. Inf. Theory.*, **57**, 4680-4688.

[7] Chen, J. and Chen, Z. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Space. *Biometrika*, **95**, 759-771.

[8] Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM, *Statistics Sinica*, **22**(2), 555-574.

[9] Chen, S., Donoho, D. and Saunders, M. (2001). Atomic Decomposition by Basis Pursuit. *Society for Industrial and Applied Mathematics*, **43**, 129-159

[10] Chen, Z. and Chen, J. (2009). Tournament screening cum EBIC for feature selection with high-dimensional feature spaces. *Science in China Series A: Mathematics.* **52**(6), 1327-1341

[11] Chen, Z. and Cui, W. (2010). A two-phase procedure for QTL mapping with regression models. *Theoretical and Applied Genetics.* **21**, 363-372

[12] Claudia (2010). Locating multiple interacting quantitative trait Loci with the zero-inflated generalized poisson regression. *Statistical applications in genetics and molecular biology*, **9**(1), Article26

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

[14] Draper, N and Smith, H (1998). Applied Regression Analysis, *Wiley: Wiley series in probability and statistics. Texts and references section*

[15] Efron et.al (2004). Least angle regression. *Ann.Statist.*, **32**, 407-499.

[16] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Am. Statist. Assoc.*, **96**, 1348-1360.

[17] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B.*, **70**(5), 849-911

[18] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101-148

[19] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, **32**, 928-961.

[20] Fan, J. and Song, R. (2010). Sure Independence Screening in Generalized Linear Models with NP-dimensionality. *Ann. Stat.*, **38**, 3567-360

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

[22] Forgel, R. and Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. *arXiv:1011.6640v1 [math.ST]*

[23] Frank, I. and Friedman, J. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-148.

[24] Golub et.al (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-536

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

[26] Huang, J., Ma, S. and Zhang, C. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica.* **18**, 1603-1618

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

[28] Kim et.al (2008). Smoothly Clipped Absolute Deviation on High Dimensions. *J. Am. Statist. Assoc.*, **103**, 1665-1673.

[29] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Ann.Statist.*, **28**, 1356-1378.

[30] Kohavi, R. and John, G. (1997). Wrappers for feature selection. *Artificial Intelligence*, **97**(1-2), 273-324.

[31] Lee et.al (2003). Gene selection: a Bayesian variable selection approach. *bioinformatics*, **19**(1), 90-97

[32] Liao, J. and Chin, K. (2007). Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *bioinformatics*, **23**(15), 1945-1951

[33] Luo, S. and Chen, Z. (2013a). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, **143**, 494-504.

[34] Luo, S. and Chen, Z. (2013b). Sequential Lasso for feature selection with ultra-high dimensional feature space. *Journal of the American Statistical Association*, Minor Revision Invited after second-round review

[35] Luo, S. and Chen, Z. (2013c). Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters. *Statistics and Its Interface*, to appear

[36] Luo, S. and Chen, Z. (2013d). Extended BIC in the Cox model with high-dimensional feature spaceds. Manuscript

[37] Mallows, C. (1973). Some comments on CP. *Technometrics*, **15**, 661-675.

[38] Mccullagh, P. and Nelder, J. (1989). Generalized linear models. Second Edition. *Chapman and Hall, London.*

[39] Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of statistics.* **34**, 1436-1462.

[40] Osborne et.al (2000). On the Lasso and its dual. *J. Comput. Graph. Stat.*, **9**, 319-337.

[41] Park, M. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Statist. Soc. Ser. B.*, **69**, 659-677.

[42] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist*, **6**, 461-464.

[43] Siegmund, D. (2004). Model selection in irregular problems: Application to mapping quantitative trait loci. *Biometrika* , **91**, 785-800.

[44] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B*, **39**, 111-147.

[45] Storey et.al (2005). Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol*, **3**, 267

[46] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *J.R. Statist. Soc. Ser. B.*, **58**, 267-288.

[47] Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using L_1 constrained quadratic programming(Lasso). *IEEE Trans. Inf. Theory.*, **55**, 2183-2202.

[48] Wang et.al (2011). A Model Selection Approach for Expression Quantitative Trait Loci (eQTL) Mapping. *Genetics*, **187**, 611-621

[49] Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist*, **38**(2), 894-942.

[50] Zhang, C. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist*, **36**(4), 1567-1594.

[51] Zhao, J. and Chen, Z. (2012). A Two-Stage Penalized Logistic Regression Approach to Case-Control Genome-Wide Association Studies. *Journal of Probability and Statistics*, Volume 2012 (2012), Article ID 642403, 15 pages

[52] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research.* **7**, 2541-2563.

[53] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist Assoc.*, **101**, 1418-1429.

[54] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B.*, **67**, 301-320.

[55] Zou, H. and Li, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *The Annals of Statistics*, **36**, 1509-1533

[56] Zou, H. and Zhang, H. (2009). On The Adaptive Elastic-Net With A Diverging Number of Parameters. *Ann. Statist.*, **37(4)**, 1733-1751.

[57] Zou, W. and Zeng, Z. (2009). Multiple Interval Mapping for Gene Expression QTL Analysis. *Genetica.* **137**(2), 125-34.