FEATURE SELECTION IN HIGH-DIMENSIONAL STUDIES

LUO SHAN

NATIONAL UNIVERSITY OF SINGAPORE

2012

FEATURE SELECTION IN HIGH-DIMENSIONAL STUDIES

LUO SHAN

(Master of Science, Peking University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2012

Acknowledgements

I am so grateful that I have this opportunity to express my sincere thanks to my teachers, friends and family members before presenting my thesis, which will be impossible without their faithful support.

I would like to express my first and foremost appreciation to my supervisor, Professor Chen Zehua, for his patient guidance, consistent support and encouragement. The regular discussions we ever had will be an eternal treasure in my future career. Professor Chen's invaluable advices, ideas and comments were motivational and inspirational. What I have learned from him is not only confined to research, but also in cultivating healthy personal characteristics.

I am also particularly indebted to another two important persons in my PhD life, Professor Bai Zhidong and Professor Louis Chen Hsiao Yun, for their help and

encouragement. Professor Bai's recognition and recommendation have brought me the chance to be a student in NUS. His unexpected questions in classes have propelled me to expand my knowledge area consistently. The habit I formed since then benefits me a lot. Professor Louis Chen's enthusiasm in teaching, doing research and amiable disposition in daily life have made my acclimation in Singapore much easier. Consciously and unconsciously, the personalities of these two famous scholars have influenced me significantly.

I also would like to thank the other staff members in our department. Illuminations from the young and talented professors whose offices are located at Level Six have occupied an important proportion in my life. Their conscientious, modesty and devotion to academic have always been good examples for me. Thanks to Mr Zhang Rong, Ms Chow Peck Ha, Yvonne for their IT technical help and attentive cares.

Thanks to my dear friends, Mr Jiang Binyan, Mr Liu Xuefeng, Mr Fang Xiao, Mr Jiang Xiaojun, Mr Liu Cheng, Ms Li Hua, Ms Zhang Rongli, Ms He Yawei, Ms Jiang Qian, Ms Fan Qiao, etc. Thanks for their accompany, which has made my life here enjoyable for most of the time.

Finally, I would like to thank my parents, my parents-in-law, my husband, my brothers and sisters, for loving me and understanding me all the time. Thanks to my lovely niece and nephew, for bringing endless happiness into this family.

Table of Contents

Summary

This thesis comprises two topics: the selection consistency of the extended Bayesian Information Criteria (EBIC) and the sequential LASSO procedure in feature selection under small-n-large-p situation in high-dimensional studies.

In the first part of this thesis, we expand the current study of the EBIC to more flexible models. We investigate the properties of EBIC for linear regression models with diverging number of parameters, generalized linear regression models with non-canonical links as well as Cox's proportional hazards model. The conditions under which the EBIC remains selection consistent are established and extensive numerical study results are provided.

In the second part of this thesis, we propose a new stepwise selection procedure,

sequential LASSO, to conduct feature selection in ultra-high dimensional feature space. The conditions for its selection consistency and sure screening property are explored. The comparison between sequential LASSO and its competitors is provided from both theoretical and computational aspects. Our results show that sequential LASSO could be a potentially promising feature selection procedure when the dimension of the feature space is ultra-high.

List of Notations

n	the number of independent observations
p_n	the dimension of the full feature space
X_n	the $n \times p_n$ design matrix with entries $\{x_{i,j}\}$
$\boldsymbol{y}_n$	the n-dimensional response vector
$\boldsymbol{\mu}_n$	the conditional expectation of $\boldsymbol{y}_n$ given $\boldsymbol{X}_n$
$\boldsymbol{\epsilon}_n$	the n-dimensional error vector
$\boldsymbol{\beta}_0$	the p_n-dimensional true coefficient vector in the linear regression system
s_{0n}	the index set of all non-zero coefficients in $\boldsymbol{\beta}_0$
p_{0n}	the cardinality of s_{0n}

$X(s)$	the sub-matrix of X_n with columns whose indices are contained in any arbitrary subset s of $\{1, 2, \ldots, p_n\}$
$\boldsymbol{I}$	the identity matrix with order n
$H_0(s)$	the projection matrix $X(s)\left(X^\tau(s)X(s)\right)^{-1}X^\tau(s)$ if it exists
$\boldsymbol{\beta}(s)$	the sub-vector of $\boldsymbol{\beta}$ with subscripts contained in s
$\lvert s\rvert$	the cardinality of s
$\lambda_{\min}(\cdot)$	the smallest eigenvalue of a square matrix
$\lambda_{\max}(\cdot)$	the largest eigenvalue of a square matrix
O	$f(n) = O(g(n))$ if there exist positive integer M and constant $C > 0$ such that $\dfrac{\lvert f(n)\rvert}{g(n)} < C$ for all $n > M$
o	$f(n) = o(g(n))$ if $\lim\limits_{n\to+\infty} \dfrac{\lvert f(n)\rvert}{g(n)} = 0$
$\lVert\boldsymbol{x}\rVert_2$	$\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$ for $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$
$\lVert\boldsymbol{x}\rVert_1$	$\lvert x_1\rvert + \lvert x_2\rvert + \cdots + \lvert x_n\rvert$ for $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$
$\lVert\boldsymbol{x}\rVert_{+\infty}$	$\max\{\lvert x_1\rvert, \lvert x_2\rvert, \cdots, \lvert x_n\rvert\}$ for $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$

List of Tables

CHAPTER 1

Introduction

In this chapter, we give an introduction to feature selection, provide a brief literature review and sketch the outline of this thesis. The introduction is given in Section 1.1. The literature review is given in Section 1.2. The objectives and organization of the thesis are outlined in Section 1.3.

1.1 Introduction to Feature Selection

Feature Selection, which is also known as variable selection, sparsity or support recovery, is a fundamental topic in both classical and modern statistical inference with applications to diverse research areas such as quantitative trait loci (QTL) mapping and genome wide association studies (GWAS). It aims to recruit the causal or relevant features ([102]) from the suspected feature space into a regression model to describe the relationship between an outcome of interest and the predictors. Because not all these predictors considered initially have important influence on the outcome in reality, statistical inference based on a full regression model is inherently unstable and not advised. By conducting a judicious feature selection, the three-fold objectives can be achieved: an improved prediction performance, more cost-effective predictors, and a better understanding of the underlying process that generated the data ([82],[83]). The selection consistency defined in [183] and prediction accuracy are two goals of feature selection. Under the assumptions where the dimension of the candidate feature space p is fixed and the sample size n is large enough, these two goals could be achieved simultaneously and effectively via criteria such as Akaike's Information Criterion (AIC) ([1]) and its variants Consistent AIC (CAIC), Consistent AIC with Fisher-Information

(CAICF) ([17]), Mallow's C_p ([120]), Cross-Validation (CV) ([154]), the Bayes Information Criterion (BIC) ([144]) and Generalized Cross-Validation (GCV) ([46]). However, under the small-n-large-p situation in high-dimensional studies, where p is much larger than n, the occurrence of over-fitting makes it necessary to address the two goals from a different point of view and to reinvestigate the feasibility of these criteria.

Recently, we have been buried in enormous amount of data from various fields such as biotechnology, finance and astronomy because of the expeditious development in information technology industry. For instance, in GWAS, it has become routine to genotype hundreds of thousands single-nucleotide polymorphism (SNP) markers ([42]). The proliferation of high-dimensional data necessitates the re-examination of conventional statistical methods because of the violation of their assumptions and the appearance of novel objectives of statistical analysis ([49]). Among these issues, feature selection has drawn much attention from statisticians.

Under the small-n-large-p situation in high-dimensional studies, the selection consistency of feature selection becomes more important and needs more attention than high prediction accuracy because it is essential to extract the useful information considering the noise accumulation and interpretation of the model. Moreover, the significance of the selection consistency in pragmatic applications scattered in different disciplines. In QTL mapping, compared with the true QTLs, markers

which are highly linked to them may have the same or even higher prediction ability, but they are less favorable in the model because of the lack of biological interpretation ([22]). In industry, the most influential and vital variables on the quality of a final product are more concerned by process engineers ([39]). In modern systems biology, it is important to connect gene expression data with clinical studies to detect the associated genes for certain disease or life-span of a species from the whole genome ([13],[43]).

It is important to mention that, in feature selection under the small-n-large-p situation in high-dimensional studies, an assumption associated with feature selection in high-dimensional studies is "sparsity" , which refers to the phenomenon that among those suspicious predictors, only a few of them are causal or relevant features. Prior information provided by biologists showed that disease related genes occupy only a small proportion of the genome. For humans, of the approximately 25,000 protein-coding genes, 2,418 are possibly associated with specific diseases ([7]). An accurate detection of possible associated genes inferred from current data-throughout will benefit the further validation experiments performed in labs.

With the appearance of high or ultra-high feature space, where p or $\ln p$ has a polynomial order of n, the model selection criteria such as C_p, AIC, CV, BIC, GCV are no longer suitable for feature selection due to the consequent challenges such as high spurious correlation and "sparsity". C_p, CV and GCV focus on prediction

accuracy, they were shown to have the asymptotic optimality in the sense that the average mean square error tends to its infimum in probability ([113]). AIC and BIC aim to obtain a model to best approximate the true model based on Kullback-Leibler divergence and Bayesian posterior probability respectively, the importance of a tradeoff between prediction accuracy and complexity of the model has been reflected in these criteria, but applications in high-dimensional studies showed that AIC and BIC tended to select far more features than the true relevant ones (See [22],[15],[151]).

In high-dimensional studies, statisticians have made great efforts to develop new techniques to diminish the impact of high spurious correlation to maintain the important information in feature selection. Correspondingly, they have also set up standards to evaluate these techniques. Aside from computational feasibility, the commonly desired characteristics include the oracle property defined in [58], selection consistency and sure screening property defined in [61]. These properties function at different stages of a complete feature selection process.

For a complete feature selection process, a natural direction in the first place is to release the computation burden efficiently through dimension reduction without losing important information. Stepwise or greedy searching algorithms such as Sure Independence Screening (SIS) and Iterative SIS (ISIS) ([61]), Forward Stepwise Regression (FSR, [54]), Orthogonal Matching Pursuit (OMP) algorithm ([159])

are commonly applied to vastly reduce the high or ultra-high dimensional feature space to a lower-dimensional space. However, this lower-dimensional space still has a much larger dimension than expected (see Theorem 1 in [166], Theorem 4.1 in [97], etc.), which requires further feature selection. The sheer number of all possible models remains huge, we can not proceed to select from them directly by all subsets selection methods because of computational intractability of such undertaking. As formally proved and presented in [93], such a subset selection is NP-hard. Feasible alternatives are penalized likelihood methods, which stem from the idea of regularization ([14]). Examples include the Least Absolute Shrinkage and Selection Operator (LASSO) ([156]), the Smoothly Clipped Absolute Deviation (SCAD) ([58]) and the adaptive LASSO ([185]), etc. Given a range of tuning parameters, they can discard the noncontributory models and thus produce a series of much less candidate models than the total number of all possible models in the solution paths. Unavoidably, they require an appropriate choice of the tuning parameters to pinpoint the best model among these sub-models.

Therefore, in high-dimensional studies, an efficient feature selection procedure usually consists of two stages: a screening stage and a selection stage, where the second stage involves a penalized likelihood feature selection procedure and a final selection criterion. Such a two-stage idea has been applied in [61], [168], [34], [166], [182], [106]. To guarantee the overall selection consistency, the sure screening

property for the procedure at the first stage, the oracle property for the penalized technique and the selection consistency for the final selection criterion at the second stage should be assured.

Apart from this two-stage selection, papers [24], [23], [167], [32] focused on conducting feature selection under the Bayesian decision theory framework. Bayesian averaging where a number of distinct models and more predictors are involved was proposed in [25]. In high-dimensional studies, the full Bayes (FB) is too flexible in selecting prior distributions and the empirical Bayes (EB) is preferable to FB in practice. Instead of setting hyper-prior parametric distributions on those parameters in the prior distributions in FB, EB users estimate the parameters from auxiliary data directly. Unfortunately, there are too many challenges involved in implementing Bayesian model choice. It was shown in [41] and [145] that there is a surprising asymptotic discrepancy between FB and EB. Resampling has also been used in feature selection, such as [76]. The most promising subset of predictors is identified as those with the highest visited probability for the samples.

1.2 Literature Review

1.2.1 Feature Selection in Linear Regression Models

Ever since feature selection associated concepts and methods were introduced in [87], researchers have made significant strides in developing efficient methods for feature selection and especially in high-dimensional situations lately. Most of these methods were initially developed based on observations from linear regression models (LMs), where the error term is usually assumed to be Gaussian.

At the screening stage, the usage of greedy algorithms proposed in [8] is appealing for their ability in dimension reduction and is appreciated if sure screening property can be guaranteed. Namely, as the sample size goes to infinity, with probability tending to 1, the procedure can successfully retain all the important features. One famous and simple method is based on marginal effects of the predictors. SIS and ISIS screen important features according to their marginal correlation ranking in LMs. They were proved to own sure screening property under mild conditions. The second popular family is the sequential or stepwise feature selection. It was shown in [166] that for LMs, Forward Selection ("Forward Stepwise Regression (FSR)" in [54]) has sure screening property when the dimension of feature space is ultra-high and the magnitudes of the effects are allowed to depend on the sample

size. Other screening procedures include OMP ([159], [30]) etc. They can be easily implemented, but these reduced models still have sizes much bigger than expected (see Theorem 1 in [166] and Theorem 3 in [97]). As pointed out in [10], [124], stepwise procedures or a single-inference procedure may lead to greatly inflated type I error, or equivalently, a huge proportion of unimportant features will be erroneously selected. Furthermore, if the size of the reduced model is too small, SIS will miss the true predictor which is marginally independent but jointly dependent of the responses. This disadvantage can be alleviated but not be eliminated by ISIS or OMP. Forward Selection pursues the minimal prediction error in each step and thus requires a cautious consideration in high-dimensional situations owing to high spurious correlation.

The penalized likelihood techniques at the second stage are formulated by adding a penalty function coupled with a tuning parameter to the likelihood function([118]), they are lauded for computational efficiency and stability. Covariates with "effects" lower than a data-driven threshold are excluded from the model for a given tuning parameter. The underlying idea is to shrink the smaller "effects" which are believed to be probably caused by noise to zero through the penalty function. Along the solution path produced by adjusting the tuning parameter, what matters for the procedure is the oracle property, meaning that the model with exactly the true important features is among the sub-models with

probability tending to 1 as the sample size n increases to infinity.

Among these penalized likelihood feature selection procedures, the LASSO was most frequently employed for its efficient computation. A relatively comprehensive study has been done on LASSO. Conditions for the existence, uniqueness and number of non-zero coefficients of the LASSO estimator were detected in [127]; the general path-following algorithm ([138]) and stagewise LASSO ([184]) were proposed to approximate the LASSO paths; the consistency and limiting distributions of the LASSO-type estimators were investigated in [109]. Although being a leading approach in feature selection, the drawback of LASSO lies in the conditions required for its oracle property, which is described as Irrepresentable Condition in [183] or Mutual Incoherence Conditions in [165] or Neighborhood Stability in [122]. It essentially requires that the uncausal features should be weakly correlated with the true causal features. Considering the incomparably large cardinality of uncausal features, this condition is too strong to be satisfied. Although it was shown in [123] that when the irrepresentable condition is violated in the presence of highly correlated variables, the LASSO estimator is still consistent in the L_2 norm sense. Given the focus of feature selection, more work need to be done on LASSO.

Inspired by the spirit of LASSO, its extensions or modified versions arose quickly. The elastic net proposed in [187] encourages a grouping effect where

strongly correlated predictors tend to be in or out of the model together. It encompasses LASSO as a special case and its oracle property was examined in [101]. It was verified that the oracle property entails similar constraints on the design matrix as LASSO. Adaptive LASSO was proposed in [185] for fixed p and its extension to small-n-large-p situation was finished by [92]. The adaptive irrepresentable condition was given for its oracle property. The adaptive elastic-net proposed in [189] has oracle property when $0 \leq \ln p / \ln n < 1$ under weak regular conditions. The SCAD can result in sparse, unbiased and continuous solutions under mild conditions, but it has computation issues because of the optimizations involving non-convex objective functions. An efficient fast algorithm was developed in [107] to implement SCAD when $p \gg n$. For other techniques, it was found in [53] that the Least Angle Regression (LARs) and the forward stagewise regression were closely related with LASSO in the sense that their resulting graphs are similar given connected true parameters and they have identical solution paths for certain designed matrices. LARs and its variants were further examined in [85], [86], [133]. The paper [98] shed light on how the LASSO and Dantzig selector proposed in [31] are related. We can refer to [62] for more details about other recently developed approaches such as non-negative garrott estimator proposed in [177].

Despite these encouraging results, it is important to note that, the oracle property of most of these procedures hinges on the choice of tuning parameter. In

practice, the tuning parameter is always chosen by a separately given criterion, such as cross validation, generalized cross validation, etc. However, whether this selected parameter satisfies the assumption required for the oracle property or not is unknown and hard to be testified. It was shown in [112] that when the prediction accuracy is used as the criterion to choose the tuning parameter, certain procedures are not consistent in terms of feature selection in general. Now it is necessary to provide a criterion to ensure the consistency of the tuning parameter, or equivalently, a final consistent selection criterion to identify the best model.

Regarding the final selection criterion, AIC and BIC fail under high-dimensional situation since they are inclined to engender models with too many misleading covariates, which are highly correlated with the response due to spurious correlation with the causal features. For their extensions, it was shown in [180] that, for finite p, to select the regularization parameter, BIC-type selector is selection consistent and AIC-type tends to overfit with positive probability. However, their theoretical behavior under high-dimensional situation remains unknown. The little bootstrap was proposed in [20] to give almost unbiased estimates for sub-model prediction error and used these to do sub-model selection. A modified BIC (mBIC) was proposed in [15] for the study of genetic QTL mapping to address the problem of likely inclusion of spurious effects. They noticed that epistatic terms appearing in a model without the related main effects cause BIC to have a strong tendency to

overestimate the number of interactions and QTL number. It was discovered in [16] that this mBIC can be connected with the well known Bonferroni correction for multiple testing. Hypothesis testing was applied in [168] to eliminate some variables at the final selection stage. A family of extended Bayesian information criteria (EBIC) was developed in [33] for feature selection in high-dimensional studies, which asymptotically includes mBIC as a special case. It was also proved in [33] that EBIC is selection consistent for LMs when the dimension of feature space is of polynomial order of the sample size and the true parameter vector is fixed.

Most importantly, we need to be aware that in real applications, cases become more complicated. For instance, in LMs, it is reasonable to assume diverging number of relevant features with magnitudes converging to zero (See [49], [166]). Feature selection under small-n-large-p situation in high-dimensional studies with non-linear regression models such as logistic regression in Generalized Linear Regression models (GLMs) and Cox's Proportional Hazards (CPH) models need to be investigated as well because of the prevalence of these models in case-control studies and survival analysis.

1.2.2 Feature Selection in Non-linear Regression Models

Feature selection in non-linear regression models is as prevalent as in LMs. For example, in cancer research, gene expression data is often reported in tandem with time to event information such as time to metastasis, death or relapse ([4]).

Given a high-dimensional feature space, feature selection in non-linear models has more challenges due to the complicated data structure and implicit estimators compared with LMs ([60]). Most feature selection techniques in these models were applications of those techniques in LMs, such as [29], [114], [174], [119], [51]. Certain famous procedures introduced in LMs have been systematically investigated in many non-linear regression models subsequently.

SIS and ISIS were extended to GLMs in [64], [65] and also to Cox model in [57]. Their sure screening property was also testified under certain conditions.

The LASSO, the SCAD and the adaptive LASSO were respectively applied in Cox model for feature selection in [157], [59] and [181]. The asymptotic selection consistency of L_1 and $L_1 + L_2$ in linear and logistic regression models was proved in [27]. For the simplicity of computation, an efficient and adaptive shrinkage method was proposed in [186] for feature selection in the Cox model, which tends to outperform the LASSO and the SCAD estimators with moderate sample sizes for $n > p$

situation. Other path solution algorithms can be found in [128] (`glmpath`) and [74] (`glmnet`). As a generalization of the likelihood or partial likelihood term in usual penalized feature selection methods, feature selection in GLMs with Lipschiz loss functions with LASSO penalty was studied in [141]. Most of these procedures have been proved to possess oracle property under regular conditions. For more complex models and data structures, the oracle properties of LASSO in non-parametric regression setting were proved in [28]. In [103], the author proposed a new LASSO-type method for censored data after one-step imputation and presented a tremendous new challenge. The analysis performed in [104] reveals the distinct advantages of the non-concave penalized likelihood methods over traditional model selection techniques, they also discussed the performance and the pros and cons of various techniques in large medical data in logistic regression.

For subset or sub-models selection criterion, the authors of [164] extended the BIC to the Cox model by changing the sample size in the penalty term to the number of uncensored events. It was also proved that EBIC is selection consistent for GLMs with canonical link functions in [35] under high dimensional situations. The consistency of EBIC for Gaussian graphical models was established in [70]. EBIC was used in [106] to determine the final model in finite mixture of sparse normal linear models in large feature spaces when multiple sub-populations are available. It can be expected that EBIC could preserve its selection consistency

for a much broader range of models with high or ultra-high dimensional feature spaces.

1.3 Objectives and Organizations

The objectives of this thesis include two main parts. The first part focuses on investigating the selection consistency of a two-stage procedure where EBIC is utilized as the final selection criterion in LMs, GLMs with general canonical link functions and CPH models. The second part of this thesis is to introduce a new feature selection procedure-sequential LASSO and to discuss its properties.

Part I includes Chapters 2, 3, 4, 5. In Chapter 2, we introduce EBIC in detail. In Chapter 3, we examine the selection consistency of the EBIC in feature selection in linear regression models under a more general scenario where both the number of relevant features and their effects are allowed to depend on the sample size in a high-dimensional or ultra-high dimensional feature space. We give the conditions under which the EBIC remains selection consistent and provide the theoretical proof. We also compare these conditions with those imposed for oracle property in penalized likelihood procedures such as in [183], [165], [107], and our proposition implies that ours are much weaker. This study in linear regression models is followed by its extension to GLM in Chapter 4 and CPH in Chapter 5.

As a preliminary work for CPH, we assume that the dimension of feature space is of polynomial order of the sample size and the true parameter vector in the model is independent of the sample size. We believe that for more complex scenarios as in LMs, the selection consistency of EBIC can be expected and verified with additional technical details. In each of Chapters 3 to 5, we also conduct extensive numerical studies to show the finite sample performances of a two-stage procedure with EBIC as the final selection criteria as supportive evidences of our theories. Both simulation results and real data analysis on QTL mapping are covered. Our numerical studies comprise different data structures in linear regression models, GLMs and CPH. Results showed that in all scenarios, the EBIC perform as well as in linear regression models under high-dimensional feature space.

Part II includes Chapters 6, 7, 8. In this part, we attempt to overcome the impact of high spurious correlation among features in feature selection using our newly developed method-sequential LASSO. In Chapter 6, its underlying theory and computation issues are stated in detail. Moreover, in Chapter 7, we have scrutinized the conditions required for its selection consistency. The EBIC as a stopping rule for sequential LASSO is proposed, the selection consistency of this integrated procedure is established. We apply this procedure to simulated and real data analysis. Compared with its competing approaches, sequential LASSO with EBIC as a stopping rule is shown to be a promising feature selection procedure in

ultra-high dimensional situations. In Chapter 8, we show that sequential LASSO enjoys sure screening property under much weaker conditions than Forward Selection.

In Chapter 9, we provide overall conclusions and discussions on open questions for future research to complete this thesis.

Part I

Extended Bayesian Information Criteria

In this part, we examine the applicability of the EBIC in more general and complicated models. A detailed introduction of the EBIC is given in Chapter 2. The necessary conditions for its selection consistency in LMs, GLMs and CPH are established in Chapters 3, 4 and 5. Our conclusion for this part is given after Chapter 5. We also conduct extensive numerical studies to demonstrate the finite sample performance of the EBIC in these chapters. Moreover, since QTL mapping is one of the motivations for this thesis, we also provide several real data applications of EBIC. The comparison between our findings and those in previous literatures is also given.

CHAPTER 2

Introduction to EBIC

2.1 Derivation of EBIC

In a parametric regression model, if the number of features (covariates) p_n or its logarithm is of the polynomial order of the sample size n, i.e., $p_n = O(n^\kappa)$ or $\ln p_n = O(n^\kappa)$ for some positive constant κ, the feature space is referred to as a high-dimensional or ultra-high dimensional feature space. Regression problems with high or ultra-high dimensional feature spaces arise in many important fields of scientific research such as genomics study, medical study, risk management, machine learning, etc. Such problems are generally referred to as small-n-large-p

problems.

The EBIC was developed in [33] for feature selection in small-n-large-p problems. The family of EBIC is indexed by a parameter γ in the range [0, 1], it includes the original BIC and mBIC as its special cases exactly or asymptotically when $\gamma = 0$ and $\gamma = 1$.

The EBIC was motivated from a Bayesian paradigm. Let $\{(y_i, x_i) : i = 1, 2, \ldots, n\}$ be independent observations. Suppose that the conditional density function of y_i given x_i is $f(y_i|x_i, \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \Theta \subset R^{p_n}, p_n$ being a positive integer. The likelihood function of $\boldsymbol{\beta}$ is given by

$$L_n(\boldsymbol{\beta}) = f(x; \boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|x_i, \boldsymbol{\beta}).$$

Denote $Y = (y_1, y_2, \ldots, y_n)$. Let s be a subset of $\{1, 2, \ldots, p_n\}$. Denote by $\boldsymbol{\beta}(s)$ the parameter $\boldsymbol{\beta}$ with those components outside s being set to 0. Let $\mathcal{S}$ be the model space under consideration, i.e, $\mathcal{S} = \{s : s \subseteq \{1, 2, \cdots, p_n\}\}$, let $p(s)$ be the prior probability of model s. Assume that, given s, the prior density of $\boldsymbol{\beta}(s)$ is $\pi(\boldsymbol{\beta}(s))$. The posterior probability of s is obtained as

$$p(s|Y) = \frac{m(Y|s)p(s)}{\sum_{s \in \mathcal{S}} m(Y|s)p(s)},$$

where $m(Y|s)$ is the likelihood of model s, given by

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

The BIC selects the model that minimizes

$$\mathrm{BIC}(s) = -2 \ln L_n \left(\hat{\boldsymbol{\beta}}(s)\right) + |s| \ln n,$$

where $\hat{\boldsymbol{\beta}}(s)$ is the maximum likelihood estimator of $\boldsymbol{\beta}(s)$ and $|s|$ is the number of components in s. When $\hat{\boldsymbol{\beta}}(s)$ is $\sqrt{n}$ consistent, $-2 \ln\left(m(Y|s)\right)$ has a Laplace approximation given by the $\mathrm{BIC}(s)$ up to an additive constant. In the derivation of BIC, this constant $p(s)$ is taken as a constant over all s. With this constant prior, BIC favors models with larger numbers of features in small-n-large-p problems (see [22], [15]).

Assume that $\mathcal{S}$ is partitioned into $\cup_{j=1}^{p_n} \mathcal{S}_j$, such that models within each $\mathcal{S}_j$ have equal dimension j. Let $\tau(\mathcal{S}_j)$ be the size of $\mathcal{S}_j$. Assign the prior distribution $P(\mathcal{S}_j)$ proportional to $\tau^{\xi}(\mathcal{S}_j)$ for some ξ between 0 and 1. For each $s \in \mathcal{S}_j$, assign equal probability, $p(s|\mathcal{S}_j) = 1/\tau(\mathcal{S}_j)$, this is equivalent to $P(s)$ for $s \in \mathcal{S}_j$ proportional to $\tau^{-\gamma}(\mathcal{S}_j)$ where $\gamma = 1 - \xi$. This extended BIC family is given by

$$\mathrm{EBIC}_{\gamma}(s) = -2 \ln L_n \left(\hat{\boldsymbol{\beta}}(s)\right) + |s| \ln n + 2\gamma \ln\left(\tau(\mathcal{S}_{|s|})\right), 0 \leq \gamma \leq 1. \qquad (2.1.1)$$

When the feature space is high-dimensional and the relevant features are fixed, the selection consistency of EBIC in linear regression models was established in [33] when $p_n = O(n^\kappa)$ and $\gamma > 1 - \frac{1}{2\kappa}$ for any positive constant κ, which suggests that the original BIC may not be selection consistent when p_n is of order higher than $O(\sqrt{n})$. In the following chapters of this part, we examine the selection consistency of the EBIC in more general models for a wider application of the EBIC.

2.2 Applications of EBIC in Feature Selection

According to definition (2.1.1), the EBIC of a particular model depends on the set of features s it contains and the value of γ. Literally, the selection consistency of EBIC states that with a properly chosen γ, the EBIC corresponding to the true set of relevant features s_{0n} is the minimum among all subsets of features having comparable sizes with s_{0n}. Such a property ensures the capability of EBIC for identifying s_{0n} correctly provided that the candidate sets are not too big and s_{0n} is included in the candidate sets. Practically, it is impossible to assess all possible models, especially in the case of high or ultra-high dimensional feature spaces. It is natural to reduce the dimension of the feature space as the first step and then to generate a model sequence by using a feasible procedure, see, e.g., [61], [34], whereafter, a model selection criterion is applied. When the model sequence

is controlled by a range of tuning parameters, the model selection criterion is equivalent to the selection of tuning parameters. For the purpose of brevity, we will incorporate the model selection into the second stage. In this section, a general two-stage procedure of this nature will be elaborated and applied in succeeding numerical studies. The procedure is as follows:

(1) Screening stage: Let $\mathscr{F}_n$ denote the set of all the features. This stage screens out obviously irrelevant features by using an appropriate screening procedure and reduces $\mathscr{F}_n$ to a small set $\mathscr{F}_n^*$.

(2) Selection stage: Use a penalized likelihood of the form

$$l_{n,\lambda}\left(X(\mathscr{F}_n^*), \boldsymbol{\beta}(\mathscr{F}_n^*)\right) = -2\ln L_n(X(\mathscr{F}_n^*), \boldsymbol{\beta}(\mathscr{F}_n^*)) + \sum_{j\in\mathscr{F}_n^*} p_\lambda(|\beta_j|),$$

where $L_n(X(\mathscr{F}_n^*), \boldsymbol{\beta}(\mathscr{F}_n^*))$ is the likelihood function of the model with all features in $\mathscr{F}_n^*$ and $p_\lambda(\cdot)$ is a penalty function with desirable properties including the property of sparsity. Choose λ by EBIC as follows. Given a range $\mathscr{R}_\lambda$, for each $\lambda \in \mathscr{R}_\lambda$, let $s_{n\lambda}$ be the set of features with non-zero coefficients when $l_{n,\lambda}(X(\mathscr{F}_n^*), \boldsymbol{\beta}(\mathscr{F}_n^*))$ is minimized. Based on (2.1.1), compute

$$\mathrm{EBIC}_\gamma(s_{n\lambda}) = -2\ln L_n(X(s_{n\lambda}), \hat{\boldsymbol{\beta}}(s_{n\lambda})) + |s_{n\lambda}|\ln n + 2\gamma\ln\binom{p_n}{|s_{n\lambda}|},$$

where $\hat{\boldsymbol{\beta}}(s_{n\lambda})$ is the maximum likelihood estimate (without penalty) of $\boldsymbol{\beta}(s_{n\lambda})$ and γ is taken to be $1 - \frac{\ln n}{C \ln p_n}$ for some $C > 2$. Let λ^* be the one which attains the minimum $\text{EBIC}_\gamma(\lambda)$. The final selected set of features is $s_{n\lambda^*}$.

It is straightforward to see that, suppose under certain conditions, the following properties hold:

(1) Sure Screening Property of the screening procedure: $P(\mathscr{F}_n^* \in \mathscr{F}_n) \to 1$, as n goes to infinity;

(2) Oracle Property of the penalized likelihood procedure: there exists $\lambda_0 \in \mathscr{R}_\lambda$ such that $P(s_{n\lambda_0} = s_{0n}) \to 1$, as n goes to infinity;

(3) Selection Consistency of the EBIC_γ : $P\left(\text{EBIC}_\gamma(s_{0n}) = \min_{\lambda \in \mathscr{R}_\lambda} \text{EBIC}_\gamma(s_{n\lambda})\right) \to 1$, as n goes to infinity.

Then the overall selection consistency of the two-stage procedure is attained. For a specified combination of techniques, for example, SIS followed by SCAD with EBIC, the coexistence of the conditions for SIS's sure screening property, SCAD's oracle property and EBIC's selection consistency can be easily verified. In this part, we will show the finite sample performance of this two-stage feature selection procedure in LMs, GLMs with non-canonical links and CPHs in sections 3.2, 4.2 and 5.2 respectively.

When this thesis was almost done, we found that the screening property is no longer necessary for the realization of regularization such as adaptive LASSO and SCAD. See [92] and [107]. We believe that better performances can be achieved, but our focus, the selection consistency of the EBIC will not be influenced.

In order to measure the closeness of a selected set to the true set of relevant features, or equivalently, the selection accuracy of a certain procedure, the two quantities, positive discovery rate (PDR) and false discovery rate (FDR) are adopted. Given a data set with n independent observations, suppose s and s_{0n} are the selected and the true set of relevant features, the empirical versions of PDR and FDR are defined as follows:

$$\mathrm{PDR}_n = \frac{|s \cap s_{0n}|}{|s_{0n}|}, \quad \mathrm{FDR}_n = \frac{|s \cap s_{0n}^c|}{|s|}. \tag{2.2.1}$$

The simultaneous convergence of PDR_n to 1 and FDR_n to 0 reflects the asymptotic selection consistency in the sense that s itself and the true relevant features it contains both have almost the same sizes as those of s_{0n}. In this thesis, we will use these two measures for the evaluation of EBIC's selection consistency in simulation studies.

CHAPTER 3

EBIC in Linear Regression Models

3.1 Selection Consistency of EBIC

In many small-n-large-p problems that the relevant (or causal, true, as referred by some other authors) features, though sparse, are relatively large in number compared with classical statistical problems, and their effects usually taper off to zero from the largest to the smallest. To reflect the estimability of the feature effects, it is reasonable to model the number of relevant features as a diverging sequence

depending on the sample size. [49] and [63] are among the earliest papers dealing with diverging number of relevant features. In this subsection, we investigate the property of the EBIC in feature selection in LMs when the number of relevant features p_{0n} diverges at the order $O(n^c)$ for some $0 < c < 1$ and $p_n = O(n^\kappa)$ for any κ or $\ln p_n = O(n^\kappa)$ for some $0 < \kappa < 1$. We give the conditions under which the EBIC remains selection consistent and provide the theoretical proof (Theorem 3.1.1).

We denote by p_n the number of features under investigation to make its dependence on n explicit. Let $(y_i, x_{i1}, \ldots, x_{ip_n}), i = 1, \ldots, n$, be independent observations. We consider the following linear model

$$y_i = \sum_{j=1}^{p_n} \boldsymbol{\beta}_{0j} x_{ij} + \epsilon_i, \ i = 1, \ldots, n, \tag{3.1.1}$$

where ϵ_i's are i.i.d. with mean zero and variance σ^2. In matrix notation, (3.1.1) is expressed as

$$\boldsymbol{y}_n = X_n \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_n,$$

where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0p_n})^\tau$, $\boldsymbol{y}_n = (y_1, \ldots, y_n)^\tau$ and $X_n = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p_n}}$. Here either p_n or $\ln p_n$ is of a polynomial order of n, and $\boldsymbol{\beta}_0$ is sparse, meaning that only a few of its components are non-zero.

We first introduce some notations. Let $s_{0n} = \{j : \boldsymbol{\beta}_{0j} \neq 0, j \in \{1, \ldots, p_n\}\}$. Let

s be any subset of $\{1, \ldots, p_n\}$. For convenience, we also refer to s as a submodel. We denote by $X(s)$ the matrix composed of the columns of X_n with indices in s. Similarly, $\boldsymbol{\beta}_0(s)$ denotes the vector consisting of components of $\boldsymbol{\beta}_0$ with indices in s. Let $|s|$ denote the number of components in s. In particular, let $p_{0n} = |s_{0n}|$. Let $H_0(s)$ be the projection matrix of $X(s)$, i.e., $H_0(s) = X(s)[X(s)^\tau X(s)]^{-1}X(s)^\tau$. Define

$$\Delta_n(s) = \|\boldsymbol{\mu}_n - H_0(s)\boldsymbol{\mu}_n\|_2^2,$$

where $\boldsymbol{\mu}_n = \mathrm{E}\boldsymbol{y}_n = X(s_{0n})\boldsymbol{\beta}_0(s_{0n})$ and $\|\cdot\|_2$ is the L_2 norm. First we consider the following condition which determines the divergence pattern of (n, p_n, p_{0n}) and the constraint on $\boldsymbol{\beta}_0$ required for the selection consistency of the EBIC.

Consistency Condition: *Let $k_n = kp_{0n}$ for any fixed $k > 1$, then*

$$\lim_{n\to\infty} \min\left\{\frac{\Delta_n(s)}{p_{0n}\ln p_n} : s_0 \not\subset s, |s| \le k_n\right\} = \infty.$$

The restriction $|s| \le k_n$ is imposed because in practice only the models with size comparable with the true model or smaller will be considered. Implicitly, the consistency condition requires that

$$\lim_{n\to+\infty} \sqrt{\frac{n}{p_{0n}\ln p_n}}\, \min\{|\boldsymbol{\beta}_{0j}| : j \in s_{0n}\} = +\infty. \tag{3.1.2}$$

This statement can be verified in the following: without loss of generality, we let $s_{0n} = \{1, 2, \cdots, p_n\}$ and $s_{-k} = s_{0n} \backslash \{k\}$, then

$$\begin{aligned} &\min\left\{\frac{\Delta_n(s)}{p_{0n}\ln p_n} : s_0 \not\subset s, |s| \le k_n\right\} \le \min\left\{\frac{\Delta_n(s_{-k})}{p_{0n}\ln p_n} : 1 \le k \le p_{0n}\right\} \\ \le &\min\left\{\frac{n\boldsymbol{\beta}_{0k}^2}{p_{0n}\ln p_n} : 1 \le k \le p_{0n}\right\} = \min\{\boldsymbol{\beta}_{0j}^2 : j \in s_{0n}\}\frac{n}{p_{0n}\ln p_n}. \end{aligned}$$

We now discuss a relationship between the consistency condition above and the well known Sparse Reisz Condition ([179]). The Sparse Reisz Condition is as follows:

$$\begin{aligned} 0 < c_{\min} &\le \min\{\lambda_{\min}(\frac{1}{n}X(s)^\tau X(s)) : |s| \le k_n\} \\ &\le \max\{\lambda_{\max}(\frac{1}{n}X(s)^\tau X(s)) : |s| \le k_n\} \le c_{\max} < \infty, \end{aligned}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and the largest eigenvalues respectively. If p_{0n} is fixed, then $\{\boldsymbol{\beta}_{0j} : j \in s_{0n}\}$ is also fixed, and $p_{0n}\ln p_n = o(n)$, (3.1.2) is always true. As shown in [33], the Sparse Reisz Condition implies the consistency condition. If p_{0n} diverges, then the Sparse Reisz Condition together with (3.1.2) imply the consistency condition. When p_{0n} diverges, conditions of the type (3.1.2) are always imposed for selection consistency in penalized likelihood procedures, see [107], [183]. As the following proposition implies, the Sparse Reisz Condition together with (3.1.2) are stronger assumptions than the consistency condition.

Proposition 3.1.1. *Assume $s_{0n} = \{1, 2, \ldots, p_{0n}\}$. Let s_{-k} be the set with the kth element of s_{0n} removed. Let $k(s) = s_{-k} \cup s$. If (3.1.2) is satisfied and*

$$\lim_{n\to+\infty} \min_{s:|s|\leq k_n, s_{0n}\subsetneq s} \frac{\max_{k\in s_{0n}} \{\|[\boldsymbol{I} - H_0(k(s))]X(\{k\})\|_2^2\}}{n} \geq c \tag{3.1.3}$$

for some constant $c > 0$, then the consistency condition holds.

Proof of Proposition 3.1.1. The first inequality in the proof of Lemma 1 in [33] implies that

$$\begin{aligned}\triangle_n(s) &\geq \max_{k\in s_{0n}} \beta_{nk}^2 \|[\boldsymbol{I} - H_0(k(s))]X(\{k\})\|_2^2 \\ &\geq \min_{k\in s_{0n}} \beta_{nk}^2 \max_{k\in s_{0n}} \left\{\|[\boldsymbol{I} - H_0(k(s))]X(\{k\})\|_2^2\right\},\end{aligned}$$

as desired. □

The sparse Riesz condition implies Proposition 3.1.1 because

$$\begin{aligned}&\lim_{n\to+\infty} \min_{s:|s|\leq k_n, s_{0n}\subsetneq s} \frac{\min_{k\in s_{0n}} \{\|[\boldsymbol{I} - H_0(k(s))]X(\{k\})\|_2^2\}}{n} \\ \geq &\lim_{n\to+\infty} \min_{s:|s|\leq k_n, s_{0n}\subsetneq s} \lambda_{\min}(\frac{1}{n} X^\tau(s_{0n}\cup s)X(s_{0n}\cup s)).\end{aligned}$$

But the inverse is not true, which will be illustrated by two counterexamples in the following.

The same as [33], when there is one relevant feature being orthogonal to all the other features, even if there is serious multi-collinearity, the Proposition 3.1.1 holds but the sparse Riesz condition fails on the left hand side. Regarding the right hand side, the sparse Riesz condition fails but the Proposition 3.1.1 still holds. When all the pairwise correlation coefficients are ρ and all the features have variance 1,

$$\begin{aligned}\lambda_{\max}(\frac{1}{n}X(s)^\tau X(s)) =& 1-\rho+\rho|s|;\\ \frac{\|[\boldsymbol{I}-H_0(k(s))]X(\{k\})\|_2^2}{n} =& (1-\rho)\frac{\rho|s|+1}{\rho|s|+(1-\rho)}.\end{aligned}$$

Hence,

$$\begin{aligned}\max_{|s|\le k_n}\lambda_{\max}(\frac{1}{n}X(s)^\tau X(s)) &= (1-\rho)+\rho k_n;\\ \lim_{n\to+\infty}\min_{s:|s|\le k_n, s_{0n}\subsetneq s}\frac{\max_{k\in s_{0n}}\{\|[\boldsymbol{I}-H_0(k(s))]X(\{k\})\|_2^2\}}{n} &> \frac{1-\rho}{2}.\end{aligned}$$

Condition (3.1.2) determines the divergence pattern of (n, p_{0n}, p_n) and the constraint on $\boldsymbol{\beta}_0$. Now consider the high and ultra-high-dimensional feature spaces separately. If $p_n = O(n^\kappa)$ for any fixed $\kappa > 0$ and $p_{0n} = n^c$ for some $0 < c < \kappa$, (3.1.2) reduces to $\frac{n^{1-c}}{\ln n}\min\{|\boldsymbol{\beta}_{0j}^2| : j \in s_{0n}\} \to \infty$. The constraint on $\boldsymbol{\beta}_{0j}$ is then $\min\{|\boldsymbol{\beta}_{0j}^2| : j \in s_{0n}\}$ must have a magnitude larger than $O(n^{-(1-c)})$. Let b be any number bigger than c and smaller than 1. Then the following provides a consistency pattern: $(n, p_{0n}, p_n) = (n, O(n^c), O(n^\kappa))$, $\min\{|\boldsymbol{\beta}_{0j}| : j \in s_{0n}\} =$

$O(n^{-(1-b)/2})$, $0 < c < \kappa$, $c < b < 1$. If $\ln p_n = O(n^\kappa)$ and $p_{0n} = n^c$, then by the same argument, (3.1.2) induces the following consistency pattern: $(n, p_{0n}, \ln p_n) = (n, O(n^c), O(n^\kappa))$, $\min\{|\boldsymbol{\beta}_{0j}| : j \in s_{0n}\} = O(n^{-(1-b)/2})$, $0 < c, \kappa < 1$, $c + \kappa < b < 1$.

We now state the main result on the selection consistency of the EBIC with diverging number of parameters and high or ultra-high dimensional feature spaces.

Theorem 3.1.1. *Assume model (3.1.1) and the consistency condition. In addition, assume that* $p_{0n} \ln p_n = o(n)$, $p_{0n} \ln n = o(n)$, $\ln p_{0n} / \ln p_n \to \delta$ *where* $0 \le \delta < 1$. *Let* $k_n = k p_{0n}$ *for any constant* $k > 1$. *Then as* n *tends to* $+\infty$,

$$P\left(\min_{s:|s|\le k_n} EBIC_\gamma(s) > EBIC_\gamma(s_{0n})\right) \to 1,$$

if $\gamma > \dfrac{1+\delta}{1-\delta} - \dfrac{\ln n}{2(1-\delta)\ln p_n}$.

The following are immediate corollaries of Theorem 3.1.1.

Corollary 3.1.1. *If* $p_n = O(n^\kappa)$ *for any constant* $\kappa > 0$, $p_{0n} = p_0$ *is fixed, the EBIC is selection consistent with* $\gamma > 1 - \dfrac{\ln n}{2 \ln p_n} = 1 - \dfrac{1}{2\kappa}$ *among all models* s *with* $|s| \le k_n$.

Corollary 3.1.2. *If* $p_n = O(n^\kappa)$ *for any constant* $\kappa > 0$, $p_{0n} = O(n^c)$, $\min\{|\boldsymbol{\beta}_{0j}| : j \in s_{0n}\} = O(n^{-(1-b)/2})$, $0 < c < \kappa$, $c < b < 1$, *then the EBIC is selection consistent with* $\gamma > \dfrac{\kappa + c - 0.5}{\kappa - c}$ *among all models* s *with* $|s| \le k_n$.

Corollary 3.1.3. *If* $\ln p_n = O(n^\kappa)$ *for* $0 < \kappa < 1$, $p_{0n} = O(n^c)$, $\min\{|\boldsymbol{\beta}_{0j}| : j \in s_{0n}\} = O(n^{-(1-b)/2})$, $0 < c, \kappa < 1$, $c + \kappa < b < 1$, *the EBIC is selection consistent with* $\gamma > 1 - \dfrac{\ln n}{2 \ln p_n}$ *among all models* s *with* $|s| \leq k_n$.

The following two lemmas are needed for the proof of Theorem 3.1.1.

Lemma 3.1.1. *If* $\dfrac{\ln j}{\ln p} \to \delta$ *as* $p \to +\infty$, *we have*

$$\ln(\frac{p!}{j!(p-j)!}) = j \ln p(1-\delta)(1+o(1)).$$

Proof of Lemma 3.1.1: Write

$$\frac{p!}{j!(p-j)!} = \frac{p(p-1)\dots(p-j+1)}{j!} = \frac{p^j \left(1-\frac{1}{p}\right)\dots\left(1-\frac{j-1}{p}\right)}{j!}.$$

Note that

$$\left(1-\frac{j-1}{p}\right)^{j-1} < \left(1-\frac{1}{p}\right)\dots\left(1-\frac{j-1}{p}\right) < \left(1-\frac{1}{p}\right)^{j-1},$$

and, see [135], that

$$\sqrt{2\pi} j^{j+1/2} e^{-j+1/(12j+1)} < j! < \sqrt{2\pi} j^{j+1/2} e^{-j+1/(12j)}.$$

We now have

$$\begin{aligned}\ln(\frac{p!}{j!(p-j)!}) \leq & j\ln p + (j-1)\ln(1-1/p) - (j+1/2)\ln j + j - \frac{1}{12j+1} - \ln\sqrt{2\pi} \\ \leq & j\ln p - (j+1/2)\ln j + j = j\ln p[1 - \frac{(j+1/2)\ln j}{j\ln p} + \frac{1}{\ln p}] \\ = & j\ln p(1-\delta)(1+o(1));\end{aligned} \tag{3.1.4}$$

$$\begin{aligned}\ln(\frac{p!}{j!(p-j)!}) \geq & j\ln p + (j-1)\ln(1-\frac{j-1}{p}) - (j+1/2)\ln j + j - \frac{1}{12j} - \ln\sqrt{2\pi} \\ \geq & j\ln p + (j-1)\ln\left(1-\frac{j-1}{p}\right) - (j+1/2)\ln j - \ln\sqrt{2\pi} \\ = & j\ln p\left(1 + \frac{(j-1)\ln\left(1-\frac{j-1}{p}\right)}{j\ln p} - \frac{(j+1/2)\ln j}{j\ln p} - \frac{\ln\sqrt{2\pi}}{j\ln p}\right) \\ = & j\ln p(1-\delta)(1+o(1)).\end{aligned} \tag{3.1.5}$$

Lemma 3.1.1 follows from (3.1.4) and (3.1.5). □

Lemma 3.1.2. *Let χ^2_k denote a χ^2 random variable with degrees of freedom k. If $m \to +\infty$ and $\frac{K}{m} \to 0$ then*

$$P(\chi^2_k \geq m) = \frac{1}{\Gamma(k/2)}(m/2)^{k/2-1}e^{-m/2}(1+o(1))$$

uniformly for all $k \leq K$.

Proof of Lemma 3.1.2: Denote $\bar{F}_k(m) = P(\chi_k^2 \geq m)$. By integration by parts, we obtain a recursive formula,

$$\bar{F}_k(m) = \frac{1}{2^{k/2}\Gamma(k/2)} \int_m^{+\infty} x^{k/2-1} e^{-x/2} dx = \frac{1}{\Gamma(k/2)} (m/2)^{k/2-1} e^{-m/2} + \bar{F}_{k-2}(m).$$

If k is even,

$$\bar{F}_k(m) = \frac{1}{\Gamma(k/2)} (m/2)^{k/2-1} e^{-m/2} [1 + \sum_{i=1}^{(k-2)/2} (\frac{(k/2-1)\dots(k/2-i)}{(m/2)^i})].$$

If k is odd,

$$\bar{F}_k(m) = \frac{1}{\Gamma(k/2)} (m/2)^{k/2-1} e^{-m/2} [1 + \sum_{i=1}^{(k-3)/2} (\frac{(k/2-1)\dots(k/2-i)}{(m/2)^i})] + \bar{F}_1(m),$$

where $\bar{F}_1(m) = P(\chi_1^2 \geq m) \approx 2\frac{\exp(-m/2)}{\sqrt{2\pi m}} = \frac{1}{\Gamma(k/2)} (m/2)^{k/2-1} e^{-m/2} \frac{2\Gamma(k/2)}{\sqrt{2\pi}(m/2)^{(k-1)/2}}$ when $m \to +\infty$. We can write

$$\bar{F}_k(m) = \frac{1}{\Gamma(k/2)} (m/2)^{k/2-1} e^{-m/2} [1 + R(k, m)].$$

It is straightforward to see that $R(k, m) \leq R(K, m) \to 0$ when $m \to +\infty$. □

Proof of Theorem 3.1.1:

Let s be any submodel. Decompose $\mathrm{EBIC}_\gamma(s) - \mathrm{EBIC}_\gamma(s_{0n})$ as $T_1 + T_2$ where

$$\begin{aligned} T_1 =& n \ln \frac{\boldsymbol{y}_n^\tau [\boldsymbol{I} - H_0(s)]\boldsymbol{y}_n}{\boldsymbol{y}_n^\tau [\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{y}_n} = n \ln \frac{\boldsymbol{y}_n^\tau [\boldsymbol{I} - H_0(s)]\boldsymbol{y}_n}{\boldsymbol{\epsilon}_n^\tau [\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n} \\ =& n \ln \left\{ 1 + \frac{\boldsymbol{y}_n^\tau [\boldsymbol{I} - H_0(s)]\boldsymbol{y}_n - \boldsymbol{\epsilon}_n^\tau [\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n}{\boldsymbol{\epsilon}_n^\tau [\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n} \right\} \\ T_2 =& (|s| - p_{0n}) \ln n + 2\gamma (\ln \tau(\mathcal{S}_{|s|}) - \ln \tau(\mathcal{S}_{p_{0n}})). \end{aligned} \tag{3.1.6}$$

Case I: $s_{0n} \not\subset s$.

Without loss of generality, assume $\sigma^2 = 1$. We can write

$$\boldsymbol{\epsilon}_n^\tau \{\boldsymbol{I} - H_0(s_{0n})\}\boldsymbol{\epsilon}_n = \sum_{i=1}^{n-p_{0n}} Z_i^2 = (n - p_{0n})(1 + o_p(1)) = n(1 + o_p(1)), \tag{3.1.7}$$

where Z_i's are i.i.d. standard normal variables, since $H_0(s_{0n})$ is a projection matrix with rank p_{0n}. We have

$$\begin{aligned} & \boldsymbol{y}_n^\tau [\boldsymbol{I} - H_0(s)]\boldsymbol{y}_n - \boldsymbol{\epsilon}_n^\tau [\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n \\ =& \Delta_n(s) + 2\boldsymbol{\mu}_n^\tau [\boldsymbol{I} - H_0(s)]\boldsymbol{\epsilon}_n + \boldsymbol{\epsilon}_n^\tau H_0(s)\boldsymbol{\epsilon}_n - \boldsymbol{\epsilon}_n^\tau H_0(s_{0n})\boldsymbol{\epsilon}_n. \end{aligned}$$

It is trivial that

$$\boldsymbol{\epsilon}_n^\tau H_0(s_{0n})\boldsymbol{\epsilon}_n = p_{0n}(1 + o_p(1)). \quad \text{(I)}$$

We will show

$$\max\{\boldsymbol{\epsilon}_n^\tau H_0(s)\boldsymbol{\epsilon}_n, |s| \le k_n\} = O_p(k_n \ln p_n), \quad \text{(II)}$$

and

$$|\boldsymbol{\mu}_n^\tau[\boldsymbol{I} - H_0(s)]\boldsymbol{\epsilon}_n| = \sqrt{\Delta_n(s) O_p(k_n \ln p_n)}, \quad \text{(III)}$$

uniformly for all s with $|s| \le k_n$. Under the assumption of the theorem, $2k_n \ln p_n = o(n)$. Then, by the asymptotic identifiability condition, (I), (II) and (III) imply that

$$\boldsymbol{y}_n^\tau[\boldsymbol{I} - H_0(s)]\boldsymbol{y}_n - \boldsymbol{\epsilon}_n^\tau[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n = \Delta_n(s)(1 + o_p(1)), \tag{3.1.8}$$

uniformly for all s with $|s| \le k_n$. It then follows from (3.1.7) and (3.1.8) that

$$T_1 = n \ln\left(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1))\right), \tag{3.1.9}$$

uniformly for all s with $|s| \le k_n$.

We now prove (II) and (III) in the following. Let $m = 2k_n[\ln p_n + \ln(k_n \ln p_n)]$. It is obvious that $\dfrac{k_n}{m} \to 0$. Note that we can express $\boldsymbol{\epsilon}_n^\tau H_0(s)\boldsymbol{\epsilon}_n = \chi_j^2(s)$ where

$j = |s|$. By the Bonferroni inequality, we have

$$\begin{aligned} &P(\max\{\boldsymbol{\epsilon}_n^\tau H_0(s)\boldsymbol{\epsilon}_n : |s| \le k_n\} \ge m) \\ =&P(\max\{\chi_j^2(s) : s \in \mathcal{S}_j, j \le k_n\} \ge m) \le \sum_{j=1}^{k_n} \tau(\mathcal{S}_j) P(\chi_j^2 \ge m). \end{aligned}$$

By the fact that $\tau(\mathcal{S}_j) = \binom{p_n}{j} \le p_n^j$ and Lemma 3.1.2, there is some c close to 1, not depending on j for $j \le k_n$, such that

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

where

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

for some q between 0 and 1, when n is large enough, since $q_n \to 0$. Thus

$$P(\max\{\boldsymbol{\epsilon}_n^\tau H_0(s)\boldsymbol{\epsilon}_n : |s| \le k_n\} \ge m) \le \frac{c}{m} \sum_{j=1}^{k_n} q^j \le \frac{c}{m} \frac{q}{1-q} \to 0; \tag{3.1.10}$$

that is,

$$\max\{\boldsymbol{\epsilon}_n^\tau H_0(s)\boldsymbol{\epsilon}_n : |s| \le k_n\} = m(1 + o_p(1)) = O_p(k_n \ln p_n),$$

which establishes (II).

For verifying (III), note that we can express

$$\boldsymbol{\mu}_n^\tau\{\boldsymbol{I} - H_0(s)\}\boldsymbol{\epsilon}_n = \sqrt{\Delta_n(s)}Z(s),$$

where $Z(s) \sim \mathrm{N}(0,1)$. For any s with $|s| \leq k_n$, we have

$$|\boldsymbol{\mu}_n^\tau\{\boldsymbol{I} - H_0(s)\}\boldsymbol{\epsilon}_n| \leq \sqrt{\Delta_n(s)}\max\{|Z(s)| : |s| \leq k_n\}.$$

Let m be the same as above. Consider $P(\max\{|Z(s)| : |s| \leq k_n\} \geq \sqrt{m})$. We have

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

since $P(\chi_1^2 \geq m) < P(\chi_j^2 \geq m)$ by Lemma 3.1.2. We have already shown that the last sum converges to zero. This establishes (III).

Now, putting (3.1.6) and (3.1.9) together, we have

$$\begin{aligned}
&\mathrm{EBIC}_\gamma(s) - \mathrm{EBIC}_\gamma(s_{0n}) \\
=& n\ln\left(1 + \frac{\Delta_n(s)}{n}(1 + o_p(1))\right) + (|s| - p_{0n})\ln n + 2\gamma(\ln\tau(\mathcal{S}_{|s|}) - \ln\tau(\mathcal{S}_{p_{0n}})) \\
\geq& n\ln(1+c) - p_{0n}(\ln n + 2\gamma\ln p_n),
\end{aligned}$$

for some positive c, when n is large enough, by the asymptotic identifiability condition. Under the assumption of the theorem, $p_{0n} \ln p_n = o(n)$ and $p_{0n} \ln n = o(n)$. Hence the above difference goes to infinity uniformly for all s with $|s| \leq k_n$ for any bounded γ.

Case II: $s_{0n} \subset s$.

When $s_{0n} \subset s$, $\{\boldsymbol{I} - H_0(s)\}X(s_{0n}) = 0$. Hence, $\boldsymbol{y}_n^T\{\boldsymbol{I} - H_0(s)\}\boldsymbol{y}_n = \boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s)]\boldsymbol{\epsilon}_n$ and

$$\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n - \boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s)]\boldsymbol{\epsilon}_n = \boldsymbol{\epsilon}_n^T\{H_0(s) - H_0(s_{0n})\}\boldsymbol{\epsilon}_n = \chi_j^2(s),$$

where $\chi_j^2(s)$ is a χ^2 random variable depending on s with degrees of freedom j and $j = |s| - p_{0n}$. We obtain that

$$\begin{aligned} n \ln \left(\frac{\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n}{\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s)]\boldsymbol{\epsilon}_n} \right) =& n \ln \left\{ 1 + \frac{\chi_j^2(s)}{\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n - \chi_j^2(s)} \right\} \\ \leq& \frac{n\chi_j^2(s)}{\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n - \chi_j^2(s)}. \end{aligned} \tag{3.1.11}$$

As $n \to \infty$, $n^{-1}\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n \to \sigma^2 = 1$, i.e.,

$$\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n = n(1 + o(1)). \tag{3.1.12}$$

Let $\tilde{\mathcal{S}}_j = \{s : s \in \mathcal{S}_{j+p_{0n}}, s_{0n} \subset s\}$. Note that $\tau(\tilde{\mathcal{S}}_j) = \binom{p_n - p_{0n}}{j} \leq p_n^j$. Let

$m_j = 2j[\ln p_n + \ln(j \ln p_n)]$. In the same way as we derive (3.1.10), we have

$$P(\max_{1\le j\le k_n - p_{0n}} \frac{\max\{\chi_j^2(s) : s \in \tilde{\mathcal{S}}_j\}}{m_j} \ge 1) \le \sum_{j=1}^{k_n-p_{0n}} P(\max\{\chi_j^2(s) : s \in \tilde{\mathcal{S}}_j\} \ge m_j)$$
$$\le \sum_{j=1}^{k_n-p_{0n}} \tau(\tilde{\mathcal{S}}_j) P(\chi_j^2 \ge m_j) \le \frac{1}{\ln p_n} \sum_{j=1}^{k_n-p_{0n}} q_j^j \to 0,$$

where

$$q_j = \sqrt{\frac{2}{j \ln p_n} + \frac{2\ln(j \ln p_n)}{j(\ln p_n)^2}} \le \sqrt{\frac{2}{\ln p_n}(1 + o(1))} \to 0.$$

Thus,

$$\max\{\chi_j^2(s) : s \in \mathcal{S}_{j+p_{0n}}, s_{0n} \subset s\} = m_j\{1 + o_p(1)\}, \tag{3.1.13}$$

uniformly for all s with $|s| \le k_n$ and $s_{0n} \subset s$.

It follows from (3.1.11), (3.1.12) and (3.1.13) that

$$n \ln\left(\frac{\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s_{0n})]\boldsymbol{\epsilon}_n}{\boldsymbol{\epsilon}_n^T[\boldsymbol{I} - H_0(s)]\boldsymbol{\epsilon}_n}\right) \le \frac{nm_j}{[n - m_j(1 + o_p(1))]}$$
$$\le m_j(1 + o_p(1)) \le 2j(1 + \delta)\ln p_n(1 + o_p(1)),$$

uniformly for all s with $|s| \le k_n$ and $s_{0n} \subset s$, noting that $m_j \le 2j[\ln p_n + \ln((k_n - p_{0n})\ln p_n)] = 2j(1 + \delta)\ln p_n(1 + o_p(1))$ and $m_j = 2j(1 + \delta)\ln p_n(1 + o_p(1))$ for $j = k_n - p_{0n}$. Thus

$$T_1 \ge -2j(1 + \delta)\ln p_n(1 + o_p(1)).$$

When $p_{0n} \leq |s| \leq k_n$ we have $\ln|s|/\ln p_n \to \delta$ uniformly, hence, by Lemma 3.1.1,

$$T_2 = j\ln n + 2\gamma(1-\delta)j\ln p_n(1+o(1)).$$

Finally we have

$$\begin{aligned}&\mathrm{EBIC}_\gamma(s) - \mathrm{EBIC}_\gamma(s_{0n})\\ \geq& j\ln n + 2\gamma(1-\delta)j\ln p_n(1+o(1)) - 2j(1+\delta)\ln p_n(1+o_p(1)) > 0,\end{aligned}$$

uniformly for all s with $|s| \leq k_n$ and $s_{0n} \subset s$, if n is big enough, when $\gamma > \dfrac{1+\delta}{1-\delta} - \dfrac{\ln n}{2(1-\delta)\ln p_n}$.

□

3.2 Numerical Study

In this section, the performance of the two-stage feature selection procedure discussed previously is investigated in linear regression models where $\ln p_n = O(n^\kappa)$ and $\boldsymbol{\beta}_0$ depends on n. In the screening stage, the SIS is used and $|\mathscr{F}_n^*|$ is taken to be $0.5n$ throughout the simulations. In the selection stage, the penalized likelihood procedure with the SCAD penalty is used. We also consider the adaptive LASSO proposed by [185]. The usage of this two methods is due to the oracle property

both of them enjoy. Since the simulation results of the adaptive LASSO are similar to those of the SCAD penalized likelihood, only the results with SCAD penalty are reported in this subsection. The R packages `glmpath` ([128]) and `plus` ([178]) are used to compute the penalized likelihood models. We are mainly concerned about the EBIC with γ slightly bigger than $1 - \frac{\ln n}{2 \ln p_n}$ (in the simulation we take $\gamma = 1 - \frac{\ln n}{4 \ln p_n}$). But we also consider $\gamma = 0$, which corresponds to the original BIC, and $\gamma = 1$, which corresponds to an asymptotic form of mBIC, for the purpose of comparison.

We take the divergence pattern as $(n, p_{0n}, p_n) = (n, c[n^{0.325}], [\exp(n^{0.35})])$ for $n = 100, 200, 500$ and $1,000$, the value of c controls the extent of sparsity, which results in the table below:

n	100	200	500	1,000
$[n^{0.325}]$	4	6	8	9
p_n	150	595	6,655	74,622

For $j \in s_{0n}$, the parameter $\boldsymbol{\beta}_{0j}$ is independently generated as $\boldsymbol{\beta}_{0j} = (-1)^u(n^{-0.1625} + |z|)$ where $u \sim$ Bernoulli(0.4) and z is a normal random variable with mean 0 and satisfies $P(|z| \geq 0.1) = 0.25$. This ensures, roughly, $\min\{|\boldsymbol{\beta}_{0j}| : j \in s_{0n}\} = O(n^{-0.1625})$. The error variance σ^2 is determined by setting the following ratio to

certain values when $n = 100$ and kept unchanged for other n's:

$$h = \frac{\boldsymbol{E}\left(\boldsymbol{\beta}_0{}^{\tau}\Sigma\boldsymbol{\beta}_0\right)}{\boldsymbol{E}\left(\boldsymbol{\beta}_0{}^{\tau}\Sigma\boldsymbol{\beta}_0\right) + \sigma^2}, \tag{3.2.1}$$

where $\boldsymbol{\beta}_0$ is the true parameter vector and Σ is the covariance matrix of the predictors. This ratio is called the heritability in broad sense in genetic studies if the response is a quantitative trait and the covariates are genotypes of quantitative trait loci. The higher the h, the easier for the relevant features to be detected. In our simulations, we let h be $0.4, 0.6, 0.8$. The following three correlation structures are considered for the covariates:

Structure I: Power decay correlation. The covariates are generated as series of normally distributed random variables with mean 0 and correlation coefficient $\rho_{ij} = 0.5^{|i-j|}$.

Structure II: Diagonal block design with equal pairwise correlation. The covariance matrix is a diagonal block matrix. Each block except the last one is of dimension 50×50. The variances in the blocks are all equal to 1 and the off-diagonal correlations are all equal to $\rho = 0.5$.

Structure III: Diagonal block design with uniformly distributed eigenvalues. The

covariance matrix of all the covariates is of the form

$$\Sigma = \begin{pmatrix} B & \dots & \dots & \dots \\ \dots & B & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & B \end{pmatrix}.$$

The block matrix B is of dimension 50×50 and is generated by the following steps: (1) A positive definite matrix is generated such that it has the smallest eigenvalue 1, the largest eigenvalue 20 and the other eigenvalues uniformly distributed between 1 and 20; (2) the matrix is converted into a correlation matrix by dividing its entries with the square roots of its diagonal elements; (3) B is taken as the correlation matrix.

For each simulation setting, the PDR_n and FDR_n averaged over 200 replicates and their standard deviations in the parenthesis are reported in Tables 3.2.1, 3.2.2 and 3.2.3. We can make the following conclusions from results in Tables 3.2.1-3.2.3: (i) The BIC (EBIC with $\gamma = 0$) does not appear to be selection consistent. Under all the settings, the FDR_n of the procedure with BIC does not reduce as n increases, it is in fact the opposite. (ii) The finite sample performance of the EBIC closely matches its asymptotic property. That is, under all the three correlation structures, for the procedure with EBIC with $\gamma = 1 - \dfrac{\ln n}{4 \ln p_n}$, the PDR_n and the

FDR_n approach rapidly to 1 and 0 respectively, as n increases from 100 to 1000, at all the three h levels. In general, the PDR_n of the procedure with BIC is higher because it always selects much more features. But, as n gets large, the PDR_n of EBIC with $\gamma = 1 - \dfrac{\ln n}{4 \ln p_n}$ quickly becomes comparable with that of the BIC. (iii) For large n, the mBIC (EBIC with $\gamma = 1$) is comparable with EBIC with $\gamma = 1 - \dfrac{\ln n}{4 \ln p_n}$, which reflects the fact that it is also selection consistent. But for small n, it loses certain power while overly controlling FDR_n.

Table 3.2.1 Results on the SIS-SCAD-EBIC Procedure with Structure I in LMs

		PDR_n			FDR_n		
n	h	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
$c=1$							
100	0.4	.726 (.242)	.449(.291)	.384(.288)	.571 (.212)	.074 (.205)	.050 (.181)
	0.6	.861(.187)	.700(.271)	.633(.301)	.478(.216)	.080(.169)	.044 (.123)
	0.8	.973(.089)	.921 (.159)	.909(.176)	.363(.204)	.0849 (.147)	.056(.119)
200	0.4	.759 (.205)	.532(.269)	.467(.269)	.662 (.177)	.034(.101)	.017(.071)
	0.6	.909(.144)	.758 (.256)	.711(.282)	.574(.185)	.079(.145)	.038(.098)
	0.8	.989(.056)	.957(.105)	.947(.128)	.389(.200)	.060(.115)	.045(.105)
500	0.4	.826 (.146)	.640(.212)	.604 (.214)	.768 (.099)	.037 (.089)	.011(.046)
	0.6	.943(.099)	.863(.164)	.836(.181)	.659(.133)	.066(.128)	.028(.079)
	0.8	.994 (.035)	.983(.060)	.980 (.067)	.504(.189)	.027(.073)	.019(.065)
1,000	0.4	1.00(.000)	.999 (.008)	.999 (.011)	.662(.024)	.019(.041)	.009 (.028)
	0.6	1.00 (.000)	1.00(.000)	1.00(.000)	.531(.037)	.019(.041)	.008(.026)
	0.8	1.00(.000)	1.00(.000)	1.00(.000)	.469(.009)	.007(.025)	.002(.014)
$c=2$							
100	0.4	.531(.183)	.243(.169)	.198(.162)	.507(.222)	.069 (.204)	.041(.172)
	0.6	.679(.166)	.416 (.213)	.349(.206)	.447(.187)	.074(.173)	.026(.093)
	0.8	.850(.153)	.708 (.225)	.628(.248)	.373(.163)	.118 (.143)	.068(.118)
200	0.4	.613(.162)	.306 (.164)	.260(.161)	.619(.162)	.028(.096)	.009 (.066)
	0.6	.720(.148)	.518(.211)	.456(.207)	.545(.181)	.036(.082)	.018 (.061)
	0.8	.895(.125)	.745(.199)	.703(.217)	.447(.164)	.086(.117)	.053(.096)
5,00	0.4	.732(.129)	.425(.174)	.371(.166)	.774(.076)	.014(.054)	.004(.025)
	0.6	.832(.104)	.635(.176)	.589(.186)	.695(.112)	.028(.064)	.009(.031)
	0.8	.956(.067)	.875(.135)	.847(.157)	.535(.159)	.098(.121)	.068(.104)
10,00	0.4	.758 (.108)	.537(.161)	.491(.164)	.825(.055)	.012(.039)	.005(.025)
	0.6	.849(.102)	.715(.134)	.689(.144)	.761(.077)	.025(.062)	.009 (.035)
	0.8	.969(.054)	.925(.084)	.906 (.106)	.581 (.146)	.095(.109)	.072(.095)

The values of γ in EBIC_γ : $\gamma_1 = 0,\ \gamma_2 = 1 - \dfrac{\ln n}{4 \ln p_n},\ \gamma_3 = 1.$

Table 3.2.2 Results on the SIS-SCAD-EBIC Procedure with Structure II in LMs

		PDR_n			FDR_n		
n	h	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
$c=1$							
100	0.4	.733(.285)	.402(.318)	.343(.291)	.427(.268)	.229(.369)	.198 (.362)
	0.6	.933(.154)	.772(.297)	.703(.321)	.339 (.213)	.117(.197)	.094(.207)
	0.8	.996(.042)	.967(.118)	.960 (.125)	.293(.203)	.053(.132)	.036(.114)
200	0.4	.868 (.203)	.534(.303)	.479(.306)	.442(.206)	.133 (.249)	.109(.246)
	0.6	.994(.039)	.931(.168)	.889(.214)	.321(.173)	.107 (.161)	.078 (.143)
	0.8	1.00(.000)	.996(.031)	.994 (.039)	.292(.165)	.025(.081)	.017(.069)
500	0.4	.948 (.093)	.754 (.178)	.723 (.184)	.689(.114)	.056(.107)	.049 (.103)
	0.6	.993(.035)	.922(.121)	.904(.132)	.626 (.127)	.031(.080)	.019(.064)
	0.8	1.00(.000)	.997 (.024)	.992(.044)	.585(.151)	.059(.109)	.031 (.083)
1,000	0.4	.939(.080)	.813(.158)	.785(.180)	.818(.046)	.073(.113)	.049(.092)
	0.6	.995 (.025)	.988 (.041)	.986(.043)	.739(.066)	.039 (.084)	.035(.079)
	0.8	.999(.010)	.998(.017)	.996(.024)	.653(.107)	.024(.069)	.017(.061)
$c=2$							
100	0.4	.430(.239)	.193(.174)	.173(.164)	.449(.294)	.310(.411)	.295(.408)
	0.6	.684 (.234)	.389(.236)	.343(.224)	.343(.220)	.164(.235)	.150(.253)
	0.8	.881(.179)	.676(.266)	.603(.284)	.308(.194)	.105(.174)	.096(.175)
200	0.4	.489(.206)	.199(.142)	.165(.133)	.416(.235)	.134 (.275)	.115(.259)
	0.6	.727(.192)	.421(.227)	.356(.214)	.351(.195)	.065(.144)	.055(.132)
	0.8	.919(.135)	.718(.254)	.672(.269)	.351(.184)	.055(.099)	.043(.088)
5,00	0.4	.664(.137)	.258(.132)	.238(.132)	.669(.145)	.031(.099)	.020(.076)
	0.6	.834(.127)	.468(.211)	.407(.209)	.609(.132)	.029(.073)	.014(.047)
	0.8	.944(.094)	.804(.244)	.778(.266)	.485(.198)	.084(.108)	.068(.095)
1,000	0.4	.675 (.133)	.311(.158)	.284(.158)	.829(.079)	.017(.055)	.014(.050)
	0.6	.882(.134)	.551(.234)	.496(.239)	.744(.115)	.060(.108)	.033(.073)
	0.8	.959 (.078)	.884(.195)	.877(.202)	.616(.178)	.069(.099)	.061 (.087)

The values of γ in EBIC_γ : $\gamma_1 = 0,\ \gamma_2 = 1 - \dfrac{\ln n}{4 \ln p_n},\ \gamma_3 = 1.$

Table 3.2.3 Results on the SIS-SCAD-EBIC Procedure with Structure III in LMs

		PDR_n			FDR_n		
n	h	γ_1	γ_2	γ_3	γ_1	γ_2	γ_3
$c=1$							
100	0.4	.915(.146)	.667(.302)	.564(.327)	.428(.191)	.041(.102)	.020(.078)
	0.6	.996(.031)	.964(.116)	.950(.133)	.360(.181)	.046(.105)	.019 (.063)
	0.8	1.00(.000)	1.00(.000)	1.00(.000)	0.326(.165)	.038(.096)	.011(.051)
200	0.4	.993(.037)	.865(.206)	.811(.252)	.575(.162)	.049(.101)	.024(.073)
	0.6	1.00(.000)	.999(.014)	.999(.014)	.536(.129)	.032(.081)	.013(.048)
	0.8	1.00(.000)	1.00(.000)	1.00(.000)	.457(.138)	.023(.065)	.009 (.042)
500	0.4	1.00(.000)	.971(.081)	.961(.090)	.768(.042)	.041(.075)	.023 (.055)
	0.6	1.00 (.000)	1.00(.000)	1.00(.000)	.704 (.058)	.022(.059)	.010 (.043)
	0.8	1.00(.000)	1.00(.000)	1.00(.000)	.608 (.091)	.016(.049)	.007(.038)
1,000	0.4	1.00(.000)	.999 (.011)	.997 (.017)	.790(.040)	.023(.046)	.008(.028)
	0.6	1.00(.000)	1.00(.000)	1.00(.000)	.740(.038)	.018(.041)	.005 (.021)
	0.8	1.00 (.000)	1.00(.000)	1.00(.000)	.705(.051)	.005(.022)	.002 (.012)
$c=2$							
100	0.4	.643 (.218)	.239(.201)	.155 (.179)	.409(.206)	.071(.185)	.028(.128)
	0.6	.911 (.141)	.589(.298)	.461(.302)	.346(.168)	.092(.163)	.045(.129)
	0.8	.995(.033)	.975 (.100)	.964 (.135)	.237(.136)	.089(.101)	.069 (.092)
200	0.4	.801(.147)	.307(.210)	.209(.179)	.536(.136)	.049 (.142)	.013 (.061)
	0.6	.974(.063)	.817(.198)	.742(.236)	.443(.147)	.076(.095)	.045 (.073)
	0.8	.999(.010)	.993(.041)	.989(.048)	.322 (.121)	.046 (.074)	.034(.063)
500	0.4	.933 (.076)	.578 (.204)	.451 (.215)	.723(.079)	.035 (.067)	.009 (.036)
	0.6	.992(.029)	.946 (.073)	.929(.094)	.642 (.091)	.062(.078)	.045(.069)
	0.8	.999(.005)	.998(.016)	.997(.017)	.498(.105)	.023(.044)	.014(.036)
1,000	0.4	.969(.049)	.779(.169)	.688(.207)	.809(.051)	.042 (.063)	.018(.039)
	0.6	.997(.013)	.976 (.054)	.973(.058)	.738(.059)	.029(.053)	.024(.042)
	0.8	.999(.004)	.998 (.011)	.998 (.012)	.608(.085)	.011(.031)	.006(.022)

The values of γ in EBIC_γ : $\gamma_1 = 0,\ \gamma_2 = 1 - \dfrac{\ln n}{4 \ln p_n},\ \gamma_3 = 1.$

CHAPTER 4

EBIC in Generalized Linear Regression Models

Generalized linear regression models (GLMs) are much more flexible in describing the relationship between a given response variable and the predictors. Feature selection in GLMs becomes naturally important in high-dimensional studies. The selection consistency of the EBIC for GLMs with canonical links was established in [35]. As pointed out in [48], the canonical link usually fails to best fit a given data set. In this chapter, we check the validity of the EBIC for feature selection in GLMs with general links in small-n-large-p problems and state our main result

in Theorem 4.1.1.

4.1 Selection Consistency of EBIC

Let $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, be the observations, where y_i is a response variable and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_n})^\tau$ is a p_n-vector of covariates. We consider the generalized linear model below:

$$y_i \sim f(y_i; \theta_i) = \exp\{\theta_i y_i - b(\theta_i)\} \text{ w.r.t. } \nu, \quad i = 1, \ldots, n,$$

where ν is a σ-finite measure. From the properties of exponential family, we have

$$\mu(\theta_i) = E(y_i) = b^{'}(\theta_i), \ \sigma^2(\theta_i) = \text{Var}(y_i) = b^{''}(\theta_i),$$

where $b^{'}$ and $b^{''}$ are the first and the second derivatives of b respectively. The θ_i is related to $\boldsymbol{x}_i$ through the relationship:

$$g(\mu(\theta_i)) = \eta_i = \boldsymbol{x}_i^\tau \boldsymbol{\beta},$$

where g is a monotone function called link function and $\boldsymbol{\beta}$ is a p_n-dimensional parameter vector. If $g(\mu(\theta_i)) = \theta_i$, i.e., $g = \mu^{-1}$, the link is called the canonical

link. Here, we consider general link functions including the canonical link. Because of the one-to-one correspondence between θ_i and η_i, there is a function h such that $\theta_i = h(\eta_i) = h(\boldsymbol{x}_i^\tau \boldsymbol{\beta})$. Thus the probability density function of y_i can be expressed as

$$f(y_i; h(\boldsymbol{x}_i^\tau \boldsymbol{\beta})) = \exp\{y_i h(\boldsymbol{x}_i^\tau \boldsymbol{\beta}) - b(h(\boldsymbol{x}_i^\tau \boldsymbol{\beta}))\}.$$

Suppose that b and g are thrice and twice differentiable respectively, which is usually the case in practical GLMs, then h is twice differentiable. Suppose its $l-$order derivatives $h^{(l)}(l = 0, 1, 2)$ exist and they are continuous and bounded. Then

$$b^{'}\left(h(X_i^\tau \boldsymbol{\beta}_0)\right) = \mu_i = \mathrm{E}(y_i), \quad b^{(2)}\left(h(X_i^\tau \boldsymbol{\beta}_0)\right) = \sigma_i^2 = \mathrm{Var}(y_i).$$

For canonical link, $h^{(1)} = 1,\ h^{(2)} = 0$.

In the above GLMs, we assume that $p_n = O(\exp\{n^\kappa\})$ for some $0 < \kappa < 1$, and that only a relatively small number of components of $\boldsymbol{\beta}$ are nonzero. Throughout the article, the following notation and convention are used. Denote by s any subset of the index set $\mathscr{S} = \{1, 2, \ldots, p_n\}$ and $|s|$ its cardinality. For convenience, s is used exchangeably to denote both an index set and the set of covariates with indices in the index set, and is also referred to as a model, i.e., the GLMs consisting of the covariates in s. Let $s_{0n} = \{j : \boldsymbol{\beta}_{0j} \neq 0, j = 1, \ldots, p_n\}$ and $p_{0n} = |s_{0n}|$. The covariates belonging to s_{0n} are called relevant features and the others irrelevant

features. s_{0n} is also referred to as the true model. Let X_i be the observation vector for the ith individual and $X_i(s)$ be its component which includes the covariates in s, and let $\boldsymbol{\beta}(s)$ be the corresponding sub vector of $\boldsymbol{\beta}$. Let S_j denote the set of $\binom{p_n}{j}$ combinations of j indices from $\mathscr{S}$. Denote $\tau(S_j) = \binom{p_n}{j}$.

The EBIC of a model s, as defined in [33], is

$$\mathrm{EBIC}_\gamma(s) = -2\ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) + |s|\ln n + 2\gamma\ln\tau(S_{|s|}), \quad \gamma \geq 0,$$

where $L_n(\hat{\boldsymbol{\beta}}(s))$ is the maximum likelihood of model s and $\hat{\boldsymbol{\beta}}(s)$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}(s)$.

Denote $\mathscr{A}_0 = \{s : s_{0n} \subsetneq s; |s| \leq kp_{0n}\}, \mathscr{A}_1 = \{s : s_{0n} \nsubseteq s; |s| \leq kp_{0n}\}$ where $k > 1$ and

$$\begin{aligned}
l_n(\boldsymbol{\beta}) =& \sum_{i=1}^{n} \ln f(y_i, \theta_i) = \sum_{i=1}^{n} \left(y_i\theta_i - b(\theta_i)\right) = \sum_{i=1}^{n} \left(y_i h(X_i^\tau\boldsymbol{\beta}) - b\left(h(X_i^\tau\boldsymbol{\beta})\right)\right), \\
s_n(\boldsymbol{\beta}) =& \frac{\partial l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n} \left(y_i - b'\left(h(X_i^\tau\boldsymbol{\beta})\right)\right) h^{(1)}(X_i^\tau\boldsymbol{\beta}) X_i, \\
H_n^E(\boldsymbol{\beta}) =& \sum_{i=1}^{n} b^{(2)}\left(h(X_i^\tau\boldsymbol{\beta})\right)\left(h^{(1)}(X_i^\tau\boldsymbol{\beta})\right)^2 X_i X_i^\tau, \\
H_n^e(\boldsymbol{\beta}) =& \sum_{i=1}^{n} \left(y_i - b'\left(h(X_i^\tau\boldsymbol{\beta})\right)\right) h^{(2)}(X_i^\tau\boldsymbol{\beta}) X_i X_i^\tau, \\
H_0(\boldsymbol{\beta}) =& -\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau} = H_n^E(\boldsymbol{\beta}) - H_n^e(\boldsymbol{\beta}).
\end{aligned} \tag{4.1.1}$$

When $s_{0n} \subseteq s$,

$$\begin{aligned} s_n(\boldsymbol{\beta}_0(s)) &= \sum_{i=1}^{n} (y_i - \mu_i) h^{(1)}(X_i^\tau \boldsymbol{\beta}_0) X_i(s), \\ H_n^E(\boldsymbol{\beta}_0(s)) &= \sum_{i=1}^{n} \sigma_i^2 \left(h^{(1)}(X_i^\tau \boldsymbol{\beta}_0) \right)^2 X_i(s) X_i^\tau(s), \\ H_n^e(\boldsymbol{\beta}_0(s)) &= \sum_{i=1}^{n} (y_i - \mu_i) h^{(2)}(X_i^\tau \boldsymbol{\beta}_0) X_i(s) X_i^\tau(s). \end{aligned} \tag{4.1.2}$$

The following assumptions are imposed for the selection consistency of EBIC. Except C1, all the other assumptions are almost similar to those in [35] when the canonical link is considered.

C1 $\ln(p_n) = O(n^\kappa), p_{0n} = O(n^b)$ where $b \geq 0, \kappa > 0$ and $b + \kappa < 1/3$;

C2 $\min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}| \geq C n^{-1/4}$ for some constant $C > 0$;

C3 For any s, the interior of $\mathscr{B}(s) = \{\boldsymbol{\beta} : \int \exp(h(X_i^\tau(s)\boldsymbol{\beta}) y) d\nu < \infty, i = 1, 2, \ldots, n\}$ is nonempty. Let $\boldsymbol{\beta}_0$ denote the true parameter of the GLMs. If $|s| \leq k p_{0n}$, where $k > 1$, then $\boldsymbol{\beta}_0(s)$ is in the interior of $\mathscr{B}(s)$.

C4 There exist positive c_1 and c_2 such that for all sufficiently large n,

$$c_1 n \leq \lambda_{\min}(H_n^E(\boldsymbol{\beta}_0(s \cup s_{0n}))) \leq \lambda_{\max}(H_n^E(\boldsymbol{\beta}_0(s \cup s_{0n}))) \leq c_2 n$$

for all s with $|s| \leq k p_{0n}$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote respectively the smallest

and largest eigenvalues;

C5 For any given $\xi > 0$, there exists a $\delta > 0$ such that when n is sufficiently large,

$$(1-\xi)H_n^E(\boldsymbol{\beta}_0(s \cup s_{0n})) \leq H_n^E(\boldsymbol{\beta}(s \cup s_{0n})) \leq (1+\xi)H_n^E(\boldsymbol{\beta}_0(s \cup s_{0n})),$$

$$(1-\xi)H_n^e(\boldsymbol{\beta}_0(s \cup s_{0n})) \leq H_n^e(\boldsymbol{\beta}(s \cup s_{0n})) \leq (1+\xi)H_n^e(\boldsymbol{\beta}_0(s \cup s_{0n}))$$

whenever $\|\boldsymbol{\beta}(s \cup s_{0n}) - \boldsymbol{\beta}_0(s \cup s_{0n})\|_2 \leq \delta$ for all s with $|s| \leq kp_{0n}$;

C6 The quantities $|x_{ij}|, |h^{'}(X_i^\tau \boldsymbol{\beta}_0)|, |h^{''}(X_i^\tau \boldsymbol{\beta}_0)|, i = 1, \ldots, n; j = 1, \ldots, p_n$ are bounded from above, and $\sigma_i^2, i = 1, \ldots, n$ are bounded both from above and below away from zero. Furthermore,

$$\max_{1 \leq j \leq p_n; 1 \leq i \leq n} \frac{x_{ij}^2 [h^{'}(X_i^\tau \boldsymbol{\beta}_0)]^2}{\sum_{i=1}^n \sigma_i^2 x_{ij}^2 [h^{'}(X_i^\tau \boldsymbol{\beta}_0)]^2} = o(n^{-1/3}),$$

$$\max_{1 \leq i \leq n} \frac{[h^{''}(X_i^\tau \boldsymbol{\beta}_0)]^2}{\sum_{i=1}^n \sigma_i^2 [h^{''}(X_i^\tau \boldsymbol{\beta}_0)]^2} = o(n^{-1/3}).$$

The positive definiteness of the information matrix $H_n(\boldsymbol{\beta})$ is fulfilled naturally for canonical links but not definitely for non-canonical links. Readers can find a thorough relevant study in [169]. This assumption is regular and can guarantee the existence and uniqueness of $\hat{\boldsymbol{\beta}}(s)$ for all s with $|s| \leq kp_{0n}$ where $k > 1$. For non-canonical links, C6 are easily satisfied by all the examples given in [169]. The verification of C6 is given in the Appendix.

We now state our main result in the following theorem:

Theorem 4.1.1. *Under Conditions C1-C6, as $n \to +\infty$, we have*

(1) $P\left(\min_{s\in\mathscr{A}_1} EBIC_\gamma(s) \le EBIC_\gamma(s_{0n})\right) \to 0$ *for any* $\gamma > 0$*;*

(2) $P\left(\min_{s\in\mathscr{A}_0} EBIC_\gamma(s) \le EBIC_\gamma(s_{0n})\right) \to 0$ *for any* $\gamma > \dfrac{1}{1-\xi}\left[1 - \dfrac{\ln n}{2\ln p_n}\right]$, *where* ξ *is an arbitrarily small positive constant.*

The proof of this theorem requires the following corollaries of Lemma 1 in [35], which is stated as follows. To avoid redundancy, this lemma is referred to Lemma 1 unless otherwise stated.

Lemma 1: Let $Y_i, i = 1, 2, \cdots, n$ be independent random variables following exponential family distributions with natural parameters θ_i. Let μ_i, σ_i^2 denote the mean and variance of Y_i respectively. Suppose that $\{\theta_i : i = 1, 2, \cdots, n\}$ is contained in a compact subset of the natural space Θ. Let $a_{ni}, i = 1, 2, \cdots, n$ be real numbers such that $\sum_{i=1}^{n} a_{ni}^2\sigma_i^2 = 1$ and $\max_{1\le i\le n}\{|a_{ni}|\} = o(n^{-1/6})$. Then for any $m = O(n^{1/3})$, we have

$$P(\sum_{i=1}^{n} a_{ni}(y_i - \boldsymbol{\mu}_i) > \sqrt{2m}) \le \exp(-m(1-\xi))$$

for any positive ξ when n is sufficiently large.

Corollary 4.1.1. *Under Conditions C1-C6,*

$$P(\max_{s\in\mathscr{A}_0, j\in s} s_{n,j}^2(\boldsymbol{\beta}_0(s)) \geq Cn^{4/3}) = o(1).$$

Proof of Corollary 4.1.1: Let $a_{ni} = x_{i,j}h^{(1)}\left(X_i^\tau(s)\boldsymbol{\beta}_0(s)\right) / \sqrt{\sum_{i=1}^n \sigma_i^2 x_{i,j}^2 \left(h^{(1)}(X_i^\tau(s)\boldsymbol{\beta}_0(s))\right)^2}$, when $s \in \mathscr{A}_0, X_i^\tau(s)\boldsymbol{\beta}_0(s) = X_i^\tau\boldsymbol{\beta}_0$, From Lemma 1 and C6, we have

$$\begin{aligned} P\left(s_{nj}\left(\boldsymbol{\beta}_0(s)\right) \geq Cn^{2/3}\right) =& P\left(\sum_{i=1}^n a_{ni}(y_i - \boldsymbol{\mu}_i) > Cn^{2/3} / \sqrt{\sum_{i=1}^n \sigma_i^2 x_{i,j}^2 \left(h^{(1)}(X_i^\tau\boldsymbol{\beta}_0)\right)^2}\right) \\ \leq& P\left(\sum_{i=1}^n a_{ni}(y_i - \boldsymbol{\mu}_i) > Cn^{1/6}\right) \leq \exp(-Cn^{1/3}). \end{aligned}$$

The first inequality holds because of the boundedness of $x_{i,j}$ and $h^{(1)}$. Consequently, when $kp_{0n}\ln p_n + \ln p_{0n} = o(n^{1/3})$, which is satisfied by C1, we have

$$\begin{aligned} \sum_{s\in\mathscr{A}_0}\sum_{j\in s} P\left(s_{nj}\left(\boldsymbol{\beta}_0(s)\right) \geq Cn^{2/3}\right) &\leq kp_{0n}p_n^{kp_{0n}}\exp(-Cn^{1/3}) \\ &= k\exp(\ln p_{0n} + kp_{0n}\ln p_n - Cn^{1/3}) = o(1). \end{aligned}$$

That is,

$$P\left(\max_{s\in\mathscr{A}_0, j\in s} s_{n,j}^2\left(\boldsymbol{\beta}_0(s)\right) \geq Cn^{4/3}\right) = o(1). \tag{4.1.3}$$

□

Corollary 4.1.2. *Under Conditions C1-C6, we have*

$$P\left(\max_{s\in\mathscr{A}_0,\|\boldsymbol{u}\|_2=1,\|\boldsymbol{\beta}(s)-\boldsymbol{\beta}_0(s)\|_2\le\delta}\boldsymbol{u}^\tau H_n^e(\boldsymbol{\beta}(s))\boldsymbol{u}\ge Cp_{0n}n^{2/3}\right)=o(1),$$
$$P\left(\max_{s\in\mathscr{A}_1,\|\boldsymbol{u}\|_2=1,\|\boldsymbol{\beta}(s\cup s_{0n})-\boldsymbol{\beta}_0(s\cup s_{0n})\|_2\le\delta}\boldsymbol{u}^\tau H_n^e(\boldsymbol{\beta}(s\cup s_{0n}))\boldsymbol{u}\ge Cp_{0n}n^{2/3}\right)=o(1). \tag{4.1.4}$$

Proof of Corollary 4.1.2: Since $\mathscr{A}_0=\{s\cup s_{0n}: s\in\mathscr{A}_1, 0<|s|\le(k-1)p_{0n}\}$, for all $s\in\mathscr{A}_1$, consider $\tilde{s}=s\cup s_{0n}$. Let

$$a_{ni}=h^{(2)}\left(X_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s})\right)\operatorname{sign}(y_i-\mu_i)/\sqrt{\sum_{i=1}^n\sigma_i^2\left(h^{(2)}\left(X_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s})\right)\right)^2},$$

since $X_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s})=X_i^\tau\boldsymbol{\beta}_0$, from Lemma 1 and C6, we have

$$P\left(\sum_{i=1}^n\left|(y_i-\mu_i)h^{(2)}\left(X_i^\tau(\tilde{s})\boldsymbol{\beta}(\tilde{s})\right)\right|\ge Cn^{2/3}\right)\le 2\exp(-Cn^{1/3}).$$

For any unit vector $\boldsymbol{u}$ with length $|\tilde{s}|$,

$$\begin{aligned}\boldsymbol{u}^\tau H_n^e(\boldsymbol{\beta}_0(\tilde{s}))\boldsymbol{u}=&\sum_{i=1}^n(y_i-\mu_i)h^{(2)}\left(X_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s})\right)\left(\boldsymbol{u}^\tau X_i(\tilde{s})\right)^2\\ \le&\sum_{i=1}^n\left|(y_i-\mu_i)h^{(2)}\left(X_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s})\right)\right|\|X_i(\tilde{s})\|_2^2\\ \le&Ckp_{0n}\sum_{i=1}^n\left|(y_i-\mu_i)h^{(2)}\left(X_i^\tau(\tilde{s})\boldsymbol{\beta}_0(\tilde{s})\right)\right|\end{aligned}$$

The last inequality holds because all $x'_{i,j}s$ are bounded. Therefore,

$$\begin{aligned}
&P\left(\max_{s\in\mathscr{A}_0,\|\boldsymbol{u}\|_2=1,\|\boldsymbol{\beta}(s)-\boldsymbol{\beta}_0(s)\|_2\leq\delta}\boldsymbol{u}^\tau H_n^e\left(\boldsymbol{\beta}(s)\right)\boldsymbol{u}\geq Cp_{0n}n^{2/3}\right)\\
\leq&P\left(\max_{s\in\mathscr{A}_0,\|\boldsymbol{u}\|_2=1}\boldsymbol{u}^\tau H_n^e\left(\boldsymbol{\beta}_0(s)\right)\boldsymbol{u}\geq C(1+\xi)^{-1}p_{0n}n^{2/3}\right)\\
\leq&2\exp(kp_{on}\ln p_n-Cn^{1/3})=o(1).
\end{aligned}$$

Similarly, we can derive the second inequality for the case $s\in\mathscr{A}_1$. □

Corollary 4.1.3. *Under Conditions C1-C6, for any* $s\in\mathscr{A}_1,\|\boldsymbol{u}\|_2=1,\dim(\boldsymbol{u})=|s\cup s_{0n}|$, *uniformly, when* $\|\boldsymbol{\beta}(s\cup s_{0n})-\boldsymbol{\beta}_0(s\cup s_{0n})\|_2\leq\delta$,

$$\boldsymbol{u}^\tau H_n\left(\boldsymbol{\beta}(s\cup s_{0n})\right)\boldsymbol{u}=\boldsymbol{u}^\tau H_n^E\left(\boldsymbol{\beta}(s\cup s_{0n})\right)\boldsymbol{u}\left(1+o_p(1)\right). \tag{4.1.5}$$

This is true when $s\cup s_{0n}$ *is replaced by* $s,\forall s\in\mathscr{A}_0$.

Proof of Corollary 4.1.3: This corollary can be seen from Corollary 4.1.2 and assumption C4. □

This corollary is important in connecting general link functions to canonical link functions. In a neighborhood of the true parameter $\boldsymbol{\beta}_0, H_n^e(\boldsymbol{\beta}(s))$ is negligible in $H_n(\boldsymbol{\beta}(s))$, which implies that $H_n(\boldsymbol{\beta}(s))$ is asymptotically locally positive definite.

Theorem 4.1.2. *Under Conditions C1-C6, as* $n\to+\infty,\|\hat{\boldsymbol{\beta}}(s)-\boldsymbol{\beta}_0(s)\|_2=$

$O_p(n^{-1/3})$, *uniformly for* $s \in \mathscr{A}_0$.

Proof of Theorem 4.1.2: For any unit vector u, let $\boldsymbol{\beta}(s) = \boldsymbol{\beta}_0(s) + n^{-1/3}\boldsymbol{u}$. Denote

$$\mathcal{T} = \left\{ \max_{s \in \mathscr{A}_0, \|\boldsymbol{u}\|_2 = 1} \boldsymbol{u}^\tau H_n^e \left(\boldsymbol{\beta}(s)\right) \boldsymbol{u} \leq C p_{0n} n^{2/3} \right\},$$

then Corollary 4.1.2 implies

$$\begin{aligned} & P\left(L_n\left(\boldsymbol{\beta}(s)\right) - L_n\left(\boldsymbol{\beta}_0(s)\right) > 0 : \ \text{ for some } \ u, s \in \mathscr{A}_0\right) \\ =& P\left(L_n\left(\boldsymbol{\beta}(s)\right) - L_n\left(\boldsymbol{\beta}_0(s)\right) > 0 : \ \text{ for some } \ u, s \in \mathscr{A}_0; \mathcal{T}\right) + o(1). \end{aligned} \tag{4.1.6}$$

With $\mathcal{T}$, when n is large enough, for all $s \in \mathscr{A}_0$, uniformly, we have

$$\begin{aligned} L_n\left(\boldsymbol{\beta}(s)\right) - L_n\left(\boldsymbol{\beta}_0(s)\right) =& n^{-1/3}\boldsymbol{u}^\tau s_n\left(\boldsymbol{\beta}_0(s)\right) - \frac{1}{2} n^{1/3} \boldsymbol{u}^\tau \left(n^{-1} H_n^E \left(\tilde{\boldsymbol{\beta}}(s)\right)\right) \boldsymbol{u} \\ & - \frac{1}{2} n^{-2/3} \left(\boldsymbol{u}^\tau H_n^e \left(\tilde{\boldsymbol{\beta}}(s)\right) \boldsymbol{u}\right) \\ =& n^{-1/3}\boldsymbol{u}^\tau s_n\left(\boldsymbol{\beta}_0(s)\right) - c_1(1-\xi) n^{1/3}/2 + O(p_{0n}) \\ \leq& n^{-1/3}\boldsymbol{u}^\tau s_n\left(\boldsymbol{\beta}_0(s)\right) - c n^{1/3}. \end{aligned}$$

Hence, for some positive constant c, we have

$$\begin{aligned} & P\left(L_n\left(\boldsymbol{\beta}(s)\right) - L_n\left(\boldsymbol{\beta}_0(s)\right) > 0 \ \text{ for some } \ \boldsymbol{u}\right) \\ \leq & P\left(\boldsymbol{u}^\tau s_n\left(\boldsymbol{\beta}_0(s)\right) \geq c n^{2/3} \ \text{ for some } \ \boldsymbol{u}\right), \end{aligned}$$

which is less than

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

From Corollary 4.1.1, $\sum_{i\in\mathscr{A}_0}\sum_{j\in s} P\left(s_{n,j}\left(\boldsymbol{\beta}_0(s)\right) \geq cn^{2/3}\right) = o(1)$. The same for the second term. Therefore,

$$P\left(L_n\left(\boldsymbol{\beta}(s)\right) - L_n\left(\boldsymbol{\beta}_0(s)\right) > 0: \quad \text{for some } \boldsymbol{u}, s \in \mathscr{A}_0\right) = o(1). \tag{4.1.7}$$

Because $L_n\left(\boldsymbol{\beta}(s)\right)$ is a concave function for any s with probability tending to 1, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}(s)$ exists and falls within an $n^{-1/3}$ neighborhood of $\boldsymbol{\beta}_0(s)$ uniformly for $s \in \mathscr{A}_0$. Thus, we have $P\left(\|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2 = O(n^{-1/3})\right) \to 1$.

□

Proof of Theorem 4.1.1: According to the definition of EBIC, for any model s, $\mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})$ if and only if

$$\ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \geq (|s| - p_{0n})\ln n/2 + \gamma\left(\ln\tau(\mathcal{S}_{|s|}) - \ln\tau(\mathcal{S}_{p_{0n}})\right). \tag{4.1.8}$$

To prove the selection consistency of EBIC, or mathematically,

$$P\left(\min_{s:|s|\leq kp_{0n}, s\neq s_{0n}} \mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})\right) \to 0 \text{ as } n \to +\infty,$$

it suffices to show that inequality (4.1.8) holds with a probability converging to 0 as the sample size goes to infinity uniformly for all $s \in \mathscr{A}_0 \cup \mathscr{A}_1$. This is completed by dealing with $s \in \mathscr{A}_0$ and $\mathscr{A}_1$ separately.

(1) When $s \in \mathscr{A}_1$, inequality (4.1.8) implies that

$$\ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \geq -p_{0n}(\ln n/2 + \gamma \ln p_n). \tag{4.1.9}$$

Therefore, if we can show as $n \to +\infty$,

$$P\left(\sup_{s\in\mathscr{A}_1} \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \geq -p_{0n}(\ln n/2 + \gamma \ln p_n)\right) \to 0, \tag{4.1.10}$$

then we will have

$$P\left(\min_{s:s\in\mathscr{A}_1} \mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})\right) \to 0 \text{ as } n \to +\infty.$$

The key becomes to assess the order for $\sup_{s\in\mathscr{A}_1} \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right)$. For any $s \in \mathscr{A}_1$, let $\tilde{s} = s \cup s_{0n}$ and $\breve{\boldsymbol{\beta}}(\tilde{s})$ be $\hat{\boldsymbol{\beta}}(s)$ augmented with zeros corresponding to the elements in $\tilde{s}\backslash s$. It can be seen that

$$\ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) = \ln L_n\left(\boldsymbol{\beta}_0(s_{0n})\right) \leq \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right),\ \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) = \ln L_n\left(\breve{\boldsymbol{\beta}}(\tilde{s})\right),$$

which leads to

$$\sup_{s\in\mathscr{A}_1} \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \leq \sup_{s\in\mathscr{A}_1} \ln L_n\left(\breve{\boldsymbol{\beta}}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right). \tag{4.1.11}$$

And also

$$\|\breve{\boldsymbol{\beta}}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \geq \|\boldsymbol{\beta}_0(s_{0n}\backslash s)\|_2 > \min_{j\in s_{0n}}\{|\boldsymbol{\beta}_{0j}|\} > Cn^{-1/4}.$$

The asymptotic positive definiteness of $H_n(\boldsymbol{\beta})$, or the concavity of $\ln L_n(\boldsymbol{\beta}(\tilde{s}))$ in $\boldsymbol{\beta}(\tilde{s})$ implies

$$\begin{aligned}
&\sup_{s\in\mathscr{A}_1} \ln L_n\left(\breve{\boldsymbol{\beta}}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \\
\leq &\sup\{\ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) : \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \geq n^{-1/4}, s\in\mathscr{A}_1\} \\
\leq &\sup\{\ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) : \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = n^{-1/4}, s\in\mathscr{A}_1\}.
\end{aligned} \tag{4.1.12}$$

To derive the order of the right hand side in the above inequality, we take

the Taylor Expansion of $\ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right)$ as follows:

$$\begin{aligned}
&\ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \\
&= \left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau s_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) - \frac{1}{2}\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau H_n^E\left(\boldsymbol{\beta}^\star(\tilde{s})\right)\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right) \\
&+ \frac{1}{2}\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau H_n^e\left(\boldsymbol{\beta}^\star(\tilde{s})\right)\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)
\end{aligned} \tag{4.1.13}$$

where $\boldsymbol{\beta}^\star(\tilde{s})$ is between $\boldsymbol{\beta}(\tilde{s})$ and $\boldsymbol{\beta}_0(\tilde{s})$. By conditions C4 and C5,

$$\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau H_n^E\left(\boldsymbol{\beta}^\star(\tilde{s})\right)\left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right) \geq c_1 n(1-\xi)\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2^2.$$

Corollary 4.1.3 implies that, for any $\boldsymbol{\beta}(\tilde{s})$ such that $\|\boldsymbol{\beta}(s) - \boldsymbol{\beta}(s_{0n})\|_2 = n^{-1/4}$, uniformly, there exists $0 < c < c_1$ such that, with probability tending to 1 as n goes to $+\infty$,

$$\ln L_n\left(\boldsymbol{\beta}(\tilde{s})\right) - \ln L_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \leq n^{-1/4}\|s_n(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} - \frac{c}{2}n^{1/2}(1-\xi). \tag{4.1.14}$$

The uniform rate for the components in the score function $s_n(\boldsymbol{\beta}_0)$ in Corollary 4.1.1 implies, the right hand side of (4.1.14) is less than $c_1 n^{5/12} - c_2 n^{1/2}$,

which is less than $-Cn^{1/2}$ for some constant $C > 0$. Combined with inequalities (4.1.11) and (4.1.12) , this leads to

$$\sup_{s\in\mathscr{A}_1} \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \leq -Cn^{1/2}.$$

Since under C1, $p_{0n}\ln n = o(n^{1/3}), p_{0n}\ln p_n = o(n^{1/3})$, we have proved inequality (4.1.10).

(2) When $s \in \mathscr{A}_0$, let $m = |s| - |s_{0n}|$, Lemma 3.1.1 implies that, asymptotically, as $n \to +\infty$, $\text{EBIC}_\gamma(s) \leq \text{EBIC}_\gamma(s_{0n})$ if and only if

$$\ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \geq m[0.5\ln n + \gamma \ln p_n]. \tag{4.1.15}$$

Therefore, it suffices to show, as $n \to +\infty$,

$$P\left(\sup_{s\in\mathscr{A}_0} \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \geq m[0.5\ln n + \gamma \ln p_n]\right) \to 0 \tag{4.1.16}$$

to obtain

$$P\left(\min_{s:s\in\mathscr{A}_0} \text{EBIC}_\gamma(s) \leq \text{EBIC}_\gamma(s_{0n})\right) \to 0 \text{ as } n \to +\infty.$$

Note that Corollary 4.1.3 implies

$$\begin{aligned}
&\ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\hat{\boldsymbol{\beta}}(s_{0n})\right) \leq \ln L_n\left(\hat{\boldsymbol{\beta}}(s)\right) - \ln L_n\left(\boldsymbol{\beta}_0(s_{0n})\right)) \\
=&(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau s_n(\boldsymbol{\beta}_0(s)) - \frac{1}{2}(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau H_n(\tilde{\boldsymbol{\beta}}(s))(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)) \\
\leq&(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau s_n(\boldsymbol{\beta}_0(s)) - \frac{1-\epsilon}{2}(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau H_n^E(\tilde{\boldsymbol{\beta}}(s))(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)),
\end{aligned} \tag{4.1.17}$$

where ξ is any arbitrarily small positive constant. The applicability of C5 to simplify the right hand side of this inequality requires $\sup_{s\in\mathscr{A}_0} \|\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)\|_2$ be approaching 0 as n goes to infinity, which is already verified in Theorem 4.1.2. Now we can apply C5. The right hand side of (4.1.17) can be upper bounded by

$$\begin{aligned}
&(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau s_n(\boldsymbol{\beta}_0(s)) - \frac{(1-\xi)(1-\epsilon)}{2}(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s))^\tau H_n^E(\boldsymbol{\beta}_0(s))(\hat{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s)) \\
\leq&\frac{1}{2(1-\xi)} s_n^\tau(\boldsymbol{\beta}_0(s))\{H_n^E(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s)),
\end{aligned}$$

where ξ is an arbitrarily small positive value. Hence, the left hand side of

(4.1.16) is no more than

$$
\begin{aligned}
&P\left(\frac{1}{2(1-\xi)} s_n^{\tau}(\boldsymbol{\beta}_0(s))\{H_n^E(\boldsymbol{\beta}_0(s))\}^{-1} s_n(\boldsymbol{\beta}_0(s)) \geq m[0.5\ln n + \gamma \ln p_n]\right) \\
\leq & |\mathscr{A}_0| \exp(-m(1-\xi)[0.5\ln n + \gamma \ln p_n]) \\
\leq & \exp\left(m[(\ln(p_n - p_{0n}) - (1-\xi)\gamma \ln p_n - \frac{(1-\xi)}{2}\ln n]\right).
\end{aligned}
\tag{4.1.18}
$$

It converges to 0 when $\gamma > \frac{1}{1-\xi}\left[1 - \frac{\ln n}{2\ln p_n}\right]$.

□

4.2 Numerical Study

Simulation Results

In this subsection, we aim to evaluate the performance of a two-stage procedure in GLMs with non-canonical links. It was shown in [35] that the EBIC is selection consistent for GLMs with canonical links. As a complementary work, we have theoretically verified the selection consistency of the EBIC in the presence of non-canonical links in Section 4.1.

Following LMs, the studies on screening and penalized likelihood procedures in feature selection in GLMs have been accomplished recently. However, to the best of our knowledge, the realization of regularization approaches such as adaptive LASSO and SCAD for GLMs with non-canonical links is unavailable. At this stage, we replace them by Forward Selection, where at each step, the variable leading to the greatest increment of the log likelihood (or equivalently, the greatest decrease of EBIC) is added into the model. The procedure continues until the total number of covariates reaches an empirically selected value. Since exhaustive searching is involved, when p_n is above 1000, the sure independence screening procedure based on the maximum marginal estimator (MMLE) proposed in [65] is applied to conduct dimension reduction before the Forward Selection. A sequence of nested models is hence generated. The one with the minimum EBIC among the model sequence is recognized as the best set of relevant features. This procedure merely requires greedy fitting of the GLMs, which can be obtained by the `glm.fit` function in R.

We consider the ultra-high dimensional feature space with diverging number of true features. Specifically, $(n, p_n, p_{0n}) = (n, [40\exp(n^{0.2})], [5n^{0.1}])$ and $n = 100, 200, 500$ are assumed in our simulation. γ in EBIC are chosen to be $(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0, 0.5 - \frac{\ln n}{4\ln p_n}, 1 - \frac{\ln n}{4\ln p_n}, 1)$. We considered binomial data with complementary log-log link function. That is, in population, given $X = x$, Y follows a binomial

distribution with probability of success $p(x) = 1 - \exp(-\exp(x^\tau \boldsymbol{\beta}_0))$. The settings of the covariates are adapted from $S1$ and $S3$ in [65].

Structure I: Let $q = 15$, which is much smaller than $[\frac{p_n}{3}]$, where $[x]$ denotes the largest integers not greater than x, denote $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{p_n}) = (\boldsymbol{X}_{i,j})$. $(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_q) \sim N(\mathbf{0}, \Sigma_\rho)$, where Σ_ρ stands for a $q \times q$ matrix with diagonal elements 1 and off-diagonal elements ρ. In our simulation study, $\rho = 0, 0.3, 0.5, 0.7$ are considered. $(\boldsymbol{X}_{q+1}, \cdots, \boldsymbol{X}_{[\frac{p_n}{3}]}) \sim N(\mathbf{0}, \boldsymbol{I})$. $\boldsymbol{X}_{i,j}, 1 \leq i \leq n, [\frac{p_n}{3}] + 1 \leq j \leq [\frac{2p_n}{3}]$ are i.i.d copies from Laplace distribution with location zero and scale 1. $\boldsymbol{X}_{i,j}, 1 \leq i \leq n, [\frac{2p_n}{3}] + 1 \leq j \leq p_n$ are i.i.d copies from a mixture normal distribution from $N(-1, 1), N(1, 0.5)$ with equal mixture proportion. The true coefficient vector $\boldsymbol{\beta}_0$ satisfies $\boldsymbol{\beta}_{0,L*j} = 1$ and 1.3 for odd and even $j \in \{1, 2, \cdots, p_{0n}\}$ respectively and 0 otherwise. Here $L = 10$.

Structure II: The only difference between setting 2 and setting 1 is $L = 5$. In setting 1, all the true features are statistically independent while in setting 2, three of them have pairwise linear correlation $\rho, \rho = 0.3, 0.5$ are considered.

Structure III: Let $q = 50, L = 10$, where q is much smaller than p_n and $p_n - q$ is much bigger than the maximum index for causal features, $L * p_{0n}$. Let the components of $\{\boldsymbol{X}_j\}_{j=1}^{p_n - q}$ and $\{\boldsymbol{\xi}_k\}_{k=p_n-q+1}^{p_n}$ be independent standard normal random

variables and

$$\boldsymbol{X}_k = \sum_{j=1}^{p_{0n}} \boldsymbol{X}_{L*j}(-1)^{j+1}/5 + \sqrt{25 - p_{0n}}/5\boldsymbol{\xi}_k,\ k = p_n - q + 1, \cdots, p_n.$$

The true coefficient vector $\boldsymbol{\beta}_0$ satisfies $\boldsymbol{\beta}_{0,L*j} = 1$ and 1.3 for odd and even $j \in \{1, 2, \cdots, p_{0n}\}$ respectively and 0 otherwise. In this setting, all the causal features are statistically independent, q highly correlated uncausal features have weak marginal but strong overall correlation with the causal features.

For each simulation setting, the PDR_n and FDR_n averaged over 200 replicates and their standard deviations in the parenthesis are reported in Tables 4.2.1, 4.2.2 and 4.2.3. The following conclusions can be made from the results in Tables 4.2.1, 4.2.2 and 4.2.3: (i) with all the four γ values, the PDR_n increases as n gets larger, (ii) with γ_1 and γ_2 (which are below the lower bound of the consistent range), the FDR_n does not show a trend to decrease while, with γ_3 and γ_4 (which are within the consistent range), the FDR_n reduces rapidly towards zero, (iii) though the PDR_n with γ_3 and γ_4 are lower than those with γ_1 and γ_2 when sample size is small, but they become comparable as the sample size increases, and (iv) the FDR_n with γ_4 is lower than that with γ_3 when sample size is small, however, the PDR_n is also lower, as sample size gets larger, both the PDR_n and FDR_n with γ_3 and those with γ_4 become comparable. These findings demonstrate that the selection consistency of EBIC is well realized in the finite sample case.

Real Data Analysis: Leukemia Data

In this subsection, we analyze a famous Leukemia data set published in [80] aiming at detecting genes which affect the category of Leukemia. It is available in R packages `Biobase` and `golubEsets`. This data set consists of expression levels of 7129 genes from 47 patients with acute lymphoblastic leukemia (ALL) and 25 with myeloid leukemia (AML). We compare our result with the findings in [80] and [111](probit model based), [115] (logistic model based) in two different ways.

Firstly, the 7129 genes are reduced to 300 genes by adopting the method in [65] to enter forward selection based on the log-likelihood of the fitted models. We then compare the top 19 genes with those in [115], the top 27 genes with those in [111] and the top 50 genes with those in [80]. The result is displayed in Table 4.2.4, $*, \triangle, \star$ on the upper right represents the common gene with [115], [111] and [80] respectively. From Table 4.2.4, we can see that genes with ID $1834, 1882, 6855$ are all detected as important genes by these four different methods.

Secondly, we use EBIC with $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (0, 0.3795, 0.8795, 1, 2)$ where the second and the third corresponds to $0.5 - \dfrac{\ln n}{4 \ln p_n}, 1 - \dfrac{\ln n}{4 \ln p_n}$ to determine the best model among the top 60 models produced by forward selection. The result is displayed in Table 4.2.5. We can see that when $\gamma \in [0, 1]$, in models with logit,

probit, cauchit, EBIC can retain one of the common important genes reported in [115], [111] and [80], while in model with complementary log-log link, EBIC can retain two of the common important genes.

We also used 8-fold cross validation to select the optimal link function among these four links. In detail, we randomly split the samples into 8 groups $G_1, G_2, \cdots, G_8$, the mean prediction error is defined to be

$$\text{MSE} = \frac{1}{8}\sum_{j=1}^{8}\sum_{i\in G_j}\left(y_i^{(j)} - \hat{y}_i^{(j)}\right), \tag{4.2.1}$$

where $\hat{y}_i^j$ is the fitted response value in the model which is selected based on observations excluding G_j. The variables included in the model are selected in two different ways: (1) under each link function, two important features are selected, their union is taken into the model; (2) selecting the top fifty important features for each link, a feature is included in the model if it belongs to at least two of these four sets. For this Leukemia data set, the optimal link function is logit for these two methods. We note that, when all the samples are applied to select the genes responsible for the classification of Leukemia, for all the link functions, the one with logit link has the maximum log likelihood among the four.

Table 4.2.1 Results on the FS-EBIC procedure with Structure I in GLMs with Cloglog Link

		γ_1		γ_2		γ_3		γ_4	
ρ	n	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n
0	100	.736	.375	.735	.362	.646	.193	.481	.074
		(.281)	(.292)	(.284)	(.291)	(.382)	(.228)	(.453)	(.141)
	200	.930	.272	.918	.223	.879	.127	.862	.078
		(.220)	(.252)	(.253)	(.215)	(.311)	(.147)	(.337)	(.108)
	500	.971	.408	.963	.371	.939	.079	.936	.026
		(.135)	(.181)	(.163)	(.152)	(.231)	(.119)	(.238)	(.062)
0.3	100	.708	.407	.708	.398	.621	.196	.471	.081
		(.298)	(.296)	(.298)	(.306)	(.384)	(.230)	(.442)	(.152)
	200	.933	.281	.924	.239	.889	.143	.855	.083
		(.202)	(.248)	(.232)	(.212)	(.303)	(.161)	(.344)	(.111)
	500	.969	.428	.959	.354	.938	.047	.933	.014
		(.130)	(.169)	(.177)	(.138)	(.238)	(.091)	(.247)	(.048)
0.5	100	.712	.401	.711	.383	.632	.201	.451	.080
		(.293)	(.295)	(.294)	(.292)	(.385)	(.223)	(.447)	(.146)
	200	.929	.281	.923	.243	.881	.128	.858	.084
		(.219)	(.257)	(.236)	(.223)	(.313)	(.130)	(.343)	(.110)
	500	.967	.434	.959	.371	.939	.043	.933	.006
		(.142)	(.166)	(.168)	(.147)	(.235)	(.085)	(.249)	(.031)
0.7	100	.674	.432	.674	.414	.606	.244	.430	.092
		(.291)	(.289)	(.291)	(.287)	(.365)	(.241)	(.432)	(.144)
	200	.931	.292	.926	.248	.888	.148	.874	.112
		(.196)	(.246)	(.218)	(.207)	(.295)	(.146)	(.314)	(.125)
	500	.970	.427	.966	.365	.937	.032	.934	.010
		(.134)	(.173)	(.150)	(.150)	(.234)	(.072)	(.240)	(.038)

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0, 0.5 - \dfrac{\ln n}{4 \ln p_n}, 1 - \dfrac{\ln n}{4 \ln p_n}, 1)$.

Table 4.2.2 Results on the FS-EBIC procedure with Structure II in GLMs with Cloglog Link

		γ_1		γ_2		γ_3		γ_4	
ρ	n	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n
0.3	100	.662	.424	.660	.409	.594	.233	.492	.132
		(.272)	(.287)	(.276)	(.286)	(.350)	(.237)	(.392)	(.195)
	200	.931	.256	.926	.231	.891	.111	.881	.068
		(.199)	(.245)	(.212)	(.222)	(.281)	(.137)	(.295)	(.101)
	500	.973	.401	.967	.339	.946	.041	.941	.018
		(.127)	(.173)	(.149)	(.134)	(.209)	(.089)	(.217)	(.055)
0.5	100	.571	.489	.570	.478	.521	.304	.442	.189
		(.259)	(.274)	(.261)	(.276)	(.303)	(.265)	(.337)	(.230)
	200	.918	.272	.910	.239	.888	.121	.869	.081
		(.204)	(.256)	(.230)	(.231)	(.267)	(.148)	(.293)	(.122)
	500	.970	.402	.964	.351	.946	.056	.942	.021
		(.129)	(.183)	(.148)	(.153)	(.199)	(.115)	(.212)	(.062)

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0, 0.5 - \dfrac{\ln n}{4 \ln p_n}, 1 - \dfrac{\ln n}{4 \ln p_n}, 1)$.

Table 4.2.3 Results on the FS-EBIC procedure with Structure III in GLMs with Cloglog Link

	γ_1		γ_2		γ_3		γ_4	
n	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n
100	.586	.506	.586	.484	.524	.332	.387	.198
	(.258)	(.252)	(.258)	(.253)	(.316)	(.252)	(.366)	(.239)
200	.796	.414	.791	.386	.767	.285	.746	.221
	(.261)	(.282)	(.274)	(.273)	(.311)	(.247)	(.334)	(.228)
500	.946	.479	.936	.416	.912	.195	.896	.171
	(.167)	(.165)	(.197)	(.150)	(.248)	(.185)	(.269)	(.176)

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0, 0.5 - \dfrac{\ln n}{4 \ln p_n}, 1 - \dfrac{\ln n}{4 \ln p_n}, 1)$.

Table 4.2.4 Leukemia Data: The Top 50 Genes Selected by Forward Selection under GLMs with Different Link Functions

Link Function	Genes ID
logit	$1834^{*,\star,\triangle}$, 4438, 4951, $6539^{\star}$, 155, 2181, $1882^{*,\star,\triangle}$, 6472, 65, 1953 3692, 706, 1787, $5191^{\star}$, 1239, 3119, 2784, 1078, 3631, 6308 $6373^{\star}$, $1909^{\star}$, 4153, $1685^{\triangle}$, $6855^{\star,\triangle}$, 7073, 5539, 2830, 4819, 6347 1081, 1095, 5328, 4279, 4373, 5737, 4366, 5280, 3307, 284 6676, 4291, 1945, 4079, 3722, 668, 782, $4196^{\star}$, 25, $4389^{\star}$
probit	$1834^{*,\star,\triangle}$, 4438, 4951, 155, 5585, 5466, 706, $7119^{\star}$, 3119, 4480 $6201^{\triangle}$, 490, 6895, $1882^{*,\star,\triangle}$, 1809, 2855, 3123, 4211^{*}, $2020^{*,\star}$, 3631 5823, 1953, $1745^{\star,\triangle}$, 65, 997, $1928^{\star}$, 3307, 1787, 538, 5539 4107, 2385, 1087, $1909^{\star}$, 5376, 5552, 6005, 1604, 3391, 5442 6702, 6309, $2348^{\star}$, 4282, 4925, 6167, 2323, 1779, 5122, $3847^{\star}$
cauchit	$1882^{*,\star,\triangle}$, 4951, $6281^{\star}$, 4499, 4443, $6539^{\star}$, 5107, $1834^{*,\star,\triangle}$, 4480, 6271 6378, 3631, $2111^{\star}$, $6201^{\triangle}$, $6373^{\star}$, 1800, 4780, 321, $4107^{\triangle}$, $1779^{\triangle}$ 6277, 1544, $5254^{\star}$, $1928^{\star}$, $1745^{\star,\triangle}$, 3163, 7073, 310, $4389^{\star}$, 5146 1927, 885, 3137, 2258, 4334, 6657, 2733, 5336, 5972, 6167 4229, $4328^{\star}$, 715, 4149, $5191^{\star}$, 6283, 200, 6702, 5794, 4190
cloglog	$1834^{*,\star,\triangle}$, $6855^{*,\star,\triangle}$, 4377, 5122, 2830, 4407, 4780, 6309, $4973^{\star}$, 715 5376, 930, 1800, $1882^{*,\star,\triangle}$, 5794, 4399, $4389^{\star}$, 922, 1962, 4267 1926, 4229, $5254^{\star}$, 770, 2141, 6923, 7073, 2828, $4847^{\star}$, 698 1779, $1928^{\star}$, 4049, 876, 6857, 6347, $6376^{\star}$, 2361, 4664, 758 3631, 6308, 4499, 4480, 5971, 6510, 5300, 3475, 3932, 6801

Table 4.2.5 Leukemia Data: The Genes Selected by EBIC under GLMs with Different Link Functions

Link Function	Genes ID				
	γ_1	γ_2	γ_3	γ_4	γ_5
logit	$1834^{*,\star,\triangle}$, 4438 (logLik=-2.296e-08)	1834, 4438 (-2.296e-08)	1834, 4438 (-2.296e-08)	1834 (-9.786)	NULL
probit	$1834^{*,\star,\triangle}$, 4438 (logLik=-3.022e-08)	1834, 4438 (-3.022e-08)	1834, 4438 (-3.022e-08)	1834 (-9.5)	NULL
cauchit	$1882^{*,\star,\triangle}$, 4951 (logLik=-2.122e-06)	1882, 4951 (-2.122e-06)	1882, 4951 (-2.122e-06)	1882, 4951 (-2.122e-06)	NULL
cloglog	$1834^{*,\star,\triangle}$, $6855^{*,\star,\triangle}$ (logLik=-6.908e-08)	1834, 6855 (-6.908e-08)	1834, 6855 (-6.908e-08)	1834, 6855 (-6.908e-08)	NULL

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (0, 0.5, 1 - \frac{\ln n}{4 \ln p_n}, 1, 2)$.

CHAPTER 5

EBIC in Cox's Proportional Hazards Models

In this chapter, we prove the selection consistency of the EBIC in Cox model as stated in Theorem 5.1.3. Meanwhile, we gave a large deviation result on the score function in Theorem 5.1.1 and a uniform convergence rate for the partial likelihood estimator in Theorem 5.1.2. The last two theorems were derived not only to assist the proof of our main theorem, but are also very important in high-dimensional studies.

5.1 Selection Consistency of EBIC

Let T and C denote the survival and censoring times, they have cumulative distribution functions F and G with associated density functions f and g respectively, and they are assumed to be conditionally independent given the covariate vector $\mathbf{Z}$. $\mathbf{Z}$ may depend on time t.

In the right-censored model, the observations from n independent individuals are triplets $(X_i, \delta_i, \mathbf{Z}_i)_{i=1}^n$, where $X_i = \min(T_i, C_i), \delta_i = I(T_i \leq C_i), \mathbf{Z}_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,p_n})$. Currently, we assume the dimension of the full covariates space $p_n = O(n^\kappa)$ for some $\kappa > 1$. Suppose there are no ties in the observation times. The complete likelihood of the observed data set is

$$L = \prod_{i:\delta_i=1} f(X_i|\mathbf{Z}_i) \prod_{i:\delta_i=0} (1 - F(X_i|\mathbf{Z}_i)) = \prod_{i:\delta_i=1} h(X_i|\mathbf{Z}_i) \prod_{i=1}^{n} (1 - F(X_i|\mathbf{Z}_i)), \tag{5.1.1}$$

where

$$h(t|\mathbf{z}) = \lim_{\Delta t \downarrow 0} P(t \leq T < t + \Delta t | T \geq t, \mathbf{Z} = \mathbf{z}) = f(t|\mathbf{z}) / (1 - F(t|\mathbf{z})) \tag{5.1.2}$$

is the conditional hazard function of T given $\mathbf{Z} = \mathbf{z}$.

The Cox's proportional hazards model assumes

$$h(t|\mathbf{z}) = h_0(t)\exp(\mathbf{z}^\tau\beta), \tag{5.1.3}$$

where $h_0(t)$ is the baseline hazard rate with cumulative baseline hazard function

$$H_0(t) = \int_0^\tau h_0(s)ds. \tag{5.1.4}$$

Without loss of generality, we assume that the support of $h_0(t)$ is $[0,1]$ and $\int_0^1 h_0(t)dt < +\infty$.

Let $t_1^0 < t_2^0 < \cdots < t_N^0$ be the ordered distinct observed failure times. Let (j) index its associated covariates $\mathbf{Z}_{(j)}$ and $\mathscr{R}(t)$ be the risk set: $\mathscr{R}(t) = \{i : X_i \geq t\}$. Partial likelihood estimation in Cox's model considers the "least informative" nonparametric modeling for $H_0(t)$: $H_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. Take the partial differential of $\ln L$ with respect to h_j, the maximizers h_j can therefore be given by

$$\hat{h}_j(\beta) = \left(\sum_{i\in\mathscr{R}(t_j^0)} \exp(\mathbf{Z}_i^\tau\beta)\right)^{-1}. \tag{5.1.5}$$

Partial likelihood estimate for the true vector β_0 is the maximizer of

$$l_n(\beta) = \sum_{j=1}^{N}\left(\mathbf{Z}_{(j)}^\tau\beta - \ln\left(\sum_{i\in\mathscr{R}(t_j^0)}\exp(\mathbf{Z}_i^\tau\beta)\right)\right). \tag{5.1.6}$$

For Cox's proportional hazards model, $\widehat{\beta}(s)$ in the EBIC (2.1.1) refers to the partial likelihood estimator given covariates contained in s.

For convenience, we denote by s_{0n} the set of nonzero predictors in β_0 with size p_{0n}. For the present, we assume β_0 is independent of sample size n. Let $k_n = Cp_{0n}$ for some $C > 1$. Define

$$\mathscr{A}_0 = \{s : s_{0n} \subset s; |s| \leq k_n\} \quad \mathscr{A}_1 = \{s : s_{0n} \nsubseteq s; |s| \leq k_n\}. \tag{5.1.7}$$

Beyond these, the following notations are used throughout this subsection. For the readability, most of them are consistent with those in [69]. Define

$$\begin{aligned} N_i(t) =& I(X_i \leq t, \delta_i = 1), \;\; Y_i(t) = I(X_i \geq t); \\ \mathscr{F}_t^{(n)} =& \sigma\left\{N_i(u), I(X_i \leq u, \delta_i = 0) : 0 \leq u \leq t, 1 \leq i \leq n\right\}. \end{aligned} \tag{5.1.8}$$

For any $s \subseteq \{1, 2, \cdots, p_n\}$, define $\mathbf{Z}(s), \beta(s)$ as the sub-vectors of $\mathbf{Z}, \beta$ with indices contained in s. For vectors $a = (a_1, a_2, \cdots, a_p), b = (b_1, b_2, \cdots, b_p)$, we write $a \otimes b$ for the $p \times p$ matrix $a^\tau b$ with (i, j)th element $a_i b_j$. For the convenience of presentation, we will use C_1, C_2, C to represent positive constants without specifying their values.

For a given index set $s \subseteq \{1, 2, \cdots, p_n\}$, define

$$
\begin{aligned}
S_n^{(0)}(\beta(s), t) =& \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\left(\mathbf{Z}_i(s)^\tau \beta(s)\right); \; S_n^{(1)}(\beta(s), t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i(s) Y_i(t) \exp\left(\mathbf{Z}_i(s)^\tau \beta(s)\right), \\
S_n^{(2)}(\beta(s), t) =& \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i(s)^{\otimes 2} Y_i(t) \exp\left(\mathbf{Z}_i(s)^\tau \beta(s)\right), \\
\mathbf{E}_n(\beta(s), t) =& \frac{S_n^{(1)}(\beta(s), t)}{S_n^{(0)}(\beta(s), t)}, \quad V(\beta(s), t) = \frac{S_n^{(2)}(\beta(s), t)}{S_n^{(0)}(\beta(s), t)} - \mathbf{E}_n\left(\beta(s), t\right)^{\otimes 2}
\end{aligned}
\tag{5.1.9}
$$

and their asymptotic versions as $s^{(j)}(\beta(s), t), j = 0, 1, 2; e(\beta(s), t)$ and $v(\beta(s), t)$. Then the partial likelihood function given covariates in s is $l_n\left(\beta(s)\right) = l_n\left(\beta(s), 1\right)$, where

$$
\begin{aligned}
l_n\left(\beta(s), t\right) =& \sum_{i=1}^{n} \int_0^t \mathbf{Z}_i^\tau(s) \beta(s) dN_i(u) - \int_0^t \ln\left(n S_n^{(0)}(\beta(s), u)\right) d\bar{N}(u), \\
U\left(\beta(s), t\right) =& \sum_{i=1}^{n} \int_0^t \left(\mathbf{Z}_i(s) - \mathbf{E}_n\left(\beta(s), u\right)\right) dM_i(u), \quad \text{where} \\
M_i(t) =& N_i(t) - \int_0^t Y_i(u) \exp\left(\mathbf{Z}_i^\tau(s) \beta(s)\right) h_0(u) du \\
& \text{is a martingale with respect to } \{\mathscr{F}_t^{(n)} : t \geq 0\}, \; \bar{N} = \sum_{i=1}^{n} N_i. \\
I\left(\beta(s), t\right) =& -\frac{\partial l_n\left(\beta(s), t\right)}{\partial \beta(s) \partial \beta(s)^\tau} = n \int_0^t V\left(\beta(s), w\right) S_n^{(0)}\left(\beta(s), w\right) h_0(w) dw, \\
\Sigma\left(\beta(s), t\right) =& \int_0^t v\left(\beta(s), w\right) s^{(0)}\left(\beta(s), w\right) h_0(w) dw.
\end{aligned}
\tag{5.1.10}
$$

Assumption 5.1.1. *There exists a compact neighborhood $\mathscr{B}$ of β_0 such that the*

following conditions are satisfied,

A(5.1.1.1) *There exists positive constants C_0, C_1 such that $\forall \beta \in \mathscr{B}$,*

$$P\left(\sup_{t\in[0,1]}\left|S_n^{(0)}(\beta,t)-s^{(0)}(\beta,t)\right|\geq\frac{C_1u_n}{\sqrt{n}}\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right)$$
$$P\left(\sup_{t\in[0,1]}\left|S_{n,j}^{(1)}(\beta,t)-s_j^{(1)}(\beta,t)\right|\geq\frac{C_1u_n}{\sqrt{n}}\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right)$$

hold for any positive u_n such that $u_n \to +\infty, n^{-1/6}u_n \to 0$ as $n \to +\infty$.

A(5.1.1.2) *The functions $s^{(0)}, s^{(1)}, s^{(2)}$ are element-wise bounded and $s^{(0)}$ is bounded away from 0, the family of functions $s^{(j)}(\cdot, t), 0 \leq t \leq 1$ is an equi-continuous family at β_0;*

A(5.1.1.3) *The process $Y(t) = (Y_1(t), \cdots, Y_n(t))^\tau$ is left continuous with right hand limits and satisfies $P(Y(t) = 1, 0 \leq t \leq 1) > 0$;*

A(5.1.1.4) *The covariate vector $Z(t)$ is left continuous.*

Conditions A(5.1.1.1) is a more elaborated version of Condition (2.2) in Section 8.2 of [69]. Note that $S_n^{(l)}, ; l = 0, 1, 2$ are summations of i.i.d random variables. It was verified in [68] that when the associated random variable satisfies Cramer Condition, we will have such tail probabilities. Moreover, with condition A(5.1.1.2), it can be deduced that

$$P\left(\sup_{t\in[0,1]}|\mathbf{E}_{nj}(\boldsymbol{\beta}_0,t)-e_j(\boldsymbol{\beta}_0,t)|\geq\frac{C_1u_n}{\sqrt{n}}\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right) \tag{5.1.11}$$

and

$$P\left(\sup_{t\in[0,1]}\left|\frac{I_{i,j}\left(\boldsymbol{\beta}_0,t\right)}{n}-\Sigma_{i,j}\left(\boldsymbol{\beta}_0,t\right)\right|\geq\frac{C_1u_n}{\sqrt{n}}\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right). \tag{5.1.12}$$

A(5.1.1.2) is condition (2.5) in Section 8.2 of [69] for fixed p_n, A(5.1.1.3) and A(5.1.1.4) are assumed in Theorem 4.1 in [3].

Assumption 5.1.2. *Let* $\varepsilon_{i,j}=\int_0^1\left(Z_{i,j}(t)-e_j(\boldsymbol{\beta}_0,t)\right)dM_i(t)$, *where* $e_j(\boldsymbol{\beta}_0,t)$ *is the* j*th component of* $e(\boldsymbol{\beta}_0,t)$. *For any set* $s\in\mathscr{A}_0$ *and any* $|s|-$*dimensional vector* $\boldsymbol{a}$ *which satisfies*

$$\boldsymbol{Var}(\sum_{i=1}^{n}\sum_{j\in s}\boldsymbol{a}_j\varepsilon_{i,j}/\sqrt{n})=1,$$

suppose the Cramer Condition in [68] holds for the linear combination of $\sum_{j\in s}\boldsymbol{a}_j\varepsilon_{i,j}$, *i.e, for any positive* u_n *such that* $u_n\to+\infty, n^{-1/6}u_n\to 0$ *as* $n\to+\infty$, *there exist positive constants* C_0 *and* C_1 *such that*

$$P\left(\left|\sum_{i=1}^{n}\sum_{j\in s}\boldsymbol{a}_j\varepsilon_{i,j}\right|\geq\sqrt{n}u_n\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right). \tag{5.1.13}$$

Without loss of generality, we assume all the diagonal elements of $\Sigma\left(\boldsymbol{\beta}_0,1\right)$ are 1. Then when $\boldsymbol{a}_j=1$ for any fixed j and 0 otherwise, (5.1.13) reduces to

$$P\left(\left|\sum_{i=1}^{n}\varepsilon_{i,j}\right|\geq\sqrt{n}u_n\right)\leq\frac{C_0}{u_n}\exp\left(-\frac{u_n^2}{2}\right),\ \forall j\in\{1,2,\cdots,p_n\}. \tag{5.1.14}$$

Theorem 5.1.1. (Large Deviation of the Score Function) *Under Assumptions 5.1.1 and 5.1.2, for any positive u_n such that $u_n \to +\infty, n^{-1/6}u_n \to 0, \ln n = o(u_n^2)$ as $n \to +\infty$, there exist positive constants c_0 such that*

$$P\left(|U_j\left(\boldsymbol{\beta}_0, 1\right)| > \sqrt{n}u_n\right) \leq c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right) \tag{5.1.15}$$

and for any unit vector $\boldsymbol{u}$ and $s \in \mathscr{A}_0$,

$$P\left(\left|\boldsymbol{u}^\tau \Sigma^{-1/2}\left(\boldsymbol{\beta}_0(s), 1\right) U\left(\boldsymbol{\beta}_0(s), 1\right)\right| > \sqrt{n}u_n\right) \leq c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right) \tag{5.1.16}$$

for any arbitrary $\varepsilon > 0$ and $j \in \{1, 2, \cdots, p_n\}$.

Proof of Theorem 5.1.1. Here we decompose the jth component of the score function as

$$\begin{aligned}\xi_j(t) =& \sum_{i=1}^{n} \int_0^\tau \left(\boldsymbol{Z}_{i,j} - e_j\left(\boldsymbol{\beta}_0, u\right)\right) dM_i(u) - \sum_{i=1}^{n} \int_0^\tau \left(\mathbf{E}_{n,j}\left(\boldsymbol{\beta}_0, u\right) - e_j\left(\boldsymbol{\beta}_0, u\right)\right) dM_i(u) \\ =& \xi_{1j}(t) - \xi_{2j}(t)\end{aligned} \tag{5.1.17}$$

To avoid confusion, let $\xi_j = \xi_j(1),\ \xi_{1j} = \xi_{1j}(1),\ \xi_{2j} = \xi_{2j}(1)$. For any fixed $s \in \mathscr{A}_0$,

let $\boldsymbol{a} = \boldsymbol{u}^{\tau}\Sigma^{-1/2}(\boldsymbol{\beta}_0(s), 1)$. Then

$$\boldsymbol{u}^{\tau}\Sigma^{-1/2}\left(\boldsymbol{\beta}_0(s), 1\right) U\left(\boldsymbol{\beta}_0(s), 1\right) = \sum_{j\in s} \boldsymbol{a}_j \xi_{1j} - \sum_{j\in s} \boldsymbol{a}_j \xi_{2j}. \tag{5.1.18}$$

The large deviation result of $\sum_{j\in s} \boldsymbol{a}_j \xi_{1j}$ is already given in Assumption 5.1.2, now it suffices to show the large deviation of $\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}$: Let u_n be of the same order as required and $\|\boldsymbol{x}\|_{+\infty}$ be the maximum absolute value in $\boldsymbol{x}$, then

$$\begin{aligned}
&P\left(\left|\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}\right| > \sqrt{n}u_n\right) \\
=&P\left(\left|\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}\right| > \sqrt{n}u_n,\ \| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty} \geq \frac{C_1 u_n}{\sqrt{n}}\right) \\
&+ P\left(\left|\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}\right| > \sqrt{n}u_n,\ \| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty} \leq \frac{C_1 u_n}{\sqrt{n}}\right) \\
=&P\left(\left|\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}\right| > \sqrt{n}u_n,\ \| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty} \geq \frac{C_1 u_n}{\sqrt{n}}\right) \\
&+ P\left(\left|\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}\right| > \sqrt{n}u_n,\ \| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty} \leq \frac{C_1 u_n}{\sqrt{n}},\right. \\
&\left.\sup_{u\in[0,1]} |S^{(0)}(\boldsymbol{\beta}_0, u) - s^{(0)}(\boldsymbol{\beta}_0, u)| \geq \frac{C_1 u_n}{\sqrt{n}}\right) \\
&+ P\left(\left|\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}\right| > \sqrt{n}u_n,\ \| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty} \leq \frac{C_1 u_n}{\sqrt{n}},\right.
\end{aligned} \tag{5.1.19}$$

$$
\begin{aligned}
&\sup_{u\in[0,1]} |S^{(0)}(\boldsymbol{\beta}_0,u) - s^{(0)}(\boldsymbol{\beta}_0,u)| \le \frac{C_1 u_n}{\sqrt{n}}\Bigg) \\
\le & P\left(\| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0,u\right) - e\left(\boldsymbol{\beta}_0,u\right)]\|_{+\infty} \ge \frac{C_1 u_n}{\sqrt{n}} \right) \\
&+ P\left(\sup_{u\in[0,1]} |S^{(0)}(\boldsymbol{\beta}_0,u) - s^{(0)}(\boldsymbol{\beta}_0,u)| \ge \frac{C_1 u_n}{\sqrt{n}} \right) \qquad (5.1.20) \\
&+ P\left(\left| \sum_{j\in s} \boldsymbol{a}_j \xi_{2j} \right| > \sqrt{n} u_n \mid \mathscr{C} \right) \\
= & P_{2,1} + P_{2,2,1} + P_{2,2,2},
\end{aligned}
$$

where

$$
\mathscr{C} = \left\{ \| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0,u\right) - e\left(\boldsymbol{\beta}_0,u\right)]\|_{+\infty} \le \frac{C_1 u_n}{\sqrt{n}},\ \sup_{u\in[0,1]} |S^{(0)}(\boldsymbol{\beta}_0,u) - s^{(0)}(\boldsymbol{\beta}_0,u)| \le \frac{C_1 u_n}{\sqrt{n}} \right\}
$$

Equation (5.1.11) and A(5.1.1.1) show that

$$
P_{2,1} \le C_0 \exp\left(-\frac{u_n^2}{2} + \kappa \ln n - \ln u_n \right)\ ;\ P_{2,2,1} \le C_0 \exp\left(-\frac{u_n^2}{2} - \ln u_n \right). \qquad (5.1.21)
$$

We can verify that condition on $\mathscr{C}$, the new martingale $\sum_{j\in s} \boldsymbol{a}_j \xi_{2j}(t)$ has bounded jumps by following the steps in the proof of Theorem 3.1 in [18]. Let $\bar{M}(t) = \sum_{i=1}^{n} M_i(t), \bar{N}(t) = \sum_{i=1}^{n} N_i(t)$, then $|\triangle(\bar{M}(t))| = |\triangle(\bar{N}(t))| \le 1$.

Firstly,

$$\left|\triangle\left(n^{-1/2}\xi_{2j}(t)\right)\right| \leq n^{-1/2}\| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty} \equiv n^{-1/2}c_n \leq \frac{C_1 u_n}{n}, \tag{5.1.22}$$

therefore,

$$\left|\triangle\left(n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right)\right| \leq \sum_{j\in s}|\boldsymbol{a}_j|\left|\triangle\left(n^{-1/2}\xi_{2j}(t)\right)\right| \leq \frac{|s|C_1 u_n}{n}. \tag{5.1.23}$$

Secondly, the predictable quadratic variation of $n^{-1/2}\xi_{2j}(t)$, denoted by $\left\langle n^{-1/2}\xi_{2j}(t)\right\rangle$ is bilinear and satisfies that

$$\begin{aligned}
\left\langle n^{-1/2}\xi_{2j}(t)\right\rangle =& n^{-1}\int_0^\tau \left(\mathbf{E}_{n,j}\left(\boldsymbol{\beta}_0, u\right) - e_j\left(\boldsymbol{\beta}_0\right), u)\right)^2 d\left\langle \bar{M}(u)\right\rangle \\
=& \int_0^\tau \left(\mathbf{E}_{n,j}\left(\boldsymbol{\beta}_0, u\right) - e_j\left(\boldsymbol{\beta}_0, u\right)\right)^2 S^{(0)}(\boldsymbol{\beta}_0, u)h_0(u)du \\
\leq& \| \sup_{u\in[0,1]} [\mathbf{E}_n\left(\boldsymbol{\beta}_0, u\right) - e\left(\boldsymbol{\beta}_0, u\right)]\|_{+\infty}^2 \int_0^\tau S^{(0)}(\boldsymbol{\beta}_0, u)h_0(u)du \equiv b_n^2(t). \\
\left\langle n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right\rangle \leq& |s|\sum_{j\in s}\boldsymbol{a}_j^2\left\langle n^{-1/2}\xi_{2j}(t)\right\rangle \leq |s|^2 b_n^2(t).
\end{aligned} \tag{5.1.24}$$

Obviously, $b_n^2(t) \leq b_n^2(1) \leq c_n^2\int_0^1 S^{(0)}(\boldsymbol{\beta}_0, u)h_0(u)du$. Note that

$$\int_0^1 S^{(0)}(\boldsymbol{\beta}_0, u)h_0(u)du \leq \int_0^1 s^{(0)}(\boldsymbol{\beta}_0, u)h_0(u)du + \sup_{u\in[0,1]}|S^{(0)}(\boldsymbol{\beta}_0, u) - s^{(0)}(\boldsymbol{\beta}_0, u)|\int_0^1 h_0(u)du. \tag{5.1.25}$$

A(5.1.1.2) and equation (5.1.23) implies that

$$\sup_{t\in[0,1]} b_n^2(t) \le c_n^2\left(C_1 + C_2\frac{C_1 u_n}{\sqrt{n}}\right) \le C\frac{u_n^2}{n}. \tag{5.1.26}$$

That is, when $|s| = O(1)$, condition on $\mathscr{C}$, there exist constants $b^2 = O\left(\frac{u_n^2}{n}\right)$, $K = O(\frac{u_n}{n})$ such that

$$\left|\triangle\left(n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right)\right| \le K;\ \left\langle n^{-1/2}\sum_{j\in s}\boldsymbol{a}_j\xi_{2j}(t)\right\rangle \le b^2. \tag{5.1.27}$$

According to Lemma 2.1 in [140], when $n^{-1/6}u_n \to 0, u_n \to +\infty$, we have

$$P_{2,2,2} \le 2\exp\left(-\frac{u_n^2}{2(Ku_n + b^2)}\right) \le C_1\exp\left(-\frac{u_n^2}{2}\right).$$

Hence, when $n^{-1/6}u_n \to 0, u_n \to +\infty, \ln n = o(u_n^2)$, there exists a positive constant c_0 independent of j and an arbitrarily small positive ε such that

$$P\left(\left|\boldsymbol{u}^\tau\Sigma^{-1/2}\left(\boldsymbol{\beta}_0(s),1\right)U\left(\boldsymbol{\beta}_0(s),1\right)\right| > \sqrt{n}u_n\right) \le c_0\exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right).$$

Let $\boldsymbol{a} = (\cdots, \boldsymbol{a}_j, \cdots)$, when $\boldsymbol{a}_j = 1$ and 0 otherwise, we have

$$P\left(|U_j\left(\boldsymbol{\beta}_0, 1\right)| > \sqrt{n}u_n\right) \leq c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2}\right)$$

over $j \in \{1, 2, \cdots, p_n\}$.

□

Assumption 5.1.3. *Assume the following conditions,*

A(5.1.3.1) Let $\lambda_{\min}$ denote the smallest eigenvalue of a square matrix, there exists a positive constant λ_1 such that

$$\lambda_{1,n} = \inf_{s_{on} \subsetneq s, |s| \leq k_n + p_{0n}} \lambda_{\min}\left(I\left(\beta_0(s), 1\right)\right) \geq n\lambda_1. \tag{5.1.28}$$

A(5.1.3.2) For any given $\varepsilon > 0$, there exists a constant $\delta > 0$ such that, when n is sufficiently large,

$$I\left(\beta(s), 1\right) \geq (1-\varepsilon) I\left(\beta_0(s), 1\right)$$

for all $\beta(s)$ such that $|s| \leq k_n$ and $\|\beta(s) - \beta_0(s)\|_2 \leq \delta$. Here, matrices $A \geq B$ means $A - B$ is semi-positive definite.

The counterpart of A(5.1.3.1) in linear regression models is the Sparse Riesz

Condition. Similar conditions were also assumed in [35] for generalized linear regression models. As was relaxed technically in linear regression models, a weaker version of A(5.1.3.1) can be expected in Cox models.

Theorem 5.1.2. (Uniform Convergence of the Partial Likelihood Estimator) *Under Assumptions 5.1.1,5.1.2 and 5.1.3, with probability tending to 1 as* $n \to +\infty$,

$$\|\hat{\beta}(s) - \beta_0(s)\|_2 = O(\psi_n)$$

uniformly for $s \in \mathscr{A}_0$, *where*

$$\frac{\lambda_{1,n}\psi_n}{\sqrt{n}} \to +\infty,\ \frac{\lambda_{1,n}\psi_n}{n^{2/3}} \to 0,\ \ln n = o(\frac{\lambda_{1,n}^2\psi_n^2}{n}). \tag{5.1.29}$$

Proof of Theorem 5.1.2. For any unit vector $\mathbf{w}(s)$, let $\boldsymbol{\beta}(s) = \boldsymbol{\beta}_0(s) + \psi_n\mathbf{w}(s)$. Under Assumption 5.1.3, for all $s \in \mathscr{A}_0$,

$$\begin{aligned} l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s)) =& \psi_n\mathbf{w}(s)^\tau U(\boldsymbol{\beta}_0(s), 1) - \frac{1}{2}\psi_n^2\mathbf{w}(s)^\tau\{I(\tilde{\boldsymbol{\beta}}(s), 1)\}\mathbf{w}(s) \\ \leq& \psi_n\mathbf{w}(s)^\tau U(\boldsymbol{\beta}_0(s), 1) - \frac{1-\varepsilon}{2}\lambda_{1,n}\psi_n^2. \end{aligned}$$

Hence, for some positive constant c, we have

$$P(l_n(\boldsymbol{\beta}(s)) - l_n(\boldsymbol{\beta}_0(s) > 0 \ \text{ for some } \ \mathbf{w}(s)) \leq P\left(\max_{j\in s, s\in\mathscr{A}_0} |U_j(\boldsymbol{\beta}_0(s), 1)| \geq \frac{1-\varepsilon}{2\sqrt{k_n}}\lambda_{1,n}\psi_n\right).$$

By noting that $k_n = O(1), p_n = O(n^\kappa)$ and letting $u_n = \frac{1-\varepsilon}{2\sqrt{nk_n}}\lambda_{1,n}\psi_n$, under (5.1.29), according to equation (5.1.15), it follows that

$$\begin{aligned} P\left(\max_{j\in s, s\in\mathscr{A}_0} |U_j(\boldsymbol{\beta}_0(s),1)| \geq \frac{1-\varepsilon}{2\sqrt{k_n}}\lambda_{1,n}\psi_n\right) \leq & \sum_{j\in s, s\in\mathscr{A}_0} P\left(|U_j(\boldsymbol{\beta}_0(s),1)| \geq \frac{1-\varepsilon}{2\sqrt{k_n}}\lambda_{1,n}\psi_n\right) \\ \leq & k_n p_n^{k_n} C_0 \exp\left(-C_1\frac{\lambda_{1,n}^2\psi_n^2}{n}\right) \\ \leq & \tilde{C}_0 \exp\left(-C_1\frac{\lambda_{1,n}^2\psi_n^2}{n} + C_2\kappa\ln n\right) \end{aligned}$$

for some positive constants $C_0, C_1, C_2, \widetilde{C}_0$. It converges to 0 as n goes to infinity. Because $l_n(\boldsymbol{\beta}(s))$ is a concave function for any $\boldsymbol{\beta}(s)$, we get the desired result.

□

Theorem 5.1.3. *Under Assumptions 5.1.1, 5.1.2 and 5.1.3, as $n \to +\infty$, we have*

(1) $P\left(\min_{s\in\mathscr{A}_1} EBIC_\gamma(s) \leq EBIC_\gamma(s_{0n})\right) \to 0$ *for any* $\gamma \geq 0$;

(2) $P\left(\min_{s\in\mathscr{A}_0, s\neq s_{0n}} EBIC_\gamma(s) \leq EBIC_\gamma(s_{0n})\right) \to 0$ *for any* $\gamma > 1 - \frac{1}{2\kappa}$.

It can be expected that under regular conditions, the original BIC, which corresponds to $\gamma = 0$, may not be selection consistent in Cox model with high dimensional feature space where $\kappa > 1$.

Proof of Theorem 5.1.3. Since by Lemma 3.1.1, $\ln \tau(\mathcal{S}_j) = j\kappa \ln n(1+o(1))$,

$$\mathrm{EBIC}_\gamma(s_{0n}) - \mathrm{EBIC}_\gamma(s) = 2\left(l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n}))\right) + (1+2\gamma\kappa)\left(|s_{0n}| - |s|\right)\ln n, \tag{5.1.30}$$

$\mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})$ implies

$$l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) \geq -\frac{1+2\gamma\kappa}{2}\left(|s_{0n}| - |s|\right)\ln n. \tag{5.1.31}$$

(1) When $s \in \mathscr{A}_1$, consider $\tilde{s} = s \cup s_{0n}$ and $\boldsymbol{\beta}(\tilde{s})$ near $\boldsymbol{\beta}_0(\tilde{s})$. Taylor expansion shows that

$$l_n\left(\boldsymbol{\beta}(\tilde{s})\right) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \leq \left(\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right)^\tau U(\boldsymbol{\beta}_0(s)) - \frac{(1-\varepsilon)\lambda_{1,n}}{2}\left\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\right\|_2^2$$

Let $\breve{\boldsymbol{\beta}}(\tilde{s})$ be augmented $\hat{\boldsymbol{\beta}}(s)$ with components in $\tilde{s} \cap s^c$ being 0, then $l_n\left(\hat{\boldsymbol{\beta}}(s)\right) = l_n\left(\breve{\boldsymbol{\beta}}(\tilde{s})\right)$ and $\|\breve{\boldsymbol{\beta}}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \geq |\boldsymbol{\beta}_{0,\min}|$, where $|\boldsymbol{\beta}_{0,\min}| = \min\left\{|\boldsymbol{\beta}_{0,j}| : j \in s_{0n}\right\}$.

The concavity of $l_n\left(\boldsymbol{\beta}(s)\right)$ implies

$$\begin{aligned}\mathscr{M}_n =& \sup\left\{l_n\left(\boldsymbol{\beta}(\tilde{s})\right) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) : s \in \mathscr{A}_1, \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 \geq |\boldsymbol{\beta}_{0,\min}|\right\} \\ \leq& \sup\left\{l_n\left(\boldsymbol{\beta}(\tilde{s})\right) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) : s \in \mathscr{A}_1, \|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = |\boldsymbol{\beta}_{0,\min}|\right\}.\end{aligned}$$

Since for any fixed $\tilde{s}$, when $\|\boldsymbol{\beta}(\tilde{s}) - \boldsymbol{\beta}_0(\tilde{s})\|_2 = |\boldsymbol{\beta}_{0,\min}|$,

$$l_n\left(\boldsymbol{\beta}(\tilde{s})\right) - l_n\left(\boldsymbol{\beta}_0(\tilde{s})\right) \leq |\boldsymbol{\beta}_{0,\min}| \|U_j(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} - \boldsymbol{\beta}_{0,\min}^2 \frac{(1-\varepsilon)\lambda_{1,n}}{2}$$

Therefore,

$$P\left(\mathscr{M}_n \geq -\boldsymbol{\beta}_{0,\min}^2 \frac{(1-\varepsilon)\lambda_{1,n}}{4}\right) \leq k_n p_n^{k_n} P\left(\|U_j(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} \geq \frac{|\boldsymbol{\beta}_{0,\min}|(1-\varepsilon)\lambda_{1,n}}{4}\right).$$

When $n^{1/6-\delta} = O\left(\frac{\lambda_{1,n}}{\sqrt{n}}\right)$ for some $0 < \delta < 1/6$,

$$\begin{aligned}
&P\left(l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) \geq -\frac{1+2\gamma\kappa}{2}(|s_{0n}| - |s|)\ln n\right) \\
\leq &P\left(\mathscr{M}_n \geq -C\ln n\right) \leq P\left(\mathscr{M}_n \geq -\boldsymbol{\beta}_{0,\min}^2 \frac{(1-\varepsilon)\lambda_{1,n}}{4}\right) \\
\leq &k_n p_n^{k_n} P(\|U_j(\boldsymbol{\beta}_0(\tilde{s}))\|_{+\infty} \geq \sqrt{n} n^{1/6-\delta}) \leq c_0 \exp\left(-c_1 n^{1/3-2\delta} + \kappa \ln n\right).
\end{aligned}$$

It converges to 0 when n goes to ∞. The desired result can be obtained.

(2) When $s \in \mathscr{A}_0$ and $s \neq s_{0n}$, let $m = |s| - |s_{0n}|$, $\mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})$ if and only if

$$\begin{aligned}
l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) \geq &m[0.5\ln n + \gamma \ln p_n] \\
\approx &\frac{m(1+2\gamma\kappa)\ln n}{2}.
\end{aligned} \tag{5.1.32}$$

From the assumptions, we can see that

$$\begin{aligned} l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) \leq & l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\boldsymbol{\beta}(s_{0n})) = l_n\left(\hat{\boldsymbol{\beta}}(s)\right) - l_n\left(\boldsymbol{\beta}_0(s)\right) \\ \leq & \frac{1}{2n(1-\varepsilon)} U^\tau\left(\boldsymbol{\beta}_0(s),1\right)\left[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}\right]^{-1} U\left(\boldsymbol{\beta}_0(s),1\right), \end{aligned} \tag{5.1.33}$$

where ε is an arbitrary positive value. Denote

$$\begin{aligned} \mathscr{T}_1 &= \left\{ \max_{s\in\mathscr{A}_0} \| [\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}]^{-1} - \Sigma^{-1}\left(\boldsymbol{\beta}_0(s),1\right) \|_{+\infty} \leq \frac{C_1 u_n}{\sqrt{n}} \right\} \\ \mathscr{T}_2 &= \left\{ \max_{s\in\mathscr{A}_0} \frac{U^\tau\left(\boldsymbol{\beta}_0(s),1\right) U\left(\boldsymbol{\beta}_0(s),1\right)}{|s|} \leq n u_n^2 \right\}. \end{aligned}$$

Equations (5.1.12) and (5.1.15) show that

$$P\left(\mathscr{T}_1^c\right) \leq \frac{C_0}{u_n} \exp\left(-\frac{u_n^2}{2} + 2\kappa \ln n\right) \; P\left(\mathscr{T}_2^c\right) \leq c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2} + \kappa \ln n\right).$$

Therefore, $P\left(\max_{s\in\mathscr{A}_0} \mathrm{EBIC}_\gamma(s) \leq \mathrm{EBIC}_\gamma(s_{0n})\right)$ is no more than

$$\begin{aligned} & P\left(\max_{s\in\mathscr{A}_0} \left(l_n(\hat{\boldsymbol{\beta}}(s)) - l_n(\hat{\boldsymbol{\beta}}(s_{0n})) \right) \geq \frac{m(1+2\gamma\kappa)\ln n}{2} \right) \\ \leq & P\left(\max_{s\in\mathscr{A}_0} U^\tau\left(\boldsymbol{\beta}_0(s),1\right)\left[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}\right]^{-1} U\left(\boldsymbol{\beta}_0(s),1\right) \geq mn(1-\varepsilon)(1+2\gamma\kappa)\ln n \right) \\ \leq & P\left(\max_{s\in\mathscr{A}_0} U^\tau\left(\boldsymbol{\beta}_0(s),1\right)\left[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}\right]^{-1} U\left(\boldsymbol{\beta}_0(s),1\right) \geq mn(1-\varepsilon)(1+2\gamma\kappa)\ln n \mid \mathscr{T}_1, \mathscr{T}_2 \right) \\ & + \frac{C_0}{u_n} \exp\left(-\frac{u_n^2}{2} + 2\kappa \ln n\right) + c_0 \exp\left(-\frac{(1-\varepsilon)u_n^2}{2} + \kappa \ln n\right). \end{aligned} \tag{5.1.34}$$

Since under $\mathscr{T}_1, \mathscr{T}_2$,

$$\begin{aligned}
&\max_{s\in\mathscr{A}_0}\left[U^\tau\left(\boldsymbol{\beta}_0(s),1\right)\left|[\frac{I\left(\boldsymbol{\beta}_0(s),1\right)}{n}]^{-1}-\Sigma^{-1}\left(\boldsymbol{\beta}_0(s),1\right)\right|U\left(\boldsymbol{\beta}_0(s),1\right)\right]\\
\leq& C\sqrt{n}u_n^3=C\frac{(n^{-1/6}u_n)^3}{\ln n}(n\ln n)=o(n\ln n),
\end{aligned}$$

the second and the third term in the right hand side of (5.1.34) both converge to 0 as n goes to $+\infty$ and the first term can be upper bounded by

$$\begin{aligned}
&P\left(\max_{s\in\mathscr{A}_0}U^\tau\left(\boldsymbol{\beta}_0(s),1\right)\Sigma^{-1}\left(\boldsymbol{\beta}_0(s),1\right)U\left(\boldsymbol{\beta}_0(s),1\right)\geq mn(1-\varepsilon)(1+2\gamma\kappa)\ln n\mid\mathscr{T}_1,\mathscr{T}_2\right)\\
\leq& CP\left(\max_{s\in\mathscr{A}_0}\boldsymbol{u}^\tau\Sigma^{-1/2}\left(\boldsymbol{\beta}_0(s),1\right)U\left(\boldsymbol{\beta}_0(s),1\right)\geq(1-\delta)\sqrt{mn(1-\varepsilon)(1+2\gamma\kappa)\ln n}\mid\mathscr{T}_1,\mathscr{T}_2\right),
\end{aligned}$$

where $\|\boldsymbol{u}\|_2=1$, δ is an arbitrary positive value. According to equation (5.1.16), it can be further bounded by $c_0^\star\exp\left[-\frac{1-\varepsilon^\star}{2}(1+2\gamma\kappa)m\ln n+m\kappa\ln n\right]$ where $c_0^\star$ is a positive constant. It converges to 0 when $\gamma>\frac{1}{1-\varepsilon^\star}-\frac{1}{2\kappa}$, where $\varepsilon^\star$ is an arbitrary positive value. The result is obtained.

□

5.2 Numerical Study

Simulation Results

In this subsection, the examination on the performance of SIS-Adaptive Lasso-EBIC procedure in Section 3.2 is extended to Cox proportional hazards model (CPH) where the dimension of feature space is assumed to be high. In our study, we let $p_n = n^{1.25}$ for $n = 100, 150, 200, 250$. Correspondingly, $p_n = 316, 524, 752, 994$. We concentrated on investigating the performances of EBIC in CPHs with different censoring proportions in our simulation study. The data structure is adapted from the set-up in [59] and [186].

1. The survival time T is generated as $\ln T = -\mathbf{Z}\boldsymbol{\beta}_0 + \ln \xi$, where $\xi \sim \exp(1)$, therefore, $h(t|X) = \exp(X^\tau \boldsymbol{\beta}_0)$. The censoring time is simulated from an exponential distribution with mean $U \exp(X^\tau \boldsymbol{\beta}_0)$, where $U \sim \text{Uniform}(1, L)$.

2. The predictors are normally distributed with mean 0 and the covariance matrix satisfies $\Sigma_{i,j} = 0.5^{|i-j|}$.

3. The true parameter vector satisfies $\boldsymbol{\beta}_{01} = \boldsymbol{\beta}_{09} = 0.8, \boldsymbol{\beta}_{04} = \boldsymbol{\beta}_{0,12} = 1, \boldsymbol{\beta}_{07} = \boldsymbol{\beta}_{0,15} = 0.6$ and 0 otherwise.

4. Let γ in EBIC be the following five values,

$$(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = \left(0, 0.5, 1 - \frac{\ln n}{4 \ln p_n}, 1, 2\right).$$

Different L leads to different censoring proportions in the data. Here the averaged censoring proportions with standard deviations in the parenthesis simulated from 200 replicates are summarized as follows:

L	$n = 100$	$n = 150$	$n = 200$	$n = 250$
2	.469 (.061)	.471 (.043)	.467 (.042)	.468 (.037)
3	.448 (.065)	.451 (.048)	.446 (.047)	.448 (.042)
4	.433 (.068)	.435 (.051)	.429 (.051)	.432 (.046)

For each simulation setting, the PDR_n and FDR_n averaged over 200 replicates and their standard deviations in the parenthesis are reported in Table 5.2.1. From Table 5.2.1, we can see the similar trend as in LMs and GLMs for EBIC with different γ values: (i) EBIC with $\gamma = 0$ (BIC) and $\gamma = 0.5$ achieve both higher PDR_n and FDR_n, (ii) as n increases, EBIC with $\gamma = 1 - \frac{\ln n}{4 \ln p_n}$ has comparable PDR_n with $\gamma = 0$ and $\gamma = 0.5$, but the FDR_n is satisfactory, (iii) EBIC with $\gamma = 1$ and $\gamma = 2$ both over controlled PDR_n, especially for $\gamma = 2$ when the sample size is small.

Real Data Analysis: DLBCL Data

In this subsection, we apply our proposed procedure in Chapter 2 to select genes related to diffuse large B-cell lymphoma (DLBCL). The data set was published and analyzed in [137] and it has also been studied in [81],[147]. In the data set, 240 patients were monitored using a Lyphochip cDNA microarray with 7399 probes. In the gene expression measurements of the 7399 genes (genes sharing the same name but having different predictors values will be considered to be different), a large number of them are missing. In our study, we apply the technique in [160] to impute the missing values. That is, they are imputed by the averaged expression levels of their nearest 8 neighboring genes according to Euclidean distance. The neighboring genes of a certain gene are constrained to the genes with all complete predictors and the components in the distance will be chosen as those which are complete in the particular gene.

In practice, after obtaining the solution path $\{s_k : 1 \leq k \leq K\}$ of a penalized likelihood procedure, for $1 \leq k \leq K$, compute $\text{EBIC}_\gamma(s_k)$, then select the set s_{*k} that minimizes EBIC among $\{s_k : 1 \leq k \leq K\}$. We tried different ways to get the solution path $\{s_k : 1 \leq k \leq K\}$. (i) LASSO is conducted on all the genes; (ii) Screening the first $0.6n$ genes according to their log-likelihoods in univariate fitted models, the partial likelihood estimate in the fitted Cox model with all these $0.6n$

genes is used as the initial value in the adaptive-Lasso, and the adaptive-Lasso is conducted on these $0.6n$ genes. The results are almost the same and are displayed in Table 5.2.2. When $\gamma \leq 0.7$, we identified HLA-DQα from Major histocompatibility complex, class II, which is also the second important gene selected by LARS-Cox procedure in [81] and one of the representative genes in [137]. Gene HLA-DPα was also detected in [147] and [137], but not in [81]. Moreover, we detected one important gene ($\|\|\| * AA824616 \| Hs.143964 \| ESTs25099$) belonging to the proliferation group.

Table 5.2.1 Results on the SIS-Adaptive-LASSO-EBIC Procedure with Different Censoring Proportions in CPH

		$n = 100$		$n = 150$		$n = 200$		$n = 250$	
L	γ	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n	PDR_n	FDR_n
	γ_1	.713	.551	.873	.488	.953	.425	.969	.422
		(.183)	(.176)	(.135)	(.193)	(.082)	(.184)	(.067)	(.196)
	γ_2	.496	.312	.755	.255	.907	.243	.948	.182
		(.260)	(.245)	(.222)	(.185)	(.127)	(.162)	(.094)	(.138)
2	γ_3	.345	.192	.659	.170	.844	.174	.933	.151
		(.269)	(.275)	(.267)	(.164)	(.186)	(.142)	(.104)	(.129)
	γ_4	.241	.115	.600	.132	.811	.148	.902	.122
		(.256)	(.242)	(.285)	(.174)	(.211)	(.143)	(.138)	(.118)
	γ_5	.009	.000	.113	.019	.339	.010	.648	.037
		(.051)	(.000)	(.211)	(.128)	(.329)	(.054)	(.331)	(.100)
	γ_1	.728	.539	.883	.469	.951	.407	.964	.399
		(.179)	(.197)	(.128)	(.197)	(.105)	(.197)	(.075)	(.198)
	γ_2	.536	.291	.778	.246	.915	.222	.953	.170
		(.257)	(.228)	(.214)	(.179)	(.133)	(.158)	(.089)	(.136)
3	γ_3	.384	.204	.683	.172	.871	.173	.936	.133
		(.289)	(.282)	(.259)	(.186)	(.174)	(.142)	(.099)	(.122)
	γ_4	.275	.099	.628	.127	.838	.143	.916	.111
		(.279)	(.228)	(.270)	(.164)	(.198)	(.133)	(.121)	(.112)
	γ_5	.017	.000	.152	.021	.405	.020	.705	.034
		(.067)	(.000)	(.238)	(.132)	(.354)	(.071)	(.312)	(.072)
	γ_1	.737	.523	.893	.460	.957	.414	.972	.418
		(.182)	(.195)	(.134)	(.196)	(.093)	(.193)	(.068)	(.202)
	γ_2	.555	.296	.798	.235	.922	.212	.954	.161
		(.251)	(.228)	(.205)	(.185)	(.121)	(.157)	(.087)	(.137)
4	γ_3	.398	.202	.712	.170	.882	.167	.938	.133
		(.296)	(.281)	(.251)	(.178)	(.165)	(.139)	(.102)	(.123)
	γ_4	.309	.121	.664	.136	.849	.132	.925	.112
		(.289)	(.236)	(.267)	(.163)	(.187)	(.128)	(.117)	(.115)
	γ_5	.020	.000	.181	.019	.462	.026	.739	.041
		(.074)	(.000)	(.272)	(.128)	(.362)	(.080)	(.294)	(.079)

$(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (0, 0.5, 1 - \dfrac{\ln n}{4 \ln p_n}, 1, 2)$.

Table 5.2.2 DLBCL Data: Genes Selected via the EBIC in CPH

GenBank ID	Signature	$\gamma \in R_1$	$\gamma \in R_2$	$\gamma \in R_3$	$\gamma \in R_4$
AA278718		+	+	−	−
AA004687		+	+	−	−
LC_24432	Proli	+	+	−	−
AA824616	Proli	+	+	+	−
X00452	MHC	+	+	+	−
AA490586	Germ	+	−	−	−
AA731721		+	−	−	−
X02530	Lymph	+	−	−	−
AA193262		+	+	−	−
AA469973		+	−	−	−

[1] $R_1 = \{0, 0.1\}, R_2 = \{0.2\}, R_3 = \{0.3, 0.4, 0.5, 0.6, 0.7\}, R_4 = \{0.8, 0.9, 1\}$;
[2] Germ=Germinal-cancer B-cell signature; MHC=MHC class II signature; Lymph=Lymph-node signature; Proli=Proliferation_signature.
[3] + / − represent the corresponding gene is included/ excluded via the EBIC with γ valued in the first row of the column.

Conclusion for Part I

Model Selection is crucial in high-dimensional studies. Under the regularization framework, researchers are capable of extracting a series of candidate models from all subsets for further statistical inference. When the purpose is prediction, model selection criteria based on minimizing prediction error such as Cross Validation (CV) are appreciated because of good prediction performance. However, this road is made by selecting a much bigger model than the true model, which curtails its prevalence when the identification of the sparse set of relevant features becomes the most significant task. The selection consistency of EBIC were proved under moderate conditions on the design matrix in different regression models, which makes EBIC more popular in high-dimensional studies.

Part II

Sequential LASSO in Feature Selection

In this part, we propose a novel procedure, sequential LASSO, for feature selection in linear regression models. In Chapter 6, the detailed procedure of sequential LASSO and its basic properties are given. In Chapter 7, we establish its selection consistency with ultra-high dimensional feature space and both the number and effects of causal features are allowed to depend on the sample size, the ensembles of the design matrix are either deterministic or random. Afterwards, we provide some special cases where the conditions required for the sequential LASSO to be selection consistent are satisfied but the conditions for the original LASSO are violated. We propose to employ the EBIC introduced in Chapter 2 as a stopping rule specifically for sequential LASSO. The selection consistency of the whole procedure is shown. Extensive simulation study results as well as an application in QTL mapping to compare sequential LASSO with other prevalent feature selection techniques are given in this chapter too. The sure screening property of sequential LASSO is provided in Chapter 8.

CHAPTER 6

Sequential LASSO and Its Basic Properties

6.1 Introduction to Sequential LASSO

Consider the linear regression model below:

$$y_i = \beta_0 + \sum_{j=1}^{p_n} \boldsymbol{\beta}_{0j} x_{ij} + \epsilon_i,\ i = 1, \ldots, n, \tag{6.1.1}$$

where ϵ_i's are i.i.d. normal variables with mean zero and variance σ^2, the x_{ij}'s are called features which are either deterministically determined or observed at random. The following particular natures are assumed for the above model. (a)

The dimensionality of the feature space is assumed as $\ln p_n = O(n^\kappa)$ for $\kappa > 0$ (ultra-high). (b) Let $s_{0n} = \{j : \boldsymbol{\beta}_{0j} \neq 0\}$ and let p_{0n} denote the cardinality of s_{0n}. It is assumed that $p_{0n} = O(n^c)$ for some $0 < c < 1$. (c) The magnitude of $\boldsymbol{\beta}_{0j}, j \in s_{0n}$, is allowed to vary with n. In matrix notation, (6.1.1) is expressed as

$$\boldsymbol{y}_n = X_n \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_n,$$

where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0p_n})^\tau$, $\boldsymbol{y}_n = (y_1, \ldots, y_n)^\tau$ and $X_n = (x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p_n}}$ and $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^\tau$. Let the columns of X_n be normalized such that $\frac{1}{n}\sum_{i=1}^n x_{ij} = 0$ and $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$ for all j. Let S denote the set of indices $\{1, 2, \ldots, p_n\}$. The sequential LASSO is described as follows. At initial step, sequential LASSO minimizes the following penalized sum of squares:

$$l_1 = \|\boldsymbol{y}_n - X_n \boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j \in S} |\beta_j|,$$

where $\|\cdot\|_2$ is the L_2 norm of a vector and λ_1 is the largest value of the penalty parameter such that at least one of the β_j's will be estimated to be non-zero. The set of indices of nonzero β_j's is denoted by s_{*1} and referred to as the active set. For $k \geq 1$, let $s_{\star k}$ be the index set of the features selected until step k. At step $k+1$,

sequential LASSO minimizes the following partially penalized sum of squares:

$$l_{k+1} = \|\boldsymbol{y}_n - X_n\boldsymbol{\beta}\|_2^2 + \lambda_{k+1} \sum_{j \in s_{*k}^c} |\beta_j|,$$

where no penalty is imposed on the β_j's for $j \in s_{\star k}$ and λ_{k+1} is the largest value of the penalty parameter such that at least one of the $\beta_j, j \notin s_{\star k}$'s will be estimated non-zero. The selected set is then updated to $s_{\star k+1}$. The sequential LASSO ensures that the feature will always remain in the model, see the basic properties below. This differs from the ordinary LASSO where a feature included in an earlier stage could be left out in a later stage in the solution path.

Let s be any subset of S. Denote by $X(s)$ the matrix consisting of the columns of X_n with indices in s. Similarly, let $\boldsymbol{\beta}(s)$ denote the vector consisting of the corresponding components of $\boldsymbol{\beta}$. Let $\mathscr{R}(s)$ be the linear space spanned by the columns of $X(s)$ and $H_0(s)$ denote its projection matrix, i.e, $H_0(s) = X(s)[X^\tau(s)X(s)]^{-1}X^\tau(s)$. Let $\boldsymbol{I}$ be the identity matrix with order $n \times n$. Some basic properties and computation algorithm of the sequential LASSO are given in the following section.

6.2 Basic Properties and Computation Algorithm

Proposition 6.2.1. *For $k \geq 1$ and any $l \in s_{*k}^c$, if $X(\{l\}) \in \mathscr{R}(s_{*k})$ then $l \notin s_{*k+1}$.*

Proof of Proposition 6.2.1: If $X(\{l\}) \in \mathscr{R}(s_{*k})$ then there exists an $\boldsymbol{a}_k$ such that $X(\{l\}) = X(s_{*k})\boldsymbol{a}_k$ and hence

$$\begin{aligned} l_{k+1} &= \|\boldsymbol{y}_n - X(s_{*k})(\boldsymbol{\beta}(s_{*k}) + \beta_l \boldsymbol{a}_k) - X(s^c_{*k}/\{l\})\boldsymbol{\beta}(s^c_{*k}/\{l\})\|_2^2 + \lambda(|\beta_l| + \sum_{j \in s^c_{*k}/\{l\}} |\beta|_j) \\ &= \|\boldsymbol{y}_n - X(s_{*k})\tilde{\boldsymbol{\beta}}(s_{*k}) - X(s^c_{*k}/\{l\})\boldsymbol{\beta}(s^c_{*k}/\{l\})\|_2^2 + \lambda(|\beta_l| + \sum_{j \in s^c_{*k}/\{l\}} |\beta|_j) \\ &\geq \|\boldsymbol{y}_n - X(s_{*k})\tilde{\boldsymbol{\beta}}(s_{*k}) - X(s^c_{*k}/\{l\})\boldsymbol{\beta}(s^c_{*k}/\{l\})\|_2^2 + \lambda \sum_{j \in s^c_{*k}/\{l\}} |\beta|_j. \end{aligned}$$

Thus when l_{k+1} is minimized, there must be $\beta_l = 0$, i.e., $l \notin s_{k+1}$. □

Proposition 6.2.1 implies that, for any k, the matrix $X(s_{*k})$ is of full column rank. It also suggests that, in the sequential LASSO procedure, any feature that is highly correlated with the features selected already will have little chance to be selected subsequently. This nature of the sequential LASSO is favorable when it is used for feature selection in ultra-high dimensional feature space where high spurious correlations present, see [61].

Proposition 6.2.2. *For $k \geq 1$, the minimization of l_{k+1} is equivalent to the minimization of*

$$\|[\boldsymbol{I} - H_0(s_{*k})][\boldsymbol{y}_n - X(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})]\|_2^2 + \lambda \sum_{j \in s^c_{*k}} |\beta_j|.$$

Proof of Proposition 6.2.2: Differentiating l_{k+1} with respect to $\boldsymbol{\beta}(s_{*k})$, we have

$$\frac{\partial l_{k+1}}{\partial \boldsymbol{\beta}(s_{*k})} = -2X^{\tau}(s_{*k})\boldsymbol{y}_n + 2X^{\tau}(s_{*k})X(s_{*k})\boldsymbol{\beta}(s_{*k}) + 2X^{\tau}(s_{*k})X(s^c_{*k})\boldsymbol{\beta}(s^c_{*k}).$$

Setting the above derivative to zero, we obtain

$$\hat{\boldsymbol{\beta}}(s_{*k}) = [X^{\tau}(s_{*k})X(s_{*k})]^{-1}X^{\tau}(s_{*k})[\boldsymbol{y}_n - X(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})]. \tag{6.2.1}$$

Substituting (6.2.1) into $\|\boldsymbol{y}_n - X_n\boldsymbol{\beta}\|_2^2$ we have

$$\begin{aligned} l_{k+1} &= \|\boldsymbol{y}_n - X(s_{*k})\boldsymbol{\beta}(s_{*k}) - X(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})\|_2^2 + \lambda \sum_{j\in s^c_{*k}} |\beta_j| \\ &= \|[\boldsymbol{y}_n - X(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})] - X(s_{*k})[X^{\tau}(s_{*k})X(s_{*k})]^{-1}X^{\tau}(s_{*k})[\boldsymbol{y}_n - X(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})]\|_2^2 + \lambda \sum_{j\in s^c_{*k}} |\beta_j| \\ &= \|[\boldsymbol{I} - H_0(s_{*k})][\boldsymbol{y}_n - X(s^c_{*k})\boldsymbol{\beta}(s^c_{*k})]\|_2^2 + \lambda \sum_{j\in s^c_{*k}} |\beta_j|. \end{aligned}$$

□

As a by-product of the above proof, the components of $\hat{\boldsymbol{\beta}}(s_{*k})$ are almost surely nonzero since $\boldsymbol{y}_n$ is a vector of continuous random variables. This implies that, in the sequential LASSO, we have $s_{*1} \subset s_{*2} \subset \cdots \subset s_{*k} \subset \ldots$; that is, the models selected in the sequential steps are nested.

For a general k, let $\tilde{\boldsymbol{y}}_n = [\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{y}_n$, $\tilde{X}_n = [\boldsymbol{I} - H_0(s_{*k})]X(s^c_{*k})$, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(s^c_{*k})$ and $\nu_{\bar{k}} = |s^c_{*k}|$. Then by Proposition 6.2.2, the minimization of l_{k+1} is equivalent

to the minimization of

$$\tilde{l}_{k+1} = \|\tilde{\boldsymbol{y}}_n - \tilde{X}_n\tilde{\boldsymbol{\beta}}\|_2^2 + \lambda\sum_{j=1}^{\nu_{\bar{k}}}|\tilde{\beta}_j|. \tag{6.2.2}$$

The following proposition is the Karush-Kuhn-Tucker (KKT) condition for the solution of the above minimization problem.

Proposition 6.2.3 (KKT condition)**.** *Let*

$$\partial|x| = \begin{cases} 1, & \text{if } x > 0, \\ -1, & \text{if } x < 0, \\ r, & \text{if } x = 0, \end{cases}$$

where r is an arbitrary number with $|r| \leq 1$. Let $\partial\|\tilde{\boldsymbol{\beta}}\|_1 = (\partial|\tilde{\beta}_1|, \ldots, \partial|\tilde{\beta}_{\nu_{\bar{k}}}|)^\tau$. Then $\tilde{\boldsymbol{\beta}}$ is a minimizer of (6.2.2) if

$$2\tilde{X}_n^\tau(\tilde{\boldsymbol{y}}_n - \tilde{X}_n\tilde{\boldsymbol{\beta}}) = \lambda\partial\|\tilde{\boldsymbol{\beta}}\|_1.$$

Proof of Proposition 6.2.3: We only need to verify that the form of $\partial\|\tilde{\boldsymbol{\beta}}\|_1$ given above is the sufficient and necessary condition for a sub gradient of $\|\tilde{\boldsymbol{\beta}}\|_1$. First, for any $\boldsymbol{\xi}$, we have

$$\|\boldsymbol{\xi}\|_1 - \|\tilde{\boldsymbol{\beta}}\|_1 = \sum_{j:\xi_j\neq\tilde{\beta}_j}(|\xi_j| - |\tilde{\beta}_j|)$$

$$\geq \sum_{j:\xi_j \neq \tilde{\beta}_j} \partial|\tilde{\beta}_j|(\xi_j - \tilde{\beta}_j) = \partial\|\tilde{\boldsymbol{\beta}}\|_1^{\tau}(\boldsymbol{\xi} - \tilde{\boldsymbol{\beta}}).$$

Thus by definition $\partial\|\tilde{\boldsymbol{\beta}}\|_1$ is a sub gradient.

Next, let $\boldsymbol{w}$ be any sub gradient of $\|\tilde{\boldsymbol{\beta}}\|_1$. We show that

$$w_j = \begin{cases} 1, & \text{if } \tilde{\beta}_j > 0, \\ -1, & \text{if } \tilde{\beta}_j < 0, \\ r, & \text{if } \tilde{\beta}_j = 0. \end{cases}$$

Suppose $\tilde{\beta}_j = 0$ and assume $|w_j| > 1$. Then we can define a new vector $\boldsymbol{\xi}$ such that $\xi_j = \text{sign}(w_j)$ and $\xi_i = \tilde{\beta}_i$ for $i \neq j$. Then we have $\|\boldsymbol{\xi}\|_1 - \|\tilde{\boldsymbol{\beta}}\|_1 = 1 < \boldsymbol{w}^{\tau}(\boldsymbol{\xi} - \tilde{\boldsymbol{\beta}}) = |w_j|$, contradicting to that $\boldsymbol{w}$ is a sub gradient.

Now suppose $\tilde{\beta}_j \neq 0$. For a positive number $\delta < |\tilde{\beta}_j|$, define $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ such that $\xi_{1j} = \tilde{\beta}_j + \delta\text{sign}(\tilde{\beta}_j)$, $\xi_{2j} = \tilde{\beta}_j - \delta\text{sign}(\tilde{\beta}_j)$ and $\xi_{1i} = \xi_{2i} = \tilde{\beta}_i, i \neq j$. Since $\boldsymbol{w}$ is a sub gradient we must have

$$\|\boldsymbol{\xi}_1\|_1 - \|\tilde{\boldsymbol{\beta}}\|_1 = \delta \geq \boldsymbol{w}^{\tau}(\boldsymbol{\xi}_1 - \tilde{\boldsymbol{\beta}}) = \delta w_j \text{sign}(\tilde{\beta}_j),$$

$$\|\boldsymbol{\xi}_2\|_1 - \|\tilde{\boldsymbol{\beta}}\|_1 = -\delta \geq \boldsymbol{w}^{\tau}(\boldsymbol{\xi}_2 - \tilde{\boldsymbol{\beta}}) = -\delta w_j \text{sign}(\tilde{\beta}_j),$$

which implies $w_j\text{sign}(\tilde{\beta}_j) = 1$ and hence $w_j = \text{sign}(\tilde{\beta}_j)$. □

Proposition 6.2.4. *Let* $s_{TEMP} = \left\{j : j \in s^c_{\star k}, |\tilde{\boldsymbol{y}_n}^{\tau}\boldsymbol{x}_j| = \max_{l \in s^c_{\star k}} |\tilde{\boldsymbol{y}_n}^{\tau}\boldsymbol{x}_l|\right\}$. *If*

s_{TEMP} is a singleton, then the $\boldsymbol{x}_j$ with $j \in s_{TEMP}$ is the only feature with non-zero estimated coefficient in the minimization of (6.2.2); otherwise, the minimization of (6.2.2) is equivalent to the minimization of

$$\|\tilde{\boldsymbol{y}} - \tilde{X}_{TEMP}\tilde{\beta}_{TEMP}\|_2^2 + \lambda_{k+1} \sum_{j \in s_{TEMP}} |\beta_j|,$$

where $\tilde{X}_{TEMP}$ consists of $\tilde{\boldsymbol{x}}$ with $j \in s_{TEMP}$, $\tilde{\beta}_{TEMP}$ is the corresponding coefficient vector.

This proposition follows from Proposition 6.2.3 and the proof of Theorem 7.1.1. Proposition 6.2.4 gives rise to the following simple computation algorithm for the sequential LASSO procedure.

Computation Algorithm:

• Initial Step: Standardize $\boldsymbol{y}_n, \boldsymbol{x}_j, j = 1, 2, \cdots, p$ such that $\boldsymbol{y}_n^\tau \mathbf{1} = 0, \boldsymbol{x}_j^\tau \mathbf{1} = 0$ and $\boldsymbol{y}_n^\tau \boldsymbol{y}_n = n, \boldsymbol{x}_j^\tau \boldsymbol{x}_j = n$. Compute $\boldsymbol{y}^\tau \boldsymbol{x}_j$ for $j \in S$. Let

$$s_{TEMP} = \left\{ j : |\boldsymbol{x}_j^\tau \boldsymbol{y}_n| = \max_{l \in S} |\boldsymbol{x}_j^\tau \boldsymbol{y}_n| \right\}.$$

If s_{TEMP} is a singleton, let $s_{\star 1} = s_{TEMP}$, otherwise, apply `glmpath` to $\boldsymbol{y}_n$ and $X(s_{TEMP})$ and extract the first feature with non-zero coefficient in the solution path, and let $s_{\star 1}$ be its active set.

• General Step: For $k \geq 1$, compute $\tilde{\boldsymbol{y}_n}^\tau \tilde{\boldsymbol{x}}_j$ for $j \in s_{\star k}^c$ where $\tilde{\boldsymbol{y}_n} = [\boldsymbol{I} -$

$H_0(s_{\star k})]\boldsymbol{y}_n, \tilde{\boldsymbol{x}}_j = [\boldsymbol{I} - H_0(s_{\star k})]\boldsymbol{x}_j$. Let

$$s_{TEMP} = \left\{ j : |\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{y}}_n| = \max_{l \in s_{\star k}^c} |\tilde{\boldsymbol{x}}_j^\tau \tilde{\boldsymbol{y}}_n| \right\}.$$

If s_{TEMP} is a singleton, let $s_{\star k+1} = s_{\star k} \cup s_{TEMP}$, otherwise, apply `glmpath` to $\tilde{\boldsymbol{y}}$ and $\tilde{X}(s_{TEMP})$ and extract the first feature with non-zero coefficient in the solution path, and let $s_{\star k+1}$ be $s_{\star k}$ union the active set. The procedure stops when $\text{EBIC}_\gamma(s_{\star k})$ with $\gamma = 1 - \ln n/2/\ln p_n$ begins to increase.

For more details on the stopping rule, see Section 7.3.1. The matrix $\boldsymbol{I} - H_0(s_{\star k+1})$ can be updated from $\boldsymbol{I} - H_0(s_{\star k})$ recursively. Suppose there are K active features with indices $j_l : l = 1, ..., K$ at step $k+1$. Denote by $J_l = j_1, \cdots, j_l$. Let $J_0 = \emptyset$. The recursive formula is given by

$$\boldsymbol{I} - H_0(s_{*k} \cup J_l) = [\boldsymbol{I} - H_0(s_{*k} \cup J_{l-1})] \left(\boldsymbol{I} - \frac{X_{j_l} X_{j_l}^\tau [\boldsymbol{I} - H_0(s_{*k} \cup J_{l-1})]}{X_{j_l}^\tau [\boldsymbol{I} - H_0(s_{*k} \cup J_{l-1})] X_{j_l}} \right).$$

The amount of computation in the above algorithm is minimal. The computation of the projection matrices does not involve any matrix inversion. The call for `glmpath` is in fact seldom invoked.

CHAPTER 7

Selection Consistency of Sequential LASSO

We establish in this chapter the selection consistency of the sequential LASSO when the dimension of the feature space is ultra-high, i.e., $\ln p_n = O(n^\kappa), \kappa > 0$, under two different settings of the feature matrix X_n: (i) X_n is deterministic and (ii) X_n is random. The deterministic case is dealt with in Section 7.1 and the random case in Section 7.2. The EBIC used as the stopping rule in the procedure of sequential LASSO is proposed in Section 7.3. The selection consistency of this whole procedure is established in Section 7.3.1 and demonstrated by extensive simulation studies and real data analysis in Section 7.3.2.

7.1 Selection Consistency with Deterministic Feature Matrix

In the deterministic case, the columns of X_n are normalized such that the sample mean and variance of each feature are 0 and n respectively. We now introduce some notations. For $s \subset S$, let $s^- = s^c \cap s_{0n}$. Recall that s_{0n} is the set of indices of the nonzero $\boldsymbol{\beta}_{0j}$'s. If $s \subset s_{0n}$ then s^- is the complement of s in s_{0n}. For $s \subset s_{0n}$, define

$$\gamma_n(j, s, \boldsymbol{\beta}) = \frac{1}{n} X^\tau(\{j\})[\boldsymbol{I} - H_0(s)]X_n\boldsymbol{\beta}.$$

In fact, $\gamma_n(j, s, \boldsymbol{\beta})$ only depends on $\boldsymbol{\beta}(s^c)$. But for the ease of notation, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}(s^c)$ will be used interchangeably. Unless otherwise stated, $\boldsymbol{\beta}$ also denotes the unknown true value of the parameter vector. The selection consistency of the sequential LASSO in the case of deterministic feature matrix is established under the following assumptions.

A1 $\max_{j \in s_{0n}^c} |\gamma_n(j, s, \boldsymbol{\beta})| < q \max_{j \in s^-} |\gamma_n(j, s, \boldsymbol{\beta})|$, for some $0 < q < 1$.

A2 (Partial positive cone condition). Let

$$\mathscr{A}_s = \{\tilde{j} : \tilde{j} \in s^c, |\gamma_n(\tilde{j}, s, \boldsymbol{\beta})| = \max_{j \in s^c} |\gamma_n(j, s, \boldsymbol{\beta})|\}$$

and $\tilde{X}(\mathscr{A}_s) = [\boldsymbol{I} - H_0(s)]X(\mathscr{A}_s)$. Then $[\tilde{X}^\tau(\mathscr{A}_s)\tilde{X}(\mathscr{A}_s]^{-1}\mathbf{1} > 0$, where $\mathbf{1}$ is

the vector with all components 1.

A3 $\frac{\sqrt{n}}{\ln p_n}\lambda_{\min}\left[\frac{1}{n}X^\tau(s_{0n})X(s_{0n})\right]\min_{j\in s_{0n}}|\boldsymbol{\beta}_{0j}| \to +\infty$, as $n \to \infty$, where $\lambda_{\min}$ denotes the smallest eigenvalue.

Assumption A1 is implied by the following condition

$$\|\tilde{X}_j^\tau \tilde{X}(s^-)[\tilde{X}^\tau(s^-)\tilde{X}(s^-)]^{-1}\|_1 < 1-\eta, \forall j \in s_{0n}^c, \tag{7.1.1}$$

where $\tilde{X}_j = [\boldsymbol{I} - H_0(s)]X(\{j\})$ and $0 < \eta < 1$. The claim above follows because

$$\begin{aligned}
|\gamma_n(j,s,\boldsymbol{\beta})| &= \frac{1}{n}|X^\tau(\{j\})[\boldsymbol{I}-H_0(s)]\boldsymbol{\mu}_n| \\
&= |\tilde{X}_j^\tau \tilde{X}(s^-)[\tilde{X}^\tau(s^-)\tilde{X}(s^-)]^{-1}\frac{1}{n}\tilde{X}^\tau(s^-)[\boldsymbol{I}-H_0(s)]\boldsymbol{\mu}_n| \\
&\le \|\tilde{X}_j^\tau \tilde{X}(s^-)[\tilde{X}^\tau(s^-)\tilde{X}(s^-)]^{-1}\|_1\frac{1}{n}\|\tilde{X}^\tau(s^-)[\boldsymbol{I}-H_0(s)]\boldsymbol{\mu}_n\|_\infty \\
&< (1-\eta)\frac{1}{n}\|\tilde{X}^\tau(s^-)[\boldsymbol{I}-H_0(s)]\boldsymbol{\mu}_n\|_\infty = (1-\eta)\frac{1}{n}\max_{j\in s^-}|X^\tau(\{j\})[\boldsymbol{I}-H_0(s)]\boldsymbol{\mu}_n| \\
&= (1-\eta)\max_{j\in s^-}|\gamma_n(j,s,\boldsymbol{\beta})|,
\end{aligned}$$

where the strict inequality holds by (7.1.1).

Under assumption A1, the $\mathscr{A}_s$ in A2 is a subset of s_{0n}. Assumption A2 holds if and only if

$$\tilde{X}_j^\tau \tilde{X}(\mathscr{A}_s\backslash\{j\})[\tilde{X}^\tau(\mathscr{A}_s\backslash\{j\})\tilde{X}(\mathscr{A}_s\backslash\{j\})]^{-1}\mathbf{1} < 1, \forall j \in \mathscr{A}_s. \tag{7.1.2}$$

We establish the equivalence of A2 and (7.1.2) below. Let $A = \tilde{X}(\mathscr{A}_s \backslash \{j\})$ and $\boldsymbol{b} = \tilde{X}_j$. Since a permutation of the rows and columns does not change the sum of the rows, it suffices to verify that the sum of the last row of $\begin{pmatrix} A^\tau A & A^\tau \boldsymbol{b} \\ \boldsymbol{b}^\tau A & \boldsymbol{b}^\tau \boldsymbol{b} \end{pmatrix}^{-1}$ is positive if and only if $\boldsymbol{b}^\tau A(A^\tau A)^{-1}\mathbf{1} < 1$. Let $E = I - A(A^\tau A)^{-1}A^\tau$ and $F = I - \boldsymbol{b}(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau$. By the formula for the inverse of blocked matrices, we have

$$\begin{pmatrix} A^\tau A & A^\tau \boldsymbol{b} \\ \boldsymbol{b}^\tau A & \boldsymbol{b}^\tau \boldsymbol{b} \end{pmatrix}^{-1} = \begin{pmatrix} (A^\tau F A)^{-1} & -(A^\tau A)^{-1}A^\tau \boldsymbol{b}(\boldsymbol{b}^\tau E \boldsymbol{b})^{-1} \\ -(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau A(A^\tau F A)^{-1} & (\boldsymbol{b}^\tau E \boldsymbol{b})^{-1} \end{pmatrix}$$

and

$$\begin{aligned} (A^\tau F A)^{-1} &= [A^\tau A - A^\tau \boldsymbol{b}(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau A]^{-1} \\ &= (A^\tau A)^{-1} + (A^\tau A)^{-1}A^\tau(\boldsymbol{b}^\tau E \boldsymbol{b})^{-1}\boldsymbol{b}^\tau A(A^\tau A)^{-1}. \end{aligned}$$

Substituting the expression of $(A^\tau F A)^{-1}$ into the first block of the last row of the above matrix, we obtain

$$-(\boldsymbol{b}^\tau \boldsymbol{b})^{-1}\boldsymbol{b}^\tau A(A^\tau F A)^{-1} = -(\boldsymbol{b}^\tau E \boldsymbol{b})^{-1}\boldsymbol{b}^\tau A(A^\tau A)^{-1}.$$

Thus the sum of the last row becomes

$$(\boldsymbol{b}^\tau E \boldsymbol{b})^{-1} - (\boldsymbol{b}^\tau E \boldsymbol{b})^{-1}\boldsymbol{b}^\tau A(A^\tau A)^{-1}\mathbf{1} = (\boldsymbol{b}^\tau E \boldsymbol{b})^{-1}[1 - \boldsymbol{b}^\tau A(A^\tau A)^{-1}\mathbf{1}]$$

which is greater than 0 if and only if $\boldsymbol{b}^\tau A(A^\tau A)^{-1}\mathbf{1} < 1$.

Condition (7.1.1) is a conditional version of ERC conditioning on the subset s of the relevant features. Condition (7.1.2) is similar to but much weaker than the *irrepresentable condition* ([183]). The above arguments suggest that Conditions A1 and A2 might be weaker than the ERC and the irrepresentable condition. This is indeed the case. We will demonstrate this by special cases in the below where the conditions for the selection consistency of the sequential LASSO hold but the ERC and the irrepresentable condition are not satisfied. If $\lambda_{\min}\left(\frac{1}{n}X^\tau(s_{0n})X(s_{0n})\right)$ is bounded away from zero, which is a common assumption in the case of ultra-high dimensional feature space, then Condition A3 is equivalent to $\frac{\sqrt{n}}{\ln p_n}\min_{j\in s_{0n}}|\boldsymbol{\beta}_{0j}| \to \infty$. If $\ln p_n = O(n^\kappa)$ with $\kappa < 1/2$ and $\min_{j\in s_{0n}}|\boldsymbol{\beta}_{0j}| \geq Cn^{-\delta}$ for some constant C and $\delta < 1/2 - \kappa$, A3 is then satisfied.

We now state and prove the major theorem in the following.

Theorem 7.1.1. *Suppose that assumptions A1-A3 hold. Let* $\ln p_n = O(n^\kappa)$*, where* $\kappa < 1/2$*. Then the sequential LASSO is selection consistent in the sense that*

$$P(s_{*k^*} = s_{0n}) \to 1, \quad as \quad n \to \infty,$$

where s_{*k^*} *is the set of features selected at the* k^**th step of the sequential LASSO such that* $|s_{*k^*}| = p_{0n}$*,* s_{0n} *is the set of relevant features and* $p_{0n} = |s_{0n}|$*.*

Proof of Theorem 7.1.1: By Proposition 6.2.3, at the $(k+1)$st step of the sequential LASSO, the solution $\hat{\boldsymbol{\beta}}$ satisfies

$$2\tilde{X}^\tau(\tilde{\boldsymbol{y}}_n - \tilde{X}_n\hat{\boldsymbol{\beta}}) = \lambda\partial\|\hat{\boldsymbol{\beta}}\|_1, \tag{7.1.3}$$

where $\tilde{\boldsymbol{y}}_n = [\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{y}_n$, $\tilde{X}_n = [\boldsymbol{I} - H_0(s_{*k})]X(s_{*k}^c)$, and $\partial\|\hat{\boldsymbol{\beta}}\|_1$ is a sub gradient of $\|\boldsymbol{\beta}\|_1$ at $\hat{\boldsymbol{\beta}}$ whose components are $1, -1$ or a number with absolute value less than or equal to 1 according as the components are positive, negative or zero. For $k = 0$, s_{*0} is taken as the empty set ϕ. Obviously, $s_{*0} \subset s_{0n}$. Assume that $s_{*k} \subset s_{0n}$ and $|s_{*k}| < p_{0n}$. Let

$$\hat{\gamma_n}(j, s_{*k}, \boldsymbol{\beta}) = \frac{1}{n}X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{y}_n = \gamma_n(j, s_{*k}, \boldsymbol{\beta}) + \frac{1}{n}X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{\epsilon}_n.$$

Define

$$\mathscr{A}_k = \{j : |\hat{\gamma_n}(j, s_{*k}, \boldsymbol{\beta})| = \max_{j \in s_{*k}^c} |\hat{\gamma_n}(j, s_{*k}, \boldsymbol{\beta})|\}.$$

We are going to show that, with probability converging to 1, $\mathscr{A}_k \subset s_{0n}$ and that $\mathscr{A}_k$ is the set of non-zero elements of the solution to equation (7.1.3). We first show that $\mathscr{A}_k \subset s_{0n}$, which is implied by $|\hat{\gamma_n}(j, s_{*k}, \boldsymbol{\beta})| > \max_{l \in s_{0n}^c} |\hat{\gamma_n}(l, s_{*k}, \boldsymbol{\beta})|$ for $j \in s_{*k}^-$ with probability converging to 1. The statement is established by showing

(i): $\frac{1}{n}X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{\epsilon}_n = O_p(n^{-1/2}\ln p_n)$ uniformly for all $j \in s_{*k}^c$.

(ii): For $j \in s_{*k}^-$, $\max_{j \in s_{*k}^-} |\gamma_n(j, s_{*k}, \boldsymbol{\beta})| \geq C_n n^{-1/2}\ln p_n$ for $C_n \to \infty$.

Note that $X^\tau(\{j\})[\boldsymbol{I}-H_0(s_{*k})]\boldsymbol{\epsilon}_n \sim N(0,\sigma^2\|\tilde{X}_j\|_2^2)$ where $\|\tilde{X}_j\|_2^2 \le \|X(\{j\})\|_2^2 = n$. Hence

$$\begin{aligned} & P\left(\frac{1}{n}|X^\tau(\{j\})[\boldsymbol{I}-H_0(s_{*k})]\boldsymbol{\epsilon}_n| > \sigma n^{-1/2}\ln p_n\right) \\ = \;& P(|X^\tau(\{j\})[\boldsymbol{I}-H_0(s_{*k})]\boldsymbol{\epsilon}_n| > \sigma n^{1/2}\ln p_n) \\ \le \;& P(|X^\tau(\{j\})[\boldsymbol{I}-H_0(s_{*k})]\boldsymbol{\epsilon}_n| > \sigma\|\tilde{X}_j\|_2\ln p_n) \\ = \;& P(|z| > \ln p_n) \le \frac{2}{\ln p_n}\exp\left\{-\frac{(\ln p_n)^2}{2}\right\}, \end{aligned}$$

where z is a standard normal random variable. Thus, by Bonferroni inequality,

$$P\left(\max_{j\in s_{*k}^c}\frac{1}{n}|X^\tau(\{j\})[\boldsymbol{I}-H_0(s_{*k})]\boldsymbol{\epsilon}_n| > \sigma n^{-1/2}\ln p_n\right) \le \frac{2}{\ln p_n}\exp\left\{-\frac{(\ln p_n)^2}{2}+\ln p_n\right\} \to 0. \tag{7.1.4}$$

Thus (i) is proved.

Let $\Delta(s_{*k}) = \boldsymbol{\mu}_n^\tau[\boldsymbol{I}-H_0(s_{*k})]\boldsymbol{\mu}_n$ where $\boldsymbol{\mu}_n = X_n\boldsymbol{\beta}_0$. We have the following inequalities:

$$\Delta(s_{*k}) = \sum_{j\in s_{*k}^-}\beta_j X^\tau(\{j\})[\boldsymbol{I}-H_0(s_{*k})]\boldsymbol{\mu}_n \le n\|\boldsymbol{\beta}(s_{*k}^-)\|_1\max_{j\in s_{*k}^-}|\gamma_n(j,s_{*k},\boldsymbol{\beta})|, \tag{7.1.5}$$

and

$$\begin{aligned}\Delta(s_{*k}) =& \boldsymbol{\beta}^\tau(s_{*k}^-)X^\tau(s_{*k}^-)[\boldsymbol{I}-H_0(s_{*k})]X(s_{*k}^-)\boldsymbol{\beta}(s_{*k}^-)\\ \geq& \lambda_{\min}\left(X^\tau(s_{*k}^-)[\boldsymbol{I}-H_0(s_{*k})]X(s_{*k}^-)\right)\|\boldsymbol{\beta}(s_{*k}^-)\|_2^2 \\ \geq& \lambda_{\min}\left(X^\tau(s_{0n})X(s_{0n})\right)\|\boldsymbol{\beta}(s_{*k}^-)\|_2^2.\end{aligned} \tag{7.1.6}$$

The second inequality above follows since $s_{*k}\cup s_{*k}^- = s_{0n}$ and $(X^\tau(s_{*k}^-)[\boldsymbol{I}-H_0(s_{*k})]X(s_{*k}^-))^{-1}$ is a sub-matrix of $(X^\tau(s_{0n})X(s_{0n}))^{-1}$ by the formula of the inverse of blocked matrices. Combining (7.1.5) and (7.1.6) yields

$$\begin{aligned}\max_{j\in s_{*k}^-}|\gamma_n(j,s_{*k},\boldsymbol{\beta})| &\geq \lambda_{\min}(\frac{1}{n}X^\tau(s_{0n})X(s_{0n}))\frac{\|\boldsymbol{\beta}(s_{*k}^-)\|_2^2}{\|\boldsymbol{\beta}(s_{*k}^-)\|_1}\\ &\geq \lambda_{\min}(\frac{1}{n}X^\tau(s_{0n})X(s_{0n}))\min_{j\in s_{0n}}|\boldsymbol{\beta}_{0j}|\\ &\equiv C_n n^{-1/2}\ln p_n, \quad \text{say,}\end{aligned}$$

with $C_n = \dfrac{n^{1/2}}{\ln p_n}\lambda_{\min}\left(\dfrac{1}{n}X^\tau(s_{0n})X(s_{0n})\right)\min_{j\in s_{0n}}|\boldsymbol{\beta}_{0j}|$. The second inequality above holds since $|s_{*k}^-|\|\boldsymbol{\beta}_0(s_{*k}^-)\|_2^2 \geq \|\boldsymbol{\beta}_0(s_{*k}^-)\|_1^2 \geq |s_{*k}^-|\min_{j\in s_{0n}}|\beta_{0j}|\|\boldsymbol{\beta}_0(s_{*k}^-)\|_1$. By A3, $C_n\to\infty$. Thus (ii) is proved.

By A1 and (ii),

$$\begin{aligned}&|\max_{j\in s_{*k}^-}|\gamma_n(j,s_{*k},\boldsymbol{\beta})| - \max_{j\in s_{0n}^c}|\gamma_n(j,s_{*k},\boldsymbol{\beta})||\\ >\ & (1-q)\max_{j\in s_{*k}^-}|\gamma_n(j,s_{*k},\boldsymbol{\beta})| \geq (1-q)C_n n^{-1/2}\ln p_n.\end{aligned}$$

This fact and (i) then imply that $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta})$ must attain the maximum within s_{*k}^-. Therefore, $\mathscr{A}_k \subset s_{*k}^- \subset s_{0n}$.

Without loss of generality, assume that $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) > 0$ for all $j \in \mathscr{A}_k$. Consider $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\xi})$ as a function of $\boldsymbol{\xi}$. Since the function is continuous, for each $j \in \mathscr{A}_k$, there exists a neighborhood $\mathscr{N}_j = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\beta}\|_2 \leq \delta_j\}$ and a constant $c_j > 0$ such that, for all $\boldsymbol{\xi} \in \mathscr{N}_j$, $\hat{\gamma}_n(j, s_{*k}, \boldsymbol{\xi}) - \max_{l \in \mathscr{A}_k^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\xi}))| > c_j$. Here $\mathscr{A}_k^c$ denotes the complement of $\mathscr{A}_k$ in s_{*k}^c by an abuse of notation. Let $\mathscr{N} = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\beta}\|_2 \leq \delta\}$ where $\delta = \min \delta_j$. Then for all $\boldsymbol{\xi} \in \mathscr{N}$, $\min_{j \in \mathscr{A}_k} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\xi}) - \max_{l \in \mathscr{A}_k^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\xi}))| > C$, where $C = \max c_j$.

Now construct $\hat{\boldsymbol{\beta}}$ as follows. Let $\hat{\boldsymbol{\beta}}(\mathscr{A}_k) = \omega[\tilde{X}^\tau(\mathscr{A}_k)\tilde{X}(\mathscr{A}_k)]^{-1}\mathbf{1}$ and $\hat{\boldsymbol{\beta}}(\mathscr{A}_k^c) = 0$, where $\omega > 0$. By A2, $\hat{\boldsymbol{\beta}}(\mathscr{A}_k) > 0$. Take ω small enough such that $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \in \mathscr{N}$. Thus we have $\min_{j \in \mathscr{A}_k} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) > \max_{l \in \mathscr{A}_k^c} |\hat{\gamma}_n(l, s_{*k}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))|$. On the other hand, for any $j \in \mathscr{A}_k$,

$$\begin{aligned} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) &= \max_{j \in s_{*k}^c} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) - \omega \frac{1}{n} \tilde{X}_j^\tau \tilde{X}(\mathscr{A}_k)[\tilde{X}^\tau(\mathscr{A}_k)\tilde{X}(\mathscr{A}_k)]^{-1}\mathbf{1} \\ &= \max_{j \in s_{*k}^c} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) - \frac{\omega}{n}. \end{aligned}$$

Let $\lambda = 2n \left[\max_{j \in s_{*k}^c} \hat{\gamma}_n(j, s_{*k}, \boldsymbol{\beta}) - \frac{\omega}{n}\right]$. Then, we have

$$2\tilde{X}_j^\tau(\tilde{\boldsymbol{y}}_n - \tilde{X}_n\hat{\boldsymbol{\beta}}) = \lambda, \quad \text{for} \quad j \in \mathscr{A}_k,$$

$$2\tilde{X}_j^\tau(\tilde{\boldsymbol{y}}_n - \tilde{X}_n\hat{\boldsymbol{\beta}}) < \lambda, \quad \text{for} \quad j \notin \mathscr{A}_k.$$

Let $\partial|\hat{\beta}_j| = 2\tilde{X}_j^\tau(\tilde{\boldsymbol{y}}_n - \tilde{X}_n\hat{\boldsymbol{\beta}})/\lambda$ for $j \notin \mathscr{A}_k$, and 1 for $j \in \mathscr{A}_k$. Then $\partial\|\hat{\boldsymbol{\beta}}\|_1$ with these components is a sub gradient of $\|\boldsymbol{\beta}\|_1$ at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ solves equation (7.1.3). From the construction of $\hat{\boldsymbol{\beta}}$, all the features corresponding to the non-zero components of $\hat{\boldsymbol{\beta}}$ belong to s_{0n}. Hence $s_{*k+1} \subset s_{0n}$. Thus we have shown that, given $s_{*k} \subset s_{0n}$, $s_{*k+1} \subset s_{0n}$ with probability converging to 1.

If p_{0n} is bounded then we have already established the selection consistency of the sequential LASSO. If p_{0n} diverges as $n \to \infty$, we need to show that $s_{*k} \subset s_{0n}$, $k = 1, \ldots, p_{0n}$, simultaneously, with probability converging to 1. Note that, under the assumptions, $s_{*k+1} \subset s_{0n}$ is equivalent to $\min_{j\in\mathscr{A}_k} \hat{\gamma_n}(j, s_{*k}, \boldsymbol{\beta}) > \max_{l\in\mathscr{A}_k^c} |\hat{\gamma_n}(l, s_{*k}, \boldsymbol{\beta}))|$, which is implied by

$$P\left(\max_{j\in s_{*k}^c} \frac{1}{n}|X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{\epsilon}_n| > \sigma n^{-1/2}\ln p_n\right) \to 0.$$

Therefore, when p_{0n} is divergent, the selection consistency is established if

$$P\left(\max_{\substack{0\le k<p_{0n}\\ j\in s_{*k}^c}} \frac{1}{n}|X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{\epsilon}_n| > \sigma n^{-1/2}\ln p_n\right) \to 0, \quad \text{as } n \to \infty.$$

It follows from (7.1.4) and the Bonferroni inequality that

$$\begin{aligned} & P\left(\max_{\substack{0\le k<p_{0n}\\ j\in s_{*k}^c}} \frac{1}{n}|X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{\epsilon}_n| > \sigma n^{-1/2}\ln p_n\right) \\ \le \ & \frac{2p_{0n}}{\ln p_n}\exp\left\{-\frac{(\ln p_n)^2}{2} + \ln p_n\right\} \le \frac{2}{\ln p_n}\exp\left\{-\frac{(\ln p_n)^2}{2} + 2\ln p_n\right\} \to 0, \end{aligned}$$

since $p_{0n} < p_n$. The proof is completed. □

7.2 Selection Consistency with Random Feature Matrix

Instead of considering X as a fixed design matrix, we now assume $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_n})^\tau$, $i = 1, \ldots, n$, are i.i.d. copies of a random vector $\boldsymbol{z} = (z_1, \ldots, z_{p_n})^\tau$. Without loss of generality, assume that $E\boldsymbol{z} = 0$ and $\mathrm{Var}(\boldsymbol{z}) = \Sigma$ with diagonal elements 1 and off-diagonal elements independent of n. Assume that

a1 The off-diagonal elements of Σ are bounded by a constant less than 1; that is, the correlation between any two features are bounded below from -1 and above from 1.

a2 $\sigma_{\max} \equiv \max_{1 \le j,k \le p_n} \sigma(z_j z_k) < \infty$ where $\sigma(z_j z_k)$ denotes the standard deviation of $z_j z_k$.

a3 $\max_{1 \le j,k \le p_n} \mathrm{E} \exp(t z_j z_k)$ and $\max_{1 \le j \le p_n} \mathrm{E} \exp(t z_j \epsilon)$ are finite for t in a neighborhood of zero.

For any $s, \tilde{s} \subset S$, denote by $\Sigma_{s\tilde{s}}$ the sub matrix of Σ with row indices in s and column indices in $\tilde{s}$. Define

$$\Gamma(j, s, \boldsymbol{\beta}) = (\Sigma_{jS} - \Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{sS})\boldsymbol{\beta}.$$

The following assumptions are imposed:

A1′ For any $s \subset s_{0n}$, $s \neq s_{0n}$, $\max_{j \in s_{0n}^c} |\Gamma(j, s, \boldsymbol{\beta})| < \max_{j \in s^-} |\Gamma(j, s, \boldsymbol{\beta})|$.

A2′ Let $\mathscr{A}_s = \{j : j \in s^c, |\Gamma(j, s, \boldsymbol{\beta})| = \max_{l \in s^c} |\Gamma(l, s, \boldsymbol{\beta})|\}$. Then

$$(\Sigma_{\mathscr{A}_s \mathscr{A}_s} - \Sigma_{\mathscr{A}_s s} \Sigma_{ss}^{-1} \Sigma_{s \mathscr{A}_s})^{-1} \mathbf{1} > 0.$$

A3′ $\dfrac{n^{1/2}}{\ln p_n} \lambda_{\min}(\Sigma_{s_{0n} s_{0n}})(\min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}|) \to +\infty$ as $n \to +\infty$.

The assumptions A1′ - A3′ are in fact the assumptions A1-A3 with the empirical variances and covariances of the features replaced by their theoretical counterparts. In order to establish the selection consistency of the sequential LASSO in the case of random feature matrix, we need to pass from assumptions A1′ - A3′ to assumptions A1-A3. The following lemma ensures that if A1′ - A3′ hold then A1-A3 hold with probability converging to 1 as n goes to infinity.

Lemma 7.2.1. *Under assumptions* $a1$-$a3$,

(i) $P(\max_{1 \leq j,k \leq p_n} \left| \dfrac{1}{n} \sum_{i=1}^n x_{ij} x_{ik} - \Sigma_{jk} \right| > n^{-1/3} \sigma_{\max}) \to 0.$

(ii) $P(\max_{1 \leq j \leq p_n} \left| \dfrac{1}{n} \sum_{i=1}^n x_{ij} \epsilon_i \right| > n^{-1/3} \sigma) \to 0.$

(iii) *Let* $\Sigma_{jl|s} = \Sigma_{jl} - \Sigma_{js} \Sigma_{ss}^{-1} \Sigma_{sl}$ *and* $\hat{\Sigma}_{jl|s} = X(\{j\})^\tau [\boldsymbol{I} - H_0(s)] X(\{l\})/n$. *Then*

$$\max_{1 \leq j,l \leq p_n} \max_{s : |s| \leq p_0} |\hat{\Sigma}_{jl|s} - \Sigma_{jl|s}| = o_p(1).$$

Proof of Lemma 7.2.1. : For any $j, k \in \{1, 2, \ldots, p_n\}$ it follows from Fill (1983) that

$$P(|\sum_{i=1}^{n} x_{ij}x_{ik} - n\Sigma_{jk}| > \sqrt{n}\sigma(z_j z_k)\psi_n) \leq C[1 - \Phi(\psi_n)] \exp[\frac{\psi_n^3}{\sqrt{n}}\lambda(\frac{\psi_n}{\sqrt{n}})] \quad (7.2.1)$$

where C is a constant, $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution, $\lambda(\cdot)$ is the Cramer series for the distribution of $z_j z_k$ which converges in a neighborhood of zero under assumption $a3$, and ψ_n is a sequence satisfying $\psi_n = o(n^{1/2})$ and $\psi_n \to \infty$.

Now take $\psi_n = n^{1/6-\delta}$ for $0 < \delta < \frac{1}{6} - \frac{\kappa}{2}$. Then $\lambda(\frac{\psi_n}{\sqrt{n}})$ is bounded and $\frac{\psi_n^3}{\sqrt{n}}$ goes to 0 as n converges to ∞. Thus (7.2.1) leads to

$$\begin{aligned} & P(|\sum_{i=1}^{n} x_{ij}x_{ik} - n\Sigma_{jk}| > n^{2/3-\delta}\sigma_{\max}) \leq P(|\sum_{i=1}^{n} x_{ij}x_{ik} - n\Sigma_{jk}| > n^{2/3-\delta}\sigma(z_j z_k)) \\ \leq \; & C_1[1 - \Phi(n^{1/6-\delta})] \leq \frac{C_1}{n^{1/6-\delta}} \exp(-\frac{1}{2}n^{1/3-2\delta}), \end{aligned}$$

where C_1 is a generic constant. Let $p_n = \exp(an^{\kappa})$ where $a > 0$ and $\kappa < \frac{1}{3}$. By Bonferroni inequality,

$$P(\max_{1 \leq j,k \leq p_n} \left| \sum_{i=1}^{n} x_{ij}x_{ik} - n\Sigma_{jk} \right| > n^{2/3-\delta}\sigma_{\max}) = o(n^{-1/6+\delta}) \to 0.$$

Hence (i) is proved. The proof of (ii) is similar and is omitted.

Note that, for $X(\{j\})$, $X(\{l\})$ and $X(s)$,

$$\frac{1}{n}X^\tau(\{j\})(\boldsymbol{I} - X(s)[X^\tau(s)X(s)]^{-1}X^\tau(s))X(\{l\})$$

is a continuous function of the means $\frac{1}{n}\sum_{i=1}^n x_{ij}x_{il}$, $\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik}$, $\frac{1}{n}\sum_{i=1}^n x_{il}x_{ik}$ and $\frac{1}{n}\sum_{i=1}^n x_{ik}x_{im}$, $k, m \in s$. Let $\bar{X}_{jls}$ denote the vector consisting of these means and $\boldsymbol{\mu}_{jls}$ its expectation. The function depends on $|s|$ but not on n. Let $g_{|s|}(\bar{X}_{jls})$ denote this function. We then have $g_{|s|}(\boldsymbol{\mu}_{jls}) = \Sigma_{jl|s}$.

By assumption $a1$, the range of $\boldsymbol{\mu}_{jls}$ for all j, l, s with fixed $|s|$ is compact. Hence $g_{|s|}$ is also uniformly continuous for all (j, l, s) with fixed $|s|$. Thus for any $\eta > 0$ there is a $\zeta > 0$ such that if $\|\bar{X}_{jls} - \boldsymbol{\mu}_{jls}\|_\infty \leq \zeta$ then $|g_{|s|}(\bar{X}_{jls}) - g_{|s|}(\boldsymbol{\mu}_{jls})| \leq \eta$, where ζ does not depend on (j, l, s). From the proof of (i), we can choose an n_0 such that when $n > n_0$,

$$P(\max_{1\leq j,k\leq p_n}\left|\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik} - \Sigma_{jk}\right| > \zeta) = o(n^{-1/6+\delta}).$$

Thus we have

$$P(\max_{j,l}|g_{|s|}(\bar{X}_{jls}) - g_{|s|}(\boldsymbol{\mu}_{jls})| > \eta) = o(n^{-1/6+\delta}).$$

By Bonferroni inequality,

$$P(\max_{j,l} \max_{s:|s|\leq p_{0n}} |g_{|s|}(\bar{X}_{jls}) - g_{|s|}(\boldsymbol{\mu}_{jls})| > \eta) \leq o(n^{-1/6+\delta})p_{0n} \to 0,$$

for $p_{0n} = O(n^{1/6-\delta})$. (iii) is proved.

□

Theorem 7.2.1. *Let* $\ln p_n = O(n^\kappa)$, $\kappa < 1/3$, *and* $p_{0n} = O(n^c)$, $\kappa/2 < c < 1/6$. *The sequential LASSO is selection consistent with random feature matrices that satisfy conditions* $a1$-$a3$ *and* $A1'$-$A3'$.

Theorem 7.2.1 is in fact a corollary of Lemma 7.2.1. It follows from the lemma immediately that if $a1$-$a3$ and A1′-A3′ are satisfied then A1-A3 hold with probability converging to 1. Thus the selection consistency of the sequential LASSO with random feature matrix is established.

In the following, we provide two special cases where the conditions for the selection consistency of the sequential LASSO can be directly verified. The first special case concerns constant positive correlation among the features. In this case, for the irrepresentable condition to be satisfied, some restriction must be imposed. But such restriction is not needed for sequential LASSO. The second special case deals with a correlation structure under which the irrepresentable condition is violated.

Special case I: Let the correlation matrix of $\boldsymbol{z}$ be given by

$$\Sigma = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\tau,$$

where I is the identity matrix of dimension p_n, $\mathbf{1}$ is a p_n-vector of all elements 1, and $0 < \rho \leq \rho_0 < 1$. Note that ρ is allowed to depend on n. But for the ease of notation we do not make this dependence explicit. In this case, the assumptions A1′-A3′ are satisfied with $\min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}| = Cn^{-1/2+\delta}$ for some constant C and an arbitrarily small positive δ. The claim is verified in the following.

For any $s \subset S$, the sub correlation matrix Σ_{ss} has eigenvalues $1-\rho$ and $1+(|s|-1)\rho$ with multiplicities $|s|-1$ and 1 respectively. The eigenvector corresponding to $1+(|s|-1)\rho$ is $\mathbf{1}$ with dimension $|s|$. The smallest eigenvalue is $1-\rho$. Thus A3′ follows immediately.

Now suppose $s \subset s_{0n}$. For any $j, k \in s^c$, we have

$$\begin{aligned}
\Sigma_{jk} - \Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{sk} &= \Sigma_{jk} - \rho^2 \mathbf{1}^\tau \Sigma_{ss}^{-1}\mathbf{1} = \Sigma_{jk} - \frac{\rho^2|s|}{1+(|s|-1)\rho} \\
&= \begin{cases} \dfrac{(1-\rho)(\rho|s|+1)}{1+(|s|-1)\rho} \equiv a, & \text{if } j = k \\ \dfrac{\rho(1-\rho)}{1+(|s|-1)\rho} \equiv b, & \text{if } j \neq k. \end{cases}
\end{aligned}$$

Therefore,

$$\begin{aligned}\gamma_n(j,s,\boldsymbol{\beta}) &= \sum_{k\in s^-}\beta_k(\Sigma_{jk}-\Sigma_{js}\Sigma_{ss}^{-1}\Sigma_{sk})\\ &= \begin{cases}(a-b)\beta_j + b\sum_{k\in s^-}\beta_k = b\sum_{k\in s^-}\beta_k + (1-\rho)\beta_j, & \text{for } j\in s^-,\\ b\sum_{k\in s^-}\beta_k, & \text{for } j\in s_{0n}^c.\end{cases}\end{aligned}$$

Thus

$$\max_{j\in s^-}|\gamma_n(j,s,\boldsymbol{\beta})| = \begin{cases}|b\sum_{k\in s^-}\beta_k| + (1-\rho)\max_{j\in s^-}\beta_j & \text{if } \sum_{k\in s^-}\beta_k > 0,\\ |b\sum_{k\in s^-}\beta_k| + (1-\rho)|\min_{j\in s^-}\beta_j| & \text{if } \sum_{k\in s^-}\beta_k < 0.\end{cases}$$

Obviously, $\max_{j\in s^-}|\gamma_n(j,s,\boldsymbol{\beta})| > \max_{j\in s_{0n}^c}|\gamma_n(j,s,\boldsymbol{\beta})|$ and hence A1$'$ is satisfied.

Finally, we have

$$\begin{aligned}&\Sigma_{\mathscr{A}_s\mathscr{A}_s} - \Sigma_{\mathscr{A}_s s}\Sigma_{ss}^{-1}\Sigma_{s\mathscr{A}_s}\\ =\ & (1-\rho)I + \rho\mathbf{1}\mathbf{1}^\tau - \rho^2\mathbf{1}\mathbf{1}^\tau\Sigma_{ss}^{-1}\mathbf{1}\mathbf{1}^\tau\\ =\ & (1-\rho)I + \rho\mathbf{1}\mathbf{1}^\tau - \frac{\rho^2|s|}{1+(|s|-1)\rho}\mathbf{1}\mathbf{1}^\tau\\ =\ & (1-\rho)I + \frac{\rho(1-\rho)}{1+(|s|-1)\rho}\mathbf{1}\mathbf{1}^\tau.\end{aligned}$$

Let ν be the number of elements in $\mathscr{A}_s$. The eigenvalue of the above matrix corresponding to the eigenvector $\mathbf{1}$ is

$$1-\rho+\frac{\nu\rho(1-\rho)}{1+(|s|-1)\rho}=a+(\nu-1)b.$$

Hence

$$(\Sigma_{\mathscr{A}_s\mathscr{A}_s}-\Sigma_{\mathscr{A}_s s}\Sigma_{ss}^{-1}\Sigma_{s\mathscr{A}_s})^{-1}\mathbf{1}=\frac{1}{a+(\nu-1)b}\mathbf{1}>0,$$

i.e., A2$'$ holds.

Note that, in the above argument, we only need $\rho=\rho_n\leq\rho_0<1$. But, for the irrepresentable condition to hold, the following restriction must be in place:

$$\rho_n<\frac{1}{1+c|s_{0n}|}$$

for some constant c, see [183]. If $|s_{0n}|\to\infty$, ρ_n must go to zero, i.e., eventually, all the features must be statistically uncorrelated.

Special case II. Without loss of generality, let $s_{0n}=\{1,\ldots,p_{0n}\}$. Assume that

(i) $|\boldsymbol{\beta}_{01}|>|\boldsymbol{\beta}_{02}|>\cdots>|\boldsymbol{\beta}_{0p_{0n}}|=Cn^{-1/2+\delta}$ for some constant C and an arbitrarily small positive δ;

(ii) The correlation matrix Σ has the following structure:

$$\Sigma_{s_{0n}s_{0n}} = I, \quad \Sigma_{js_{0n}} = \frac{1}{p_{0n}}\text{sign}\boldsymbol{\beta}_0^{\tau}, \text{ for } j \in s_{0n}^c.$$

In the following, we show that in this case the irrepresentable condition is violated but conditions A1$'$-A3$'$ hold, and if in addition $a2$ and $a3$ are assumed, the sequential LASSO is selection consistent. Obviously, for any $j \in s_{0n}^c$,

$$\Sigma_{js_{0n}}\Sigma_{s_{0n}s_{0n}}^{-1}\text{sign}\boldsymbol{\beta}(s_{0n}) = 1,$$

i.e., the *irrepresentable condition* does not hold. Let $s_{*0} = \phi$. Suppose $s_{*k} = \{1, \ldots, k\}$ for $k < p_{0n}$. For any $j \in s_{0n}^c$,

$$\begin{aligned}
\Gamma(j, s_{*k}, \boldsymbol{\beta}) &= [(\Sigma_{js_{*k}}, \Sigma_{js_{*k}^-}, \Sigma_{js_{0n}^c}) - \Sigma_{js_{*k}}\Sigma_{s_{*k}s_{*k}}^{-1}(\Sigma_{s_{*k}s_{*k}}, \Sigma_{s_{*k}s_{*k}^-}, \Sigma_{s_{*k}s_{0n}^c})] \begin{pmatrix} \boldsymbol{\beta}(s_{*k}) \\ \boldsymbol{\beta}(s_{*k}^-) \\ \boldsymbol{\beta}(s_{0n}^c) \end{pmatrix} \\
&= \Sigma_{js_{*k}^-}\boldsymbol{\beta}(s_{*k}^-) = \sum_{j \in s_{*k}^-} |\beta_j|/p_{0n} < |\beta_{k+1}| = \Gamma(k+1, s_{*k}, \boldsymbol{\beta}) \\
&= \max_{j \in s_{*k}^-} |\Gamma(j, s_{*k}, \boldsymbol{\beta}).
\end{aligned}$$

Thus A1$'$ is satisfied. The validity of A2$'$ is obvious since $\mathscr{A}_{s_{*k}}$ contains only one element for each $k < p_{0n}$. A3$'$ reduces to $\frac{\sqrt{n}}{\ln p_n}\min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}| \to \infty$ which holds obviously. $a1$ follows from (ii). Then, when $a2$ and $a3$ are also satisfied, the

sequential LASSO is selection consistent.

7.3 Application of Sequential LASSO in Feature Selection

7.3.1 EBIC as a Stopping Rule

In real applications, when there is no prior information on p_{0n}, an appropriate criterion to halt the sequential LASSO procedure is demanding. We propose to employ the EBIC (2.1.1) introduced in Chapter 2 as a stopping rule. In detail, we let sequential LASSO stop at the $\hat{k}$th step, where $\hat{k} = \min\{k : k \geq 1, \text{EBIC}_\gamma(s_{\star k}) < \text{EBIC}_\gamma(s_{\star k+1})\}$. In this section, we will show that sequential LASSO with this stopping rule is selection consistent. That is, $\lim P(s_{\star\hat{k}} = s_{0n}) \to 1$. This main result is expressed in a different manner in Theorem 7.3.1.

Recall that $s_{\star k}$ denotes the selected set of features after k steps in sequential LASSO. With the selection consistency of sequential LASSO, we assume $s_{\star\tilde{p}_0} = s_{0n}$. For $k \geq 1$, define

$$\begin{aligned} \text{EBIC}_\gamma(k) =& n \ln\left(\frac{\| \left[\boldsymbol{I} - H_0(s_{\star k})\right] \boldsymbol{y}_n \|_2^2}{n}\right) + k \ln n + 2\gamma \ln \binom{p_n}{k} \\ \approx& n \ln\left(\frac{\| \left[\boldsymbol{I} - H_0(s_{\star k})\right] \boldsymbol{y}_n \|_2^2}{n}\right) + k \left(\ln n + 2\gamma \ln p_n\right), \gamma \geq 0. \end{aligned} \tag{7.3.1}$$

The second approximation applies when $\ln p_n = O(n^\kappa)$ for some $\kappa > 0$ and k is of a polynomial order of n (see Lemma 3.1.1). For simplicity, we shall use this approximation in our following theorems.

A4 $\min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}| \geq Cn^{-\kappa}$.

Theorem 7.3.1. *Under the assumptions in Theorem 7.1.1 or 7.2.1, when A4 holds, we have the following conclusions,*

(i) $\forall 0 \leq k < \tilde{p}_0$, $\lim_{n \to +\infty} P\left(EBIC_\gamma(k) > EBIC_\gamma(k+1)\right) = 1$ *when* $\gamma > 0$.

(ii) $\lim_{n \to +\infty} P\left(\min_{\tilde{p}_0 \leq k \leq cp_{0n}} EBIC_\gamma(k) \geq EBIC_\gamma(p_{0n})\right) = 1$ *for any fixed constant* $c > 1$ *when* $\gamma > 1 - \dfrac{\ln n}{2 \ln p_n}$.

The proof of (ii) was already provided in the second part of the proof of Theorem 3.1.1 and hence is omitted here.

Proof of (i) in Theorem 7.3.1. For any subset $\boldsymbol{J}$ of $\{1, 2, \cdots, p_n\}$, let $\boldsymbol{\mu}_n = X_n \boldsymbol{\beta}_0$, define

$$\begin{aligned} \triangle^{\mu}(\boldsymbol{J}) =& \boldsymbol{\mu}_n^\tau \left[\boldsymbol{I} - H_0(\boldsymbol{J})\right] \boldsymbol{\mu}_n, \\ \triangle^{\epsilon}(\boldsymbol{J}) =& \boldsymbol{\epsilon}_n^\tau \left[\boldsymbol{I} - H_0(\boldsymbol{J})\right] \boldsymbol{\epsilon}_n, \\ \triangle^{\mu,\epsilon}(\boldsymbol{J}) =& \boldsymbol{\mu}_n^\tau \left[\boldsymbol{I} - H_0(\boldsymbol{J})\right] \boldsymbol{\epsilon}_n. \end{aligned} \tag{7.3.2}$$

Suppose $\boldsymbol{J}_1, \boldsymbol{J}_2$ are two subsets of $\{1, 2, \cdots, p_n\}$, decompose $\text{EBIC}_\gamma(\boldsymbol{J}_1) - \text{EBIC}_\gamma(\boldsymbol{J}_2)$

as $T_1 + T_2$ where

$$\begin{aligned} T_1 =& n \ln \left\{ 1 + \frac{(\| \left[\boldsymbol{I} - H_0(\boldsymbol{J}_1) \right] \boldsymbol{y}_n \|_2^2 - \| \left[\boldsymbol{I} - H_0(\boldsymbol{J}_2) \right] \boldsymbol{y}_n \|_2^2)}{\| \left[\boldsymbol{I} - H_0(\boldsymbol{J}_2) \right] \boldsymbol{y}_n \|_2^2} \right\} \\ =& n \ln \left\{ 1 + \frac{\left\{ \triangle^{\mu} (\boldsymbol{J}_1) - \triangle^{\mu} (\boldsymbol{J}_2) \right\} + 2 \left\{ \triangle^{\mu,\epsilon} (\boldsymbol{J}_1) - \triangle^{\mu,\epsilon} (\boldsymbol{J}_2) \right\} + \left\{ \triangle^{\epsilon} (\boldsymbol{J}_1) - \triangle^{\epsilon} (\boldsymbol{J}_2) \right\}}{\triangle^{\mu} (\boldsymbol{J}_2) + 2 \triangle^{\mu,\epsilon} (\boldsymbol{J}_2) + \triangle^{\epsilon} (\boldsymbol{J}_2)} \right\}; \\ T_2 =& (|\boldsymbol{J}_1| - |\boldsymbol{J}_2|)(\ln n + 2\gamma \ln p_n). \end{aligned} \tag{7.3.3}$$

Now we aim to prove $\text{EBIC}_\gamma(k) \leq \text{EBIC}_\gamma(k+1)$ occurs with probability tending to 0 for any $0 \leq k < \tilde{p}_0$. Let $\mathscr{A}_k$ be the one defined in the proof of Theorem 7.1.1. Note that when $\boldsymbol{J}_1 = s_{\star k},\ \boldsymbol{J}_2 = s_{\star k+1},\ T_2 = -|\mathscr{A}_k|(\ln n + 2\gamma \ln p_n)$. It suffices to show that, uniformly for $0 \leq k < \tilde{p}_0$,

$$P\left(T_1 \leq |\mathscr{A}_k| \left(\ln n + 2\gamma \ln p_n\right)\right) \to 0. \tag{7.3.4}$$

(1) If $k < \tilde{p}_0 - 1$, $s_{0n} \cap s_{\star k+1}^c \neq \emptyset$ and thus $\mathscr{A}_{k+1} \neq \emptyset$. Targeting at simplifying T_1, firstly, we prove the following two important conclusions:

$$\begin{aligned} &(\mathbf{I}) : \max_{0 \leq k \leq \tilde{p}_0 - 1} \left(\triangle^{\epsilon}(s_{\star k}) - \triangle^{\epsilon}(s_{\star k+1}) \right) = O_p(\ln n); \\ &(\mathbf{II}) : \max_{0 \leq k \leq \tilde{p}_0 - 1} \frac{\triangle^{\mu,\epsilon}(s_{\star k}) - \triangle^{\mu,\epsilon}(s_{\star k+1})}{\triangle^{\mu}(s_{\star k}) - \triangle^{\mu}(s_{\star k+1}} = o_p(1). \end{aligned}$$

Proof of (I): For $x > 0$, since $\mathscr{A}_k \subseteq s_{0n}$,

$$\begin{aligned}
&P\left(\max_{0\le k\le p_{0n}-1}(\triangle^{\epsilon}(s_{\star k})-\triangle^{\epsilon}(s_{\star k+1}))\ge x\right)\\
=&P(\max_{0\le k\le n-1,1\le j\le s_{0n}} Z_{k,j}\ge x)\quad \text{where}\quad Z_{k,j}\sim\chi^2(|\mathscr{A}_k|)\\
\le& np_{0n}P(\chi^2(1)\ge x)\qquad \text{by}\quad \text{Bonferroni Inequality}\\
=&2np_{0n}(1-\Phi(\sqrt{x}))\le\frac{C}{\sqrt{x}}\exp(-\frac{x}{2}+\ln n+\ln p_{0n}),
\end{aligned}$$

Let $x = 4(\ln n + \ln p_{0n}) = O(\ln n)$, the right hand side of the inequality converges to 0 as $n \to +\infty$, the result follows.

Moreover, the following inequality will be proved in the Appendix:

$$\frac{\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1})}{\ln n}\to+\infty. \tag{7.3.5}$$

Now we prove (II): since $\dfrac{\triangle^{\mu,\epsilon}(s_{\star k})-\triangle^{\mu,\epsilon}(s_{\star k+1})}{\sqrt{\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1})}}$ can be expressed as a standard normally distributed random variable, by following the same arguments as in the proof of (I), we know that with probability tending to 1,

$$\max_{k\le\tilde{p}_0-1}\frac{\triangle^{\mu,\epsilon}(s_{\star k})-\triangle^{\mu,\epsilon}(s_{\star k+1})}{\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1})}=\sqrt{\frac{O_p(\ln n)}{\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k})}}=o_p(1).$$

Then (II) holds.

The conclusions (I) and (II) together with (7.3.5) imply

$$T_1 \geq n \ln \left\{ 1 + \frac{\triangle^{\mu}(s_{\star k}) - \triangle^{\mu}(s_{\star k+1})}{2 \max\{\triangle^{\mu}(s_{\star k+1}), \triangle^{\epsilon}(s_{\star k+1})\}} (1 + o_p(1)) \right\}. \tag{7.3.6}$$

We also proved in the Appendix the following inequality:

$$\forall 0 \leq k < \tilde{p}_0 - 1, \triangle^{\mu}(s_{\star k}) - \triangle^{\mu}(s_{\star k+1}) \geq \frac{\triangle^{\mu}(s_{\star k+1})^2}{(1+\lambda_0)^2 n \|\boldsymbol{\beta}(s^-_{\star k+1})\|_1^2}, \tag{7.3.7}$$

where λ_0 is a constant larger than 1. Now we investigate (7.3.6) by looking into two situations separately,

a. If $\max\{\triangle^{\mu}(s_{\star k+1}), \triangle^{\epsilon}(s_{\star k+1})\} = \triangle^{\mu}(s_{\star k+1})$, plug this inequality into (7.3.6), from inequality (7.1.6), we have, with probability tending to 1,

$$\begin{aligned} \frac{T_1}{\ln p_n} \geq & \frac{n}{\ln p_n} \ln \left\{ 1 + \frac{\triangle^{\mu}(s_{\star k+1})}{2(1+\lambda_0)^2 n \|\boldsymbol{\beta}(s^-_{\star k+1})\|_1^2} \right\} \\ \geq & \frac{n}{\ln p_n} \ln \left\{ 1 + \frac{\lambda_{\min}\left(X^{\tau}(s_{0n}) X(s_{0n})\right) \min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}|}{2(1+\lambda_0)^2 n} \right\}. \end{aligned}$$

Note that $\dfrac{\lambda_{\min}\left(X^{\tau}(s_{0n}) X(s_{0n})\right) \min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}|}{(1+\lambda_0)^2 n}$ is upper bounded by a positive constant, hence there exists a constant $C > 0$ such that the right hand side is larger than

$$C \frac{n}{\ln p_n} \frac{\lambda_{\min}\left(X^{\tau}(s_{0n}) X(s_{0n})\right) \min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}|}{(1+\lambda_0)^2 n},$$

which is larger than $C\frac{\sqrt{n}C_n}{(1+\lambda_0)^2}$ where $C_n \to +\infty$ according to A3. Inequality (7.3.4) is proved.

b. If $\max\{\triangle^{\mu}(s_{\star k+1}), \triangle^{\epsilon}(s_{\star k+1})\} = \triangle^{\epsilon}(s_{\star k+1})$, which is $n(1+o_p(1))$,

$$\begin{aligned}\frac{T_1}{\ln p_n} \geq & \frac{n}{\ln p_n} \ln\left\{1 + \frac{\triangle^{\mu}(s_{\star k+1})^2}{2(1+\lambda_0)^2 n^2 \|\boldsymbol{\beta}(s^-_{\star k+1})\|_1^2}\right\} \\ \geq & \frac{n}{\ln p_n} \ln\left\{1 + \frac{\lambda^2_{\min}\left(X^{\tau}(s_{0n})X(s_{0n})\right) \min^3_{j\in s_{0n}} |\boldsymbol{\beta}_{0j}|}{2(1+\lambda_0)^2 n^2}\right\} \\ \geq & C\frac{n}{\ln p_n} \frac{\lambda^2_{\min}\left(X^{\tau}(s_{0n})X(s_{0n})\right) \min^3_{j\in s_{0n}} |\boldsymbol{\beta}_{0j}|}{2(1+\lambda_0)^2 n^2} \\ \geq & \frac{C_n^2 \ln p_n \min_{j\in s_{0n}} |\boldsymbol{\beta}_{0j}|}{2(1+\lambda_0)^2},\end{aligned}$$

where $C_n \to +\infty$. Under A4, we can easily obtain inequality (7.3.4).

(2) If $k = \tilde{p}_0 - 1,\ \triangle^{\mu}(s_{\star k+1}) = 0$, and

$$\triangle^{\mu}(s_{\star k}) \geq \lambda_{\min}\left(X^{\tau}(s_{0n})X(s_{0n})\right) \min_{j\in s_{0n}} |\boldsymbol{\beta}_{0j}|^2.$$

Therefore, under A3, with probability tending to 1,

$$\begin{aligned}\frac{T_1}{\ln p_n} \geq & \frac{n}{\ln p_n} \ln\left\{1 + \frac{\lambda_{\min}\left(X^{\tau}(s_{0n})X(s_{0n})\right) \min_{j\in s_{0n}} |\boldsymbol{\beta}_{0j}|^2}{n}(1+o_p(1))\right\} \\ \geq & C\frac{\lambda_{\min}\left(X^{\tau}(s_{0n})X(s_{0n})\right) \min_{j\in s_{0n}} |\boldsymbol{\beta}_{0j}|^2}{\ln p_n} \geq C_n\sqrt{n} \min_{j\in s_{0n}} |\boldsymbol{\beta}_{0j}|,\end{aligned}$$

where $C_n \to +\infty$. Under A4, inequality (7.3.4) is obtained.

□

7.3.2 Numerical Study

Simulation Results

In this subsection, we report our simulation study results on the comparison of the following methods with our proposed sequential LASSO where EBIC is used as the stopping rule as introduced in Subsection 7.3.1:

(1) ALasso+CV: adaptive LASSO with 5-fold cross validation criterion to select the final set and for adaptive Lasso, the marginal effect of each covariate is the initial estimator, as described in [92];

(2) SCAD + CV: the same as ALasso+ CV except the regularization method is changed to SCAD, as recommended in [175];

(3) SIS+SCAD + CV: as described in [61];

(4) SLasso+ EBIC: sequential LASSO with EBIC where $\gamma = 1 - \dfrac{\ln n}{2 \ln p_n}$ as the stopping rule;

(5) FSR+ EBIC: the same as SLasso+ EBIC except the sequential method is changed to Forward Selection.

The R packages `parcor`, `ncvreg`, `SIS` are applied for the realization of ALasso + CV, SCAD+CV, SIS+SCAD+CV. The model path from SLasso can also be obtained by updating the `penalty.factor` in function `glmnet`.

In addition to the selection accuracy, we also consider the prediction error of the model selected by sequential LASSO and its competitors in this section, which is computed in the following way. For each replication, we simulate two independent data sets under the same data structure with the same sample size n, one set is used to conduct the feature selection and estimate the coefficients in the linear regression model. For each method, if the estimator is already calculated simultaneously, it will be used directly. Otherwise, the Least Squares Estimate (LSE) will be the alternative. Another data set is used to compute the predicted MSE (PMSE), which is defined as $\|\boldsymbol{y}_n - \hat{\boldsymbol{y}}_n\|_2^2/n$, where $\hat{\boldsymbol{y}}_n$ is the fitted observations for this data set from the selected model. The lower PMSE means better prediction ability.

We consider two different scenarios for the parameters in the regression model in our simulation. In scenario I, we stick to the diverging pattern in our theoretical results in Simulation Study A and Simulation Study B. In scenario II, there is no diverging pattern and examples are borrowed from the literatures, which are in Simulation Study C and Simulation Study D.

In scenario I, the diverging pattern of p_n and p_{0n} are $(p_{0n}, p_n) = ([4n^{0.16}], [5\exp(n^{0.3})])$. We let $n = 100, 200$ and $\rho = 0.5$. Two types of coefficients for causal features are considered.

Type I: The coefficients are generated as independent random variables distributed as $(-1)^u(4n^{-0.15} + |z|)$, where $u \sim Bernoulli(0.4)$ and z is a normal random variable with mean 0 and satisfies $P(|z| \geq 0.1) = 0.25$. The coefficients

take both positive and negative values and are roughly of order $O(n^{-0.15})$.

Type II: The coefficient are generated as $2j^{0.5}n^{-0.15}$, $1 \leq j \leq p_{0n}$. The coefficients are all positive and the minimum magnitude has order $O(n^{-0.15})$ while the maximum magnitude has order $O(n^{-0.07})$. Once this $\boldsymbol{\beta}$ is generated, it will be fixed for all the replications.

Once the true coefficient vector $\boldsymbol{\beta}$ and Σ, the covariance matrix of X is fixed, the error variance σ^2 is determined by the following equality:

$$h = \frac{\boldsymbol{\beta}^\tau \Sigma \boldsymbol{\beta}}{\boldsymbol{\beta}^\tau \Sigma \boldsymbol{\beta} + \sigma^2} = 0.8.$$

Simulation Study A Four commonly adopted data structures are considered:

Structure A1: All the p_n features are statistically independent with mean zero and variance 1. *Structure A2:* The Σ satisfies $\Sigma_{ij} = \rho^{|i-j|}$ for all $i, j = 1, 2, \cdots, p_n$ and $s_{0n} = \{1, 2, \cdots, p_{0n}\}$. *Structure A3:* The Σ satisfies $\Sigma_{ij} = \rho^{|i-j|}$ for all $i, j = 1, 2, \cdots, p_n$. The true features are scattered in clusters. Specifically, for $n = 100, 200$,

n	p_{0n}	s_{0n}
100	8	$\{19, 20, 21; 39, 40, 41; 60, 61\}$
200	9	$\{19, 20, 21; 39, 40, 41; 59, 60, 61\}$

Structure A4: The $X_{i,j}$ and $\epsilon_{i,j}$ are generated independently from a shifted exponential distribution $Exp(1) - 1$.

Simulation Study B Instead of the identical correlation structure for the relevant and irrelevant features in Simulation Study A, we distinguish them in Simulation Study B. In detail, three different structures are considered:

Structure B1: Let $Z_1, \cdots, Z_{p_n}$ and , $W_1, \cdots, W_{p_{0n}}$ be i.i.d. random vectors with distribution $N(0, I)$. The feature vectors are generated as:

$$X_j = \frac{Z_j + W_j}{\sqrt{2}}, \text{ for } j \in s_{0n}; \quad X_j = \frac{Z_j + \sum_{k \in s_{0n}} Z_k}{\sqrt{1 + p_{0n}}} \text{ for } j \notin s_{0n}.$$

Structure B2: The features in s_{0n} have constant pairwise correlation. Let $X_j, j \in s_{0n}$ be the causal feature vectors generated accordingly. For $j \notin s_{0n}$, the feature vectors are generated as:

$$X_j = \boldsymbol{\epsilon}_j + \frac{\sum_{k \in s_{0n}} X_k}{p_{0n}},$$

where $\boldsymbol{\epsilon}_j$'s are independent vectors from $\mathrm{N}(0, 0.08 * \mathbf{I}_n)$. Here the variance of $\boldsymbol{\epsilon}_j$ is set to 0.08 in order for the second term, which is correlated with causal features, to dominate the variance.

Structure B3: The features are generated in the same way as in Structure B2 except that the causal features are generated according to the covariance matrix Σ with $\Sigma_{ij} = \rho^{|i-j|}$ and s_{0n} set to $\{1, 2, \cdots, p_{0n}\}$.

Simulation Study C In this study, we consider variants of the settings in [92]. For all the examples, the standard deviation of the error term is $\sigma = 1.5$ and $(n, p_n, p_{0n}) = (100, 200, 15)$. Instead of positive signs for all the relevant features as in [92], we assume that the signs of the coefficients are i.i.d samples from $(-1)^u$, where $u \sim Bernoulli(p)$. Intuitively, compared with the original data settings, in these new examples, the marginal effect of each relevant feature and the total effect of all the relevant features are both weakened. The number of replication is 500.

Structure C1: $\Sigma = \begin{pmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{pmatrix}$, where Σ_1, Σ_2 have order $15 \times 15, (p_n - 15) \times (p_n - 15)$. They both have diagonal elements 1 and off-diagonal elements $\rho = 0.2$. The true coefficient vector is $\boldsymbol{\beta}_{0j} = 2.5$ for $1 \leq j \leq 5$, 1.5 for $6 \leq j \leq 10$, 0.5 for $11 \leq j \leq 15$ and 0 otherwise. *Structure C2:* The same as Structure C1 except that $\rho = 0.5$. *Structure C3:* Σ has the same structure as Structure C1 except that Σ_1, Σ_2 has order $25 \times 25, (p_n - 25) \times (p_n - 25)$ and $\Sigma_{ij} = \rho^{|i-j|}$ in both Σ_1, Σ_2. $\boldsymbol{\beta}_{0j} = 2.5$ for $1 \leq j \leq 5$, 1.5 for $11 \leq j \leq 15$ and 0 otherwise. $\rho = 0.2$. *Structure C4:* The same as Structure C3 except that $\rho = 0.5$. *Structure C5:* The same as Structure C1 except that Σ_1, Σ_2 has order $p_n \times p_n, 0$. *Structure C6:* The same as Structure C5 except that $\rho = 0.5$.

Simulation Study D In this study, the examples are adapted from those given in [97], which are briefly summarized as follows: in the linear model $\boldsymbol{y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ for $1 \leq i \leq n$.

Structure D1: For $1 \leq j \leq p_n$, $X_{i,j} = d_{i,j} + \eta w_i$, where $d_{i,j}$'s are i.i.d from distribution $N(0,1)$, w_i's are i.i.d from $N(0,1)$, $\boldsymbol{\epsilon}_i$'s are i.i.d from $N(0,\sigma^2)$. $d_{i,j}, w_i, \boldsymbol{\epsilon}_i$ are mutually independent. The parameters are $(\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \boldsymbol{\beta}_{03}, \boldsymbol{\beta}_{04}, \boldsymbol{\beta}_{05}) = (3, -3.5, 4, -2.8, 3.2)$ and $\boldsymbol{\beta}_{0j} = 0$ for $j > 5$. $(n, p_n, \sigma, \eta) = (100, 1000, 1, 2)$.

Structure D2: The same as D1 except that $\boldsymbol{\beta}_{0j} = 3.2$ for $j = 1, 2, 3, 4;\ 4.4,$ for $j = 5, 6;\ 3.5$ for $j = 7, 8, 9$ and $(n, p_n, \sigma, \eta) = (100, 1000, 1.5, 1)$.

Structure D3: For $1 \leq j \leq 10, X_{i,j}$'s are i.i.d from $N(0,1)$, $X_{i,j} = 0.25Z_{i,j} + \sqrt{0.75}\sum_{t=1}^{10} X_{i,t}$ for $j > 10$, where $Z_{i,j}$'s are i.i.d from $N(0,1)$ and are independent of $X_{i,j}, 1 \leq j \leq 10$. The true coefficient vector is

$$(\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \cdots, \boldsymbol{\beta}_{0,10}) = (3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)$$

and $(n, p_n, \sigma) = (100, 1000, 1)$.

It was verified in [97] that, for Structures D1 and D2, all the irrelevant features are weakly correlated with the relevant features in terms of $\|\Sigma^{-1}_{s_{0n},s_{0n}}\Sigma_{s_{0n},j}\|_1 < 1$ for all $j \notin s^c_{0n}$, but the sparse Riesz Condition was violated because $\max\limits_{1\leq|s|\leq\nu} \lambda_{\max}(\Sigma_{s,s}) = \dfrac{1+\nu\eta^2}{1+\eta^2}$, which is not bounded when $|s|$ diverges with n. In Structure D3,

$$|\Sigma_{j,s_{0n}}\Sigma^{-1}_{s_{0n},s_{0n}} Sign(\boldsymbol{\beta})|_1 > 1$$

for all $j \notin s_{0n}^c$ and $\min_{i \notin s_{0n}} |\boldsymbol{E}(X_i^\tau \boldsymbol{y})| \geq \max_{i \in s_{0n}} |\boldsymbol{E}(X_i^\tau \boldsymbol{y})|$. That is, the conditions for the selection consistency of sequential LASSO are violated.

The averaged PDR_n and FDR_n and their standard deviations in the parenthesis are reported in the Tables 7.3.1, 7.3.3, 7.3.2, 7.3.4, 7.3.5 and 7.3.6. Here are the conclusions we draw by investigating the results in Tables 7.3.1-7.3.6.

If the selection accuracy of the relevant features is of interest, for general cases, where all the features have a unique covariance structure as in simulation study A and cases where partially orthogonality condition is satisfied, as in simulation study C. No matter whether the signs of the signals' effects are consistent or not, SLasso+EBIC and FSR+EBIC are comparably the best two methods in terms of high PDR and low FDR. When the model is very sparse, like in simulation study A, SCAD+CV tends to select 6 to 7 more features ($p_{0n} = 8$) than the true size of relevant features. When the model becomes less sparse, like in simulation C, where $p_{0n} = 15, p_n = 200$, SCAD+CV is much better. ALasso+CV selects too many features into the model for all the cases, they have very high PDRs and also high FDRs, while SIS+SCAD is too conservative, especially when the model is less sparse. For examples within which uniform constant ρ or power decay correlation $\rho^{|i-j|}$ is one of the parameters, we also conduct simulations when $\rho = 0.3, 0.5$, our findings are similar as $\rho = 0.5$, which is presented here.

Structure A2 with Type II coefficient is one exception in which SIS+SCAD is among the top three best methods. In this situation, each relevant feature is

weakly or at most moderately correlated with at most two relevant features. Also, there are at most two irrelevant features which are weakly or at most moderately correlated with only one or two relevant features. The marginal effect of each relevant feature is also strengthened.

In simulation study B, all the irrelevant features have weak marginal but strong overall correlation with the relevant features. SLasso + EBIC is better than FSR+EBIC in terms of higher PDR and lower FDR, and the differences are quite significant. The same pattern prevails under all three structures B1, B2 and B3. But SCAD+CV performs slightly better than SLasso+EBIC. It is also remarkable to notice that in structure D3 of simulation study D, even when the sufficient conditions for the selection consistency of sequential Lasso is violated, SLasso+EBIC still has the best performance among all the methods.

As a stepwise feature selection procedure, the selection consistency of Forward Selection (FSR) is still unknown. In [166], the author proved the sure screening property of FSR with EBIC, which results in a good starting point for adaptive Lasso and SCAD with EBIC to determine the final model. The proposed procedure in [166] selects the model with minimum EBIC from the top n models generated from FSR. Consistent with [97], we find that this procedure always chooses the largest model under consideration if the model sequence is close to n. In principle, FSR serves as a screening procedure in [166]. Since adaptive Lasso in [92] and SCAD in [107] are capable of handling the ultra high feature space directly without

losing their oracle properties, the screening becomes less necessary. Our simulation result shows that the selection consistency of FSR with a certain stopping rule can be expected.

If prediction error is of interest, we can see that SCAD + CV has the smallest PMSE as well as small model size (MSize) for almost all the settings. ALasso+CV is comparable with SCAD + CV and is only slightly worse than SCAD + CV in terms of PMSE but has a much larger MSize. FSR+EBIC and SLasso+EBIC, which are specifically devised for the identification of relevant features, have much larger PMSE than the other three methods in all the settings, but their MSize are in general smaller. They themselves are comparable while SLasso+EBIC is slightly better than FSR+EBIC. The performance of SIS+SCAD+CV is between the two extremes. The simulation study demonstrates that, for the purpose of prediction, SCAD+CV is the best method, and that FSR + EBIC and SLasso+EBIC are not good for prediction.

Real Data Analysis: Rat Data

In this subsection, we apply the methods we are considering in Section 7.3.1 to a data set published in [143]. The original data set consists of expression levels of 120 rats from 31099 probes, including TRIM 32, which was recently found to cause Bardet-Biedl syndrom ([37]) and was the response variable in our analysis. It is important to detect the genes which are related to TRIM 32, such study has

been done by statisticians in various literatures such as [92], [107], [94], [56], [155] and [91]. Instead of the 31098 probes presented in the data, we will focus on the 18975 probes which are differently expressed as in all the above-mentioned articles. All the probes are standardized to have mean 0 and standard deviation 1 before further analysis.

Table 7.3.8 displays the genes detected by all the considered methods when the top 3000 probes with the largest variances are considered. It is worthy to note that, the only two genes, 1383110_at, 1392692_at, selected by SLasso+ EBIC, are also simultaneously selected by FSR+EBIC and ALasso+CV. One of these two, 1383110_at is detected by all the six methods (three different methods: Lasso+CV, Scaled Lasso, Scaled MC+, with two different p for each method) in [155], it was also reported in [92] for Lasso+CV and ALasso+CV. Another one, 1392692_at is also reported as one detected by Lasso+CV in [155].

For the purpose of comparison, we follow the ideas in the references, select 3000 probes with the largest variances firstly, then select the top p covariates with the largest correlation coefficients with TRIM32. These p covariates are used for computing the averaged model size and predictive mean square errors. For each replication, the 120 rats are randomly partitioned into two groups with sample size 100 and 20 respectively, they are referred to training data set and testing data. The number of replication is 100 and p is set to be 1000,2000,3000. The results are listed in Table 7.3.9. Compared with the findings in previous studies, we find

that our method detects much less genes.

Table 7.3.1 Results on Comparisons of SLasso and its Competitors: Structure A and Type I Coefficients with Size $n = 100$

Methods	Structure	MSize	PDR	FDR	MSE
ALasso+CV	A1	37.78 (12.54)	.999 (.009)	.767(.072)	13.714 (2.837)
SCAD+CV		14.9 (3.77)	.998 (.022)	.43(.142)	1.523 (2.943)
SIS+SCAD		4.99 (.07)	.488 (.098)	.219(.157)	25.135 (5.091)
FSR+EBIC		8.26 (1.37)	.963 (.149)	.063(.098)	67.973 (9.563)
SLasso+EBIC		8.39 (1.31)	.968 (.134)	.071(.098)	68.06 (9.541)
ALasso+CV	A2	26.77 (15.81)	.861 (.112)	.674(.142)	17.07 (3.334)
SCAD+CV		13.69 (6.08)	.724 (.189)	.513(.179)	18.178 (4.162)
SIS+SCAD		4.89 (.31)	.484 (.066)	.21(.086)	16.68 (3.142)
FSR+EBIC		4.96 (1.73)	.579 (.201)	.057(.104)	36.975 (8.388)
SLasso+EBIC		4.57 (1.59)	.511 (.176)	.091(.133)	36.939 (8.254)
ALasso+CV	A3	31.15 (13.18)	.967 (.063)	.706(.129)	18.632 (3.99)
SCAD+CV		15.02 (5.24)	.851 (.138)	.498(.161)	20.583 (5.519)
SIS+SCAD		4.85 (.39)	.347 (.069)	.427(.106)	23.728 (5.129)
FSR+EBIC		6.24 (1.74)	.726 (.201)	.064(.096)	39.126 (8.956)
SLasso+EBIC		5.98 (1.72)	.674 (.188)	.088(.116)	38.868 (8.885)
ALasso+CV	A4	37.29 (14.74)	.998 (.018)	.756(.087)	14.857 (4.042)
SCAD+CV		14.45 (4.13)	.996 (.021)	.403(.169)	12.207 (4.161)
SIS+SCAD		4.96 (.21)	.49 (.103)	.208(.168)	24.89 (5.388)
FS+EBIC		8.39 (1.39)	.964 (.132)	.072(.088)	65.511 (11.288)
SLasso+EBIC		8.37 (1.55)	.958 (.145)	.075(.092)	65.289 (11.232)

Table 7.3.2 Results on Comparisons of SLasso and its Competitors: Structure A and Type II Coefficients with Size $n = 100$

Methods	Structure	MSize	PDR	FDR	MSE
ALasso+CV	A1	35.46 (13.15)	.977 (.05)	.75(.091)	14.568 (3.117)
SCAD+CV		14.59 (3.753)	.967 (.063)	.438(.136)	12.027 (2.702)
SIS+SCAD		4.96(.21)	.431 (.082)	.303(.132)	21.122 (4.93)
FSR+EBIC		7.54 (1.42)	.888 (.146)	.055(.104)	20.48 (4.174)
SLasso+EBIC		7.72 (1.26)	.902 (.111)	.056(.081)	20.278 (3.575)
ALasso+CV	A2	21.66 (11.71)	.951 (.065)	.565(.187)	28.646 (5.42)
SCAD+CV		12.05 (4.92)	.744 (.129)	.431(.218)	34.789 (7.02)
SIS+SCAD		4.94 (.24)	.532 (.065)	.139(.097)	32.129 (5.174)
FSR+EBIC		5.36 (1.19)	.617 (.117)	.067(.106)	40.529 (8.624)
SLasso+EBIC		5.04 (1.26)	.572 (.122)	.077(.116)	42.224 (9.458)
ALasso+CV	A3	25.78 (11.28)	.968 (.058)	.646(.139)	20.92 (3.59)
SCAD+CV		14.09 (4.66)	.827 (.112)	.479(.175)	23.117 (5.111)
SIS+SCAD		4.86 (.38)	.323 (.083)	.469(.128)	27.779 (6.898)
FSR+EBIC		6.31 (1.17)	.733 (.131)	.063(.101)	42.073 (8.191)
SLasso+EBIC		6.16 (1.47)	.692 (.148)	.089(.106)	43.327 (8.517)
ALasso+CV	A4	35.98 (13.44)	.978 (.054)	.756(.082)	15.535 (4.129)
SCAD+CV		14.13 (3.62)	.96 (.07)	.423(.143)	13.594 (4.358)
SIS+SCAD		4.89 (.34)	.425 (.088)	.304(.134)	22.682 (6.318)
FS+EBIC		7.72 (1.32)	.886 (.124)	.073(.093)	21.45 (5.584)
SLasso+EBIC		7.81 (1.34)	.893 (.127)	.078(.097)	21.516 (5.653)

Table 7.3.3 Results on Comparisons of SLasso and its Competitors: Structure A and Type I Coefficients with Size $n = 200$

Methods	Structure	MSize	PDR	FDR	MSE
ALasso+CV	A1	49.05 (17.96)	1.00 (.000)	.791(.077)	10.937 (1.463)
SCAD+CV		13.65 (4.79)	1.00 (.000)	.283(.183)	8.638 (.897)
SIS+SCAD		8.74 (.46)	.793 (.077)	.181(.086)	12.355 (3.309)
FSR+EBIC		9.37 (.65)	1.00 (.000)	.035(.06)	58.279 (6.013)
SLasso+EBIC		9.37 (.66)	1.00 (.000)	.035(.061)	58.282 (6.012)
ALasso+CV	A2	40.57 (19.77)	.941 (.072)	.735(.14)	15.297 (2.03)
SCAD+CV		23.71 (7.32)	.931 (.11)	.612(.127)	14.159 (2.928)
SIS+SCAD		8.06 (.79)	.661 (.028)	.255(.076)	14.715 (1.715)
FSR+EBIC		7.92 (1.69)	.846 (.179)	.035(.07)	36.3 (6.372)
SLasso+EBIC		7.83 (2.09)	.796 (.19)	.073(.1)	35.832 (6.151)
ALasso+CV	A3	62.88 (19.69)	1.00 (.000)	.845(.042)	9.816 (1.41)
SCAD+CV		14.53 (5.37)	.999 (.008)	.313(.201)	7.569 (1.237)
SIS+SCAD		7.85 (.96)	.437 (.068)	.495(.081)	15.528 (1.858)
FSR+EBIC		9.11 (.99)	.977 (.096)	.032(.059)	41.214 (4.95)
SLasso+EBIC		8.96 (1.62)	.941 (.162)	.051(.079)	40.37 (5.611)
ALasso+CV	A4	50.23 (16.35)	1.00 (.000)	.801(.068)	11.27 (2.001)
SCAD+CV		13.86 (4.00)	1.00 (.000)	.303(.173)	9.294 (1.788)
SIS+SCAD		8.69 (.56)	.764 (.086)	.206(.097)	13.558 (3.708)
FS+EBIC		9.61 (.94)	1.00 (.000)	.056(.078)	59.198 (6.788)
SLasso+EBIC		9.56 (.84)	1.00 (.000)	.052(.074)	59.174 (6.799)

Table 7.3.4 Results on Comparisons of SLasso and its Competitors: Structure A and Type II Coefficients with Size $n = 200$

Methods	Structure	MSize	PDR	FDR	MSE
ALasso+CV	A1	47.8 (16.98)	.995 (.023)	.79(.072)	12.646 (1.796)
SCAD+CV		15.99 (6.20)	.988 (.034)	.375(.195)	10.653 (1.317)
SIS+SCAD		8.27 (.86)	.648 (.083)	.288(.106)	13.92 (3.073)
FSR+EBIC		9.04 (.78)	.966 (.051)	.034(.058)	18.349 (2)
SLasso+EBIC		9.03 (.78)	.966 (.051)	.033(.058)	18.329 (1.996)
ALasso+CV	A2	26.4 (13.25)	.972 (.05)	.593(.18)	27.529 (3.399)
SCAD+CV		19.08 (7.67)	.837 (.109)	.54(.189)	30.767 (4.664)
SIS+SCAD		8.5 (.56)	.796 (.061)	.156(.051)	24.9 (2.518)
FSR+EBIC		7.11 (1.14)	.754 (.108)	.04(.073)	35.261 (4.272)
SLasso+EBIC		6.86 (1.34)	.709 (.113)	.06(.085)	36.411 (4.734)
ALasso+CV	A3	34.24 (15.34)	.986 (.037)	.693(.125)	19.966 (2.312)
SCAD+CV		18.42 (6.77)	.895 (.081)	.508(.166)	20.096 (2.766)
SIS+SCAD		8.39 (.63)	.576 (.069)	.384(.05)	17.891 (2.17)
FSR+EBIC		7.76 (.91)	.829 (.09)	.034(.063)	38.836 (4.236)
SLasso+EBIC		7.73 (1.17)	.802 (.098)	.058(.083)	39.176 (4.492)
ALasso+CV	A4	47.08 (16.78)	.995 (.023)	.787(.072)	12.931 (2.468)
SCAD+CV		15.76 (4.71)	.986 (.038)	.39(.168)	11.115 (2.159)
SIS+SCAD		8.32 (.85)	.641 (.089)	.301(.107)	14.62 (3.599)
FS+EBIC		9.25 (1.08)	.959 (.06)	.059(.08)	18.815 (2.653)
SLasso+EBIC		9.26 (1.03)	.962 (.057)	.058(.079)	18.797 (2.645)

Table 7.3.5 Results on Comparisons of SLasso and Its Competitors: Structure B with Type I coefficients

Methods	Structure	MSize	PDR	FDR	MSE
		$n = 100$			
ALasso+ CV	B1	15.12 (8.12)	.999 (.009)	.356(.242)	4.676 (.82)
SCAD+CV		8.09 (.69)	.979 (.099)	.029(.099)	4.45 (.933)
SIS+SCAD		4.97 (.26)	.543 (.068)	.126(.1)	17.033 (3.178)
FSR+EBIC		8.15 (.83)	.745 (.19)	.27(.169)	68.821 (9.704)
SLasso+EBIC		8.25 (.96)	.932 (.153)	.092(.143)	64.088 (8.149)
ALasso+ CV	B2	11.17 (5.16)	1.00 (.000)	.188(.225)	2.312 (.419)
SCAD+CV		7.99 (.07)	.999 (.009)	.000(.000)	2.582 (1.049)
SIS+SCAD		4.87 (.36)	.527 (.076)	.132(.115)	10.02 (2.652)
FSR+EBIC		6.69 (3.45)	.779 (.398)	.049(.086)	32.469 (8.603)
SLasso+EBIC		6.54 (3.46)	.784 (.409)	.028(.059)	31.718 (8.4)
ALasso+ CV	B3	11.21 (5.17)	1.00 (.000)	.19(.226)	5.196 (.842)
SCAD+CV		7.99 (.1)	.999 (.012)	.000(.000)	5.147 (1.031)
SIS+SCAD		4.98 (.14)	.504 (.034)	.19(.046)	10.324 (1.601)
FSR+EBIC		7.29 (1.8)	.782 (.167)	.124(.104)	39.785 (11.271)
SLasso+EBIC		7.54 (1.71)	.911 (.188)	.027(.057)	35.651 (7.288)
		$n = 200$			
ALasso+ CV	B1	25.53 (15.91)	.956 (.071)	.507(.283)	4.205 (.539)
SCAD+CV		9.09 (1.09)	.972 (.121)	.031(.124)	3.963 (.62)
SIS+SCAD		8.92 (.39)	.864 (.064)	.128(.046)	4.498 (1.987)
FSR+EBIC		9.19 (.91)	.708 (.206)	.311(.183)	58.925 (6.083)
SLasso+EBIC		9.22 (.99)	.873 (.209)	.148(.19)	55.636 (5.817)
ALasso+ CV	B2	13.3 (6.41)	1.00 (.000)	.215(.242)	2.186 (.267)
SCAD+CV		9 (0)	1.00 (.000)	.000(.000)	2.32 (.753)
SIS+SCAD		8.72 (.74)	.449 (.064)	.535(.061)	3.327 (1.679)
FSR+EBIC		9.33 (.61)	.993 (.043)	.037(.074)	31.14 (4.188)
SLasso+EBIC		9.25 (.56)	1.00 (.000)	.023(.052)	30.799 (3.538)
ALasso+ CV	B3	15.68 (9.54)	.986 (.044)	.276(.284)	5.303 (.622)
SCAD+CV		8.99 (.100)	.999 (.011)	.000(.000)	5.199 (.71)
SIS+SCAD		7.77 (.83)	.681 (.066)	.206(.07)	7.975 (1.258)
FSR+EBIC		9.37 (.59)	.943 (.086)	.091(.1)	35.593 (4.762)
SLasso+EBIC		9.26 (.59)	1.00 (.000)	.024(.054)	33.636 (3.546)

Table 7.3.6 Results on Comparisons of SLasso and its Competitors: Structure C

Methods	Structure	MSize	PDR	FDR	MSE
ALasso+CV	C1	39.59(8.37)	.956 (.048)	.621(.085)	3.918 (.689)
SCAD+CV		10.81 (1.01)	.687 (.041)	.042(.064)	5.392 (1.288)
SIS+SCAD		5 (0)	.333 (0)	0(0)	10.802 (1.129)
FSR+EBIC		11.49 (1.09)	.755 (.061)	.012(.038)	58.15 (4.487)
SLasso+EBIC		10.97 (1.77)	.708 (.101)	.027(.059)	58.613 (5.778)
ALasso+CV	C2	50.71 (13.19)	.846 (.078)	.735(.062)	4.645 (.863)
SCAD+CV		12.07 (2.26)	.711 (.064)	.1(.102)	3.843 (1.077)
SIS+SCAD		2.78 (.63)	.145 (.035)	.201(.163)	20.282 (1.403)
FSR+EBIC		10.22 (2.15)	.672 (.138)	.011(.036)	54.249 (8.709)
SLasso+EBIC		9.10 (3.11)	.589 (.194)	.022(.054)	49.325 (13.343)
ALasso+CV	C3	54.69 (11.14)	.909 (.045)	.741(.05)	7.041 (1.381)
SCAD+CV		17.52 (4.09)	.838 (.053)	.25(.147)	4.141 (.943)
SIS+SCAD		5 (0)	.332 (.009)	.004(.028)	18.665 (1.569)
FSR+EBIC		12.43 (1.67)	.773 (.115)	.067(.09)	41.013 (3.291)
SLasso+EBIC		10.81 (3.15)	.639 (.206)	.115(.121)	40.565 (4.267)
ALasso+CV	C4	56.49 (14.67)	.881 (.077)	.752(.059)	6.802 (1.371)
SCAD+CV		17.12 (3.95)	.792 (.065)	.277(.136)	3.858 (.997)
SIS+SCAD		4.75 (.55)	.259 (.041)	.175(.127)	15.327 (2.527)
FSR+EBIC		10.93 (2.22)	.696 (.128)	.04(.062)	58.294 (4.162)
SLasso+EBIC		10.52 (2.63)	.66 (.155)	.053(.073)	57.878 (4.517)
ALasso+CV	C5	47.77 (9.03)	.873 (.064)	.717(.053)	8.05 (1.131)
SCAD+CV		14.05 (2.96)	.708 (.095)	.223(.13)	8.104 (1.4)
SIS+SCAD		5 (0)	.241 (.038)	.276(.114)	20.799 (2.507)
FSR+EBIC		11.6 (1.76)	.693 (.093)	.097(.085)	52.102 (4.754)
SLasso+EBIC		11.41 (2.14)	.655 (.101)	.128(.097)	51.154 (5.147)
ALasso+CV	C6	45.34 (13.43)	.858 (.072)	.695(.079)	5.514 (.853)
SCAD+CV		13.00 (2.82)	.731 (.078)	.135(.115)	4.237 (.893)
SIS+SCAD		3.03 (.67)	.198 (.042)	.016(.062)	21.581 (1.881)
FSR+EBIC		11.19 (1.29)	.705 (.066)	.05(.067)	55.287 (3.741)
SLasso+EBIC		11.19 (2.07)	.677 (.109)	.083(.082)	54.693 (5.639)

Table 7.3.7 Results on Comparisons of SLasso and its Competitors: Structure D

Methods	Structure	MSize	PDR	FDR	MSE
ALasso+ CV	D1	28.92 (8.75)	1.00 (.000)	.807(.077)	3.241 (2.227)
SCAD+CV		4.88 (.33)	.975 (.066)	.000(.000)	7.333 (9.033)
SIS+SCAD		3.07 (.84)	.366 (.075)	.378(.163)	34.422 (5.55)
FSR+EBIC		5.26 (.74)	.992 (.08)	.047(.088)	267.434 (37.092)
SLasso+EBIC		5.09 (.96)	.972 (.147)	.035(.082)	264.058 (44.554)
ALasso+ CV	D2	45.71 (7.34)	1.00 (.000)	.798(.034)	8.956 (6.331)
SCAD+CV		9.00 (.19)	.998 (.016)	.002(.022)	15.109 (11.959)
SIS+SCAD		4.98 (.16)	.136 (.112)	.754(.205)	239.719 (63.737)
FSR+EBIC		11.5 (2.16)	.942 (.205)	.254(.19)	61.597 (38.211)
SLasso+EBIC		12.38 (2.87)	.871 (.297)	.373(.224)	77.068 (64.969)
ALasso+ CV	D3	69.64 (5.92)	.852 (.051)	.877(.012)	7.893 (4.293)
SCAD+CV		8.77 (2.57)	.583 (.104)	.308(.125)	28.737 (10.868)
SIS+SCAD		4.29 (.71)	.000 (.000)	1.00(.000)	58.334 (10.717)
FSR+EBIC		18.15 (2.98)	.785 (.122)	.561(.075)	45.766 (15.38)
SLasso+EBIC		9.82 (3.49)	.754 (.31)	.262(.146)	85.231 (33.219)

Table 7.3.8 Rat Data: The Gene Probes Selected by All Considered Methods

Methods	Probes ID
ALasso+CV	1387060_at, 1388538_at, 1380070_at, 1370052_at, 1382452_at, 1379079_at, 1397489_at, 1374131_at, 1383110_at, 1389584_at, 1392692_at, 1379971_at 1385687_at, 1369353_at, 1374106_at, 1383673_at, 1379495_at, 1383749_at 1382835_at, 1395415_at, 1383996_at.
SCAD+CV	1394689_at, 1370434_a_at, 1375724_at, 1378765_at, 1375139_at, 1388538_at 1370052_at, 1382452_at, 1377781_at, 1383841_at, 1380311_at, 1379460_at, 1385921_at, 1384886_at, 1384136_at, 1387111_at, 1390789_at, 1376693_at, 1389584_at, 1389231_at, 1390788_a_at 1367741_at, 1374106_at, 1387455_a_at 1383749_at, 1379803_at, 1383996_at, 1382633_at
SIS+SCAD	1377546_at, 1396809_at, 1381430_at, 1393543_at, 1372481_at
FSR+EBIC	1383110_at, 1392692_at, 1389584_at
SLasso+EBIC	1383110_at, 1392692_at

Table 7.3.9 Rat Data: The Averaged Number of Selected Genes and Prediction Error with Different Numbers of The Considered Genes

	$p = 1000$		$p = 2000$		$p = 3000$	
Methods	MSize	MSE	MSize	MSE	MSize	MSE
ALasso+CV	46.27	.507	45.61	.525	40.78	.552
	(21.17)	(.375)	(30.17)	(.399)	(29.34)	(.428)
SCAD+CV	14.56	.61	16.25	.628	15.64	.635
	(4.91)	(.456)	(5.89)	(.465)	(6.21)	(.489)
SIS+SCAD	4.08	.566	4.08	.566	4.08	.566
	(.8)	(.337)	(.8)	(.337)	(.8)	(.337)
FSR+ EBIC	3.27	.871	3.06	.888	2.91	.907
	(.89)	(.463)	(.89)	(.465)	(.81)	(.473)
SLasso+EBIC	2.94	1.069	2.61	1.025	2.45	1.03
	(1.09)	(.501)	(.96)	(.463)	(.87)	(.47)

CHAPTER 8

Sure Screening Property of Sequential LASSO

When the dimension of the predictor space is ultra-high or the number of true features diverges as sample size increases, some researchers may argue that it is unrealistic to guarantee the selection consistency for a stepwise feature selection algorithm ([61], [166]). In this chapter, we will show that sequential LASSO can also serve as an efficient screening procedure as SIS and Forward Selection for its sure screening property stated in Theorem 8.0.2. For simplicity, we assume that the variables enter the model one by one.

Assumption 8.0.1. *We assume that all the predictors are standardized to have mean* 0 *and standard deviation* 1 *and the following conditions are satisfied,*

(A1.) There exists $p_0 > 0, a \geq 0, 0 \leq b < 0.5, L_\beta, U_\beta > 0$ *such that*

$$p_{0n} \leq p_0 n^a;\ \boldsymbol{\beta}_{0,\min} \geq L_\beta n^{-b};\ \boldsymbol{\beta}_{0,\max} \leq U_\beta,$$

where $\boldsymbol{\beta}_{0,\min} = \min_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}|$, $\boldsymbol{\beta}_{0,\max} = \max_{j \in s_{0n}} |\boldsymbol{\beta}_{0j}|$.

(A2.) There exists $c_1 > 0$ *such that* $\lambda_{\max}(X^\tau(s_{0n})X(s_{0n})) \leq nc_1$.

(A3.) Define

$$\lambda_m = \min_{|s| \leq m} \lambda_{\min}\left(X^\tau(s \cup s_{0n})X(s \cup s_{0n})\right),$$

where $\lambda_{\min}$ *refers to the minimum eigenvalue, there exists* $c_2 > 0$ *such that*

$$\min\{\lambda_m : m = O(n^{a+2b})\} \geq nc_2.$$

Parallel to the selection consistency of sequential LASSO, our results in Theorem 8.0.2 apply to both deterministic and randomly generated design matrices. For random design matrix, assume the covariance matrix of the covariates is Σ. As a corollary of Lemma 1 in [166], assume there exists a constant p and $\kappa > 0$ such that

$$\ln(p_n) \leq pn^\kappa, \kappa + 3a + 6b < 1,$$

then $(A3.)$ in Assumption 8.0.1 holds if $\lambda_{\min}(\Sigma) > c$ for some constant $c > 0$.

Theorem 8.0.2. *Under Assumption 8.0.1, let λ_0 be the value defined in the Appendix, define*

$$K = \left[\frac{(1+\lambda_0)^2 c_1 U_\beta^2 p_0}{c_2^2 L_\beta^2} n^{a+2b} \right] + 1,$$

where $[\cdot]$ denotes the integer part of a real number, we have

$$P(\exists \{\lambda_i^\star\}_{i=1}^n \ \ such\ that\ \ s_{0n} \subseteq s_{\star K}) \to 1.$$

Proof of Theorem 8.0.2. By examining the proof of inequality (7.3.7), we can easily find that this inequality holds for any k as long as $s_{\star k}^- \neq \emptyset$. Therefore, if $s_{\star K}^- \neq \emptyset$, when $0 \leq i \leq K-1$,

$$\begin{aligned} \frac{\triangle^\mu(s_{\star i+1})^2}{(1+\lambda_0)^2 n \|\boldsymbol{\beta}_0(s_{\star i+1}^-)\|_1^2} \geq & \frac{\lambda_K^2 \|\boldsymbol{\beta}_0(s_{\star i+1}^-)\|_2^4}{(1+\lambda_0)^2 n |s_{\star i+1}^-| \|\boldsymbol{\beta}_0(s_{\star i+1}^-)\|_2^2} = \frac{\lambda_K^2 \|\boldsymbol{\beta}_0(s_{\star i+1}^-)\|_2^2}{(1+\lambda_0)^2 n |s_{\star i+1}^-|} \\ \geq & (\frac{\lambda_K^2}{(1+\lambda_0)^2 n}) \left(\frac{\|\boldsymbol{\beta}_0(s_{\star i+1}^-)\|_1}{|s_{\star i+1}^-|} \right)^2 \geq \frac{(\lambda_K \boldsymbol{\beta}_{0,\min})^2}{(1+\lambda_0)^2 n}. \end{aligned} \tag{8.0.1}$$

Therefore,

$$\triangle^\mu(S) - \triangle^\mu(s_{\star K}) = \sum_{i=0}^{K-1} (\triangle^\mu(\mathcal{S}_i) - \triangle^\mu(s_{\star i+1})) \geq \frac{K(\lambda_K \boldsymbol{\beta}_{0,\min})^2}{(1+\lambda_0)^2 n}. \tag{8.0.2}$$

On the other hand,

$$\triangle^\mu(S) - \triangle^\mu(s_{\star K}) \leq \triangle^\mu(S) \leq n c_1 \|\boldsymbol{\beta}_0\|_2^2 \leq c_1 n p_{0n} (\boldsymbol{\beta}_{0,\max})^2.$$

Under Assumption 8.0.1, by plugging in the K which is defined in Theorem 8.0.2, we have

$$\frac{K(\lambda_K \boldsymbol{\beta}_{0,\min})^2}{(1+\lambda_0)^2 n} \geq c_1 n p_{0n} (\boldsymbol{\beta}_{0,\max})^2.$$

It is a contradiction. Therefore, $s^-_{\star K} = \emptyset$. Equivalently, all the true causal features are covered in the first K steps. □

For the purpose of comparison, let K be the smallest number of steps Forward Selection needs to bring in all the true causal features, a, b, κ has the same meaning as we defined in Assumption 8.0.1. If $\lambda_{\min}(\Sigma) > c$ is assumed, the following relationship is needed in [166]:

$$a \leq 2b;\ \kappa + 6a + 12b < 1;\ K = O(n^{2a+4b}).$$

We can see from Theorem 8.0.2 that there is indeed a remarkable improvement for sequential LASSO compared with Forward Selection.

After successfully selecting all the relevant features, we make use of EBIC_γ to define a subset $\hat{s}$ of $s_{\star K}$ by

$$\hat{s} = \{j : \text{EBIC}_\gamma(s_{\star K} - \{j\}) > \text{EBIC}_\gamma(s_{\star K}), j \in s_{\star K}\}. \tag{8.0.3}$$

Theorem 8.0.3. *Under Assumption 8.0.1, suppose* $\ln p_n = O(n^\kappa)$, *when* $\kappa < 1 - 2b$

and $\gamma > 0$*, we have*

$$\lim_{n\to+\infty} P(\hat{s} = s_{0n}) = 1.$$

Proof of Theorem 8.0.3. By definition (8.0.3), we have

$$\begin{aligned} P\left(\hat{s} \neq s_{0n}\right) =& P\left(\exists j \in s_{\star K}, |\boldsymbol{\beta}_{0j}| = 0, \mathrm{EBIC}_\gamma(s_{\star K} - \{j\}) > \mathrm{EBIC}_\gamma(s_{\star K})\right) \\ &+ P\left(\exists j \in s_{\star K}, |\boldsymbol{\beta}_{0j}| \neq 0, \mathrm{EBIC}_\gamma(s_{\star K} - \{j\}) < \mathrm{EBIC}_\gamma(s_{\star K})\right) \\ =& P_1 + P_2. \end{aligned} \tag{8.0.4}$$

Now we calculate P_1, P_2 separately: note that $s_{0n} \subset s_{\star K}$, hence $\triangle^\mu(s_{\star K}) = 0$, let $\boldsymbol{J}_1 = s_{\star K}, \boldsymbol{J}_2 = s_{\star K} - \{j\}$, T_2 in (7.3.3) equals to $\ln n + 2\gamma \ln p_n$.

1. If $|\boldsymbol{\beta}_{0j}| = 0$, by noting $s_{0n} \subset s_{\star K} - \{j\}$, we know that $\triangle^\mu(s_{\star K} - \{j\}) = 0$. Therefore, T_1 in (7.3.3) equals to

$$n \ln\left\{1 - \frac{\triangle^\epsilon(s_{\star K} - \{j\}) - \triangle^\epsilon(s_{\star K})}{\triangle^\epsilon(s_{\star K} - \{j\})}\right\}.$$

By using similar techniques as in Proof of (I) in Theorem 7.3.1, we know that

$$\max_{j\in s_{\star K}} \left(\triangle^\epsilon(s_{\star K} - \{j\}) - \triangle^\epsilon(s_{\star K})\right) = O_p(\ln n),$$

$$\min_{j\in s_{\star K}} \triangle^\epsilon(s_{\star K} - \{j\}) = n\left(1 + o_p(1)\right).$$

By noting that $\lim_{x\to 0}\dfrac{\ln(1-x)}{-x}=1$ and $\dfrac{\ln n}{n}\to 0$, we have $T_1+T_2>0$ with probability tending to 1 when $\gamma>0$, that is, $P_1\to 0$.

2. If $|\boldsymbol{\beta}_{0j}|\neq 0$, denote

$$A_{j,s}=X_j^\tau[\boldsymbol{I}-H_0(s-\{j\})]X_j\;\; B_{j,s}=X_j^\tau[\boldsymbol{I}-H_0(s-\{j\})]\boldsymbol{\epsilon}_n.$$

Note that $\max_{1\le j\le p_n}\left(\triangle^\epsilon(s_{\star K}-\{j\})-\triangle^\epsilon(s_{\star K})\right)=O_p(\ln p_n)$, therefore,

$$\begin{aligned}-T_1=&n\ln\left\{1+\frac{\boldsymbol{\beta}_{0j}^2A_{j,s_{\star K}}+2\boldsymbol{\beta}_{0j}B_{j,s_{\star K}}+\triangle^\epsilon(s_{\star K}-\{j\})-\triangle^\epsilon(s_{\star K})}{\triangle^\epsilon(s_{\star K})}\right\}\\=&n\ln\left\{1+\frac{\boldsymbol{\beta}_{0j}^2A_{j,s_{\star K}}+2\sqrt{\boldsymbol{\beta}_{0j}^2A_{j,s_{\star K}}}O_p(\ln n)+O_p(\ln p_n)}{n}\right\}.\end{aligned}$$

By definition and A3,

$$A_{j,s_{\star K}}\ge\lambda_{\min}\left(X^\tau(s_{\star K})X(s_{\star K})\right)\ge nc_2.$$

Therefore, we know that there exists a positive constant C such that

$$-T_1\ge n\ln\left\{1+Cn^{-2b}\right\}\ge C'n^{1-2b}\ge T_2.$$

when $\gamma>0$ and $\kappa+2b<1$. Hence, $T_1+T_2<0$ with probability tending to 1, which means $P_2\to 0$.

□

Theorem 8.0.2 together with Theorem 8.0.3 provides a different way of sequential LASSO coupled with EBIC to achieve selection consistency. Intuitively, this new procedure requires much weaker assumptions than Theorems 7.1.1 and 7.2.1 because it allows irrelevant features to enter the model in the stepwise selection process. However, it is challenging to find a feasible way to approach K. We leave this work to the future.

Conclusion and Discussion for Part II

Stepwise feature selection procedures are appealing for their computational advantages. In this part, we proposed a new stepwise feature selection procedure, sequential LASSO and explored its properties.

In the literature, Efron et al proposed a sequential procedure called least angle regression (LAR) in [53]. With slight modification, the algorithm of LAR can also compute the solution path of LASSO sequentially, which made LASSO more popular. The classical forward stepwise regression (FSR) has been recently re-examined in [166] on its properties in feature selection with ultra-high dimensional feature space. A different version of forward stepwise regression referred to as

forward selection in [170] has been re-considered recently and dubbed as orthogonal matching pursuit (OMP), see [158], [159], [30]. Their differences can be seen from numerical study results in Section 7.3.2. We can also analyze these differences theoretically.

First, consider the difference between the sequential LASSO and FSR. After the sub model s_{*k} is selected, the sequential LASSO selects the next feature among the features that maximize

$$g_1(j) = |X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{y}_n|,$$

see the proof of Theorem 7.1.1. The FSR selects the next feature by minimizing $\text{RSS}(j) = \boldsymbol{y}_n^\tau[\boldsymbol{I} - H_0(s_{*k} \cup \{j\})]\boldsymbol{y}_n$ which is equivalent to maximizing

$$g_2(j) = \frac{|X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{y}_n|}{\sqrt{X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]X(\{j\})}}.$$

The equivalence is established by the following identity,

$$\boldsymbol{I} - H_0(s_{*k} \cup \{j\}) = [\boldsymbol{I} - H_0(s_{*k})]\left(\boldsymbol{I} - \frac{X(\{j\})X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]}{X^\tau(\{j\})[\boldsymbol{I} - H_0(s_{*k})]X(\{j\})}\right).$$

The sequential LASSO selects the next feature that has the highest correlation with the current residual $[\boldsymbol{I} - H_0(s_{*k})]\boldsymbol{y}_n$, but the FSR selects the next feature that has the highest inflated correlation with an inflating factor $[X^\tau(\{j\})[\boldsymbol{I} -$

$H_0(s_{*k})]X(\{j\})]^{-1/2}$. If $X(\{j\})$ is orthogonal to $\mathscr{R}(s_{*k})$, the factor is a constant (note that the $X(\{j\})$'s are standardized), but larger than the constant otherwise. The more correlated the $X(\{j\})$ is with the features in s_{*k}, the larger the inflating factor. If two features have the same absolute correlation with the current residual, the FSR will select the one that is more correlated with the features in s_{*k}. If one feature has a lower correlation with the current residual but is more correlated with the features in s_{*k} than another feature, it might turn out that this feature has a higher inflated correlation and is selected by FSR. Obviously, this is a disadvantage of FSR, especially when high spurious correlations present in small-n-large-p problems.

The OMP selects the next feature (or features) maximizing $g_1(j)$. At steps where there is only one feature that maximizes $g_1(j)$, the sequential LASSO and the OMP select the same next feature. But at steps where there are more than one features that maximize $g_1(j)$, there is a difference between the sequential LASSO and the OMP. The OMP selects all those features. But the sequential LASSO selects them all subject to a partial positive cone condition, see the proof of Theorem 7.1.1. If the partial positive cone condition is not satisfied, the sequential LASSO generally does not select all those features. The sequential LASSO can be easily extended as a sequential penalized likelihood method for generalized linear models but there is no obvious way by which the OMP can be extended.

Under the well-known irrepresentable condition, the LASSO has been shown in

[183] to possess the property of selection consistency while the penalty parameter is properly chosen . If the covariance matrix of the vector of the covariates has eigenvalues bounded both from above and away from zero in addition to some other assumptions, it is established in [166] that the FSR has the sure screening property when the procedure is carried out at a certain step before the number of steps reaches the sample size. The OMP has been studied under conditions called Exact Recovery Condition (ERC) in [31],[158] and Mutual Incoherence Property (MIP) in [30]. The ERC is similar to the irrepresentable condition but much stronger. The MIP is the condition that $\rho_{\max} < \frac{1}{2k-1}$ where $\rho_{\max}$ is the largest absolute correlation among all pairs of covariates and k is the number of causal covariates. The ERC implies MIP, see [158], [30]. Both the sure screening property and the selection consistency of OMP have been examined in [30] under MIP together with other conditions. Our theories suggest that the conditions for the sequential LASSO to be selection consistent may be much weaker than the original LASSO (Theorems 7.1.1 and 7.2.1).

The sequential LASSO bears some similarity with OMP. At steps where a partial positive cone condition is satisfied, the sequential LASSO selects new features with the same criterion as OMP. The properties established for the sequential LASSO then apply to OMP. Thus, we reveal some new properties of OMP other than those discovered in [158], [159] and [30]. The stopping rule is given by the extended BIC (EBIC) proposed in [33]. The selection consistency of EBIC in the

same situation is established in Section 3.1. The selection consistency of this whole procedure is shown and provided in Theorem 7.3.1. Thus, coupled with EBIC the sequential LASSO provides a practically applicable selection consistent method for feature selection in small-n-large-p problems.

For the ultra-high dimensional case, generally, from the proof of Theorem 7.1.1, we note that the tuning parameter λ in each step of sequential LASSO has to be of order $O(n)$, which is the same as $\|\boldsymbol{y}_n - X(S)\boldsymbol{\beta}(S)\|_2^2$. It means that the penalty on the complexity of the model is as important as the prediction error, which is one major difference between sequential LASSO and forward regression. It also indicates that we have to sacrifice the prediction error to single out only one causal feature, which is acceptable if we only concentrate on the selection of causal features. In [109], the authors showed theoretically that the LASSO estimator and the least squares estimator have ignorable bias if and only if $\lambda = o(n)$.

CHAPTER 9

Conclusions and Future Work

9.1 Conclusions of This Thesis

In Part I of this thesis, we extended the current study of EBIC to more complex models such as linear regression models with ultra-high dimensional space, generalized linear regression models with non-canonical links and Cox's proportional hazard models. We proved the selection consistency of EBIC in these models under acceptable conditions and applied EBIC to a general feature selection procedure in high-dimensional studies. Our extensive numerical study strongly recommends that in high dimensional studies, EBIC with a proper chosen γ is effective in model selection.

In Part II of this thesis, we managed to overcome the impact of high spurious correlation among features in feature selection using our newly proposed method-sequential LASSO. As argued by many researchers, high spurious correlation among features is an intrinsic phenomenon in feature selection and it is difficult to avoid. This thesis provides a promising and feasible direction for future research. Our theory verified that the assumptions to avoid spurious correlation in previous studies can be relaxed a lot for sequential LASSO being selection consistent.

The finite sample performance of a feature selection procedure is assessed by the positive discovery rate (PDR) and false discovery rate (FDR) as defined in [33]. Equivalently, the asymptotic property of selection consistency means that PDR converges to 1 and FDR converges to 0 simultaneously as the sample size goes to infinity. When EBIC's selection consistency was evaluated, we compared its performance mainly with BIC and mBIC in simulation studies because EBIC is an extension of BIC and mBIC. It manifests from the simulation study results that the finite sample performance of the EBIC closely matches its asymptotic property. Under all data structures and models, for the EBIC with a theoretically suggested γ value, the PDR and the FDR approach rapidly to 1 and 0 respectively, as the sample size increases. The BIC does not appear to be selection consistent and mBIC loses certain power while overly controlling FDR when the sample size is small.

Extensive simulation studies were conducted to show comparisons between sequential LASSO and other techniques. Instead of imposing the unique correlation structure on all predictors, we distinguish the linear relationships among the true causal features from those between true features and uncausal features. PDR and FDR are applied to show the selection consistency of sequential LASSO with EBIC as the stopping rule. Prediction errors are provided as well. We can see that sequential LASSO has the best behavior from the aspect of identifying relevant features. It can screen out the uncausal features even when they are strongly correlated with the true causal features. But it is not strongly recommended if pursuing high prediction accuracy is the goal of study.

Feature selection in regression problems under high or ultra-high feature spaces arise in many important fields of scientific research. Our study provides an integrated approach to conduct feature selection in these regression problems.

9.2 Open Questions for Future Research

In conclusion, we have made contributions in feature selection under high or ultrahigh dimensional feature spaces. For this forefront and challenging problem, more effort is indispensable in the future. Firstly, since our main objective was to develop a good statistical method, especially suitable for QTL mapping, real data analysis is not extensively conducted in our work. To make our statements more

convincing and persuasive, future work should involve more diverse applications, such as QTL mapping in etiological studies and eQTL mapping in micro-array data analysis. Secondly, our current work mainly focused on single response. Consequently, it is a natural question to check the applicability of these procedures when we are facing multiple responses.

Being a preliminary work, for survival models, our work is limited to a representative model-Cox proportional hazards model. Moreover, our work on CPH is constrained to the situation where dimension of the feature space is of polynomial order of sample size and constant number of true features. Like in linear regression models, a direct extension of this work would be to consider more general parameter settings and models.

In our simulations in Section 7.3.2, we can see that for ALasso+CV and SCAD+CV, their high PDR_n's confirm the screening property of the determined set. More generally, it is much easier to ensure the screening property for many greedy algorithms or regularization methods. Note that for most situations, their averaged model sizes are much less than the sample size, which motivates us to apply EBIC based methods to conduct further trimming to achieve a much lower FDR_n without sacrificing the PDR_n significantly. Specifically, for a reduced set s,

Method I: define

$$\mathscr{R}_1(s) = \{j : \text{EBIC}_\gamma(s - \{j\}) > \text{EBIC}_\gamma(s), j \in s\}\,; \tag{9.2.1}$$

Method II: define

$$\mathscr{D}^1(s) = \left\{ j : \mathrm{EBIC}_\gamma(s - \{j\}) = \min_{j \in s} \mathrm{EBIC}_\gamma(s - \{j\}) \right\},$$

$$\cdots\cdots$$

$$\mathscr{D}^k(s) = \left\{ j : \mathrm{EBIC}_\gamma(s - \bigcup_{j=1}^{k-1} \mathscr{D}^j(s) - \{j\}) = \min_{j \in s} \mathrm{EBIC}_\gamma(s - \bigcup_{j=1}^{k-1} \mathscr{D}^j(s) - \{j\}) \right\}.$$

$$\mathscr{R}_2(s) = \left\{ s - \bigcup_{j=1}^{t} \mathscr{D}^j(s) : \mathrm{EBIC}_\gamma \left(s - \bigcup_{j=1}^{t} \mathscr{D}^j(s) \right) = \min_{1 \le t \le |s|} \mathrm{EBIC}_\gamma \left(s - \bigcup_{j=1}^{t} \mathscr{D}^j(s) \right) \right\}. \tag{9.2.2}$$

$\mathscr{R}_1(s), \mathscr{R}_2(s)$ are the final sets of selected features. These two trimming methods are closely related. Note that event $\{s_{0n} = \mathscr{R}_1(s)\}$ is equivalent to

$$\left\{ \min_{j \in s_{0n}} \mathrm{EBIC}_\gamma(s - \{j\}) > \mathrm{EBIC}_\gamma(s) \right\} \bigcap \left\{ \max_{j \in s_{0n}^c} \mathrm{EBIC}_\gamma(s - \{j\}) \le \mathrm{EBIC}_\gamma(s) \right\}$$
$$\subseteq \left\{ \min_{j \in s_{0n}} \mathrm{EBIC}_\gamma(s - \{j\}) > \max_{j \in s_{0n}^c} \mathrm{EBIC}_\gamma(s - \{j\}) \right\} \subseteq \left\{ \mathscr{D}^1(s) \subseteq s_{0n}^c \right\}.$$

That is, if Method I is selection consistent for any $s_{0n} \subseteq s_0 \subseteq s$, then Method II is selection consistent if $\mathrm{EBIC}_\gamma(s_{0n}) = \min_{s_{0n} \subseteq s} \mathrm{EBIC}_\gamma(s)$. However, Method I takes much less time in computation. Our preliminary simulation results shows that Method I works as well as expected. More works are required to make this result persuasive.

It is persuasive through our theoretical and numerical studies that sequential LASSO is stable in identifying the relevant features with complex data features

in LMs. Another possible avenue of future work is to incorporate similar ideas in feature selection in GLMs and survival models, sparse graphical models, multiresponse linear regression models, and so on.

Bibliography

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

[2] Anderson, T.W.(2003). An Introduction to Multivariate Statistical Analysis. *New York: Wiley 3rd ed.* 86.

[3] Anderson, P.K., and Gill, R.D. (1982). Cox's Regression Model for counting processes: a large sample study. *Ann. Statist.*,**10**, 1100-1120.

[4] Annest, A., Bumgarner, R.E., Raftery, A.E., and Yeung, K.Y. (2009) Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data. *Bioinformatics*, 10-72.

[5] Bai, Z.D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, **9**, 611-677.

[6] Bai, Z.D., and Silverstein, J. W. (2006). Spectral Analysis of Large Dimensional Random Matrices. *Science Press,* Beijing.

[7] Barabasi, A.L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev.Genet.*, **12**, 56-68.

[8] Barron, A.R., Cohen, A., Dahmen, W., and Devore, R.A.(2008). Approximation and learning by greedy algorithms.*Ann.Statist.*, **36**,64-94.

[9] Bazaraa, M.S., Sherali, H.D., and Shetty, C.M. (2006). Nonlinear Programming: Theory and Algorithms. *Wiley-Interscience.* 165-218, 257-314.

[10] Benjamini Y., and Hochberg Y. (1995). Controlling the false discovery rate — A practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B*, **57**, 289-300.

[11] Berger, J. O. (2006). Some recent developments in Bayesian model selection. Presentation at *International Chinese Statistical Association 2006 Applied Statistics Symposium*, University of Connecticut, USA.

[12] Berger, J. O., and Pericchi, L. R. (2001). Objective Bayesian method for model selection: Introduction and comparison. In *Model Selection*, P. Lahiri (ed.), Institute of Mathematical Statistics Lecture Notes Monograph Series volume **38.**

[13] Bickel, P.J., Brown, J.B., Huang, H.Y., and Li, Q.H.(2009). An overview of recent developments in genomics and associated statistical methods. *Philos. Trans. R. Soc. London, Ser. A.* , **367**, 4313-4337.

[14] Bickel, P.J., and Li, B.(2006). Regularization in Statistics. *Test*, **2**,271-344.

[15] Bogdan, M., Ghosh, J. K., and Doerge, R.W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, **167**, 989-99.

[16] Bogdan,M., Ghosh, J.K., and Zak-Szatkowska, M.(2008). Selecting ExplanatoryVariables with the Modified Version of the Bayesian Information Criterion*Qual. Reliab. Eng. Int.*, **24**, 627-641.

[17] Bozdogan,H.(1987). Model selection and AIC: the general theory and its analytical extensions. *Psychometrika* , **52**, 345-370.

[18] Bradic. J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.*, **39**, 3092-3120.

[19] Breheny, P., and Huang, J. (2011). Coordinate descent algorithms for non-convex penalized regression, with applications to biological feature selection. *Ann.Appl.Statist.*, **5**, 232-253.

[20] Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.*, **87**, 738-754.

[21] Breiman, L. (1996). Heursitics of instability and stabilization in model selection.*Ann. Statist.*, **24**, 2350-2383.

[22] Broman, K. W., and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Statist. Soc. Ser. B*, **64**, 641-656.

[23] Brown, P.J., Fearn, T., and Vannucci, M. (1999).The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika*, **86**,635-648.

[24] Brown, P.J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction.*J. R. Statist. Soc. Ser. B*, **60**, 627-641.

[25] Brown, P.J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors.*J. R. Statist. Soc. Ser. B* , **64**, 519-536.

[26] Bühlmann, P. ,Kalisch, M., and Maathuis, M.(2010). Variable selection in high-dimensional models: partially faithful distributions and the PC-simple algorithms. *Biometrika* , **97**, 261-278.

[27] Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization.*Electron.J.Statist.*, **2**, 1153-1194.

[28] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, **1**, 169-194.

[29] Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* , **92**, 303-316.

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

[31] Candes, E. , and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313-2351.

[32] Casella, G., Gilron, F.J., Martinez, M.L., and Moreno,E.(2009). Consistency of Bayesian procedures for variable selection. *Ann.Statist.*, **37**,1207-1228.

[33] Chen, J. H., and Chen, Z. H. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* , **95**, 759-771.

[34] Chen, Z.H., and Chen, J.H. (2009). Tournament screening cum EBIC for feature selection with high dimensional feature spaces. *Sci.China, Ser.A. Math.*, **52**, 1327-1341.

[35] Chen, J.H., and Chen, Z.H. (2011). Extended BIC for small-n-large-p sparse GLM. Accepted by Statistical Sinica.

[36] Chen, Z.H., Chen, J.H., and Liu, J. (2006). A tournament approach to the detection of multiple associations in genome-wide studies with pedigree data. Working Paper 2006-09, www.stats.uwaterloo.ca. Department of Statistics and Actuarial Sciences, University of Waterloo.

[37] Chiang, A. P., Beck,J.S., Yen, H.-J. , Tayeh, M.K.,Scheetz, T.E., Swiderski, R. E.,Nishimura, D. Y.,Braun, T. A.,Kim, K-Y.,Huang, J. ,Elbedour, K.,Carmi,R., Slusarski, D. C. ,Casavant, T.L.,Stone, E. M., and Sheffield, V. C. (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a BardetCBiedl syndrome gene (BBS11). *PNAS.* , **103**, 6287-6292.

[38] Cho, H., and Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *J.R.Statist.Soc.Ser.B*, **74**, 1-30.

[39] Chong, I., and Jun, C. (2005). Performance of some variable selection methods when multicollinearity is present.*Chemom. Intell. Lab.Syst.*, **78**,103-112.

[40] Choi, N.H., Li, W., and Zhu, J. (2010). Variable Selection With the Strong Heredity Constraint and Its Oracle Property. *J.Am.Statist.Assoc.*, **105**, 354-364.

[41] Clyde, M. A., Berger, J. O., Bullard, F., Ford, E. B., Jefferys, W. H., Luo, R., Paulo, R., and Loredo, T. (2006). Current challenges in Bayesian model choice. *Astronomical Society of the Pacific Conference Series, Statistical Challenges in modern astronomy* **371**, 224.

[42] Conneely, K.A., and Boehnke, M.(2007). So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests. *Am. J. Hum. Genet.*, **81**,1158-1168.

[43] Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184-194.

[44] Cox, D. R. (1972). Regression models and life tables (with Discussion).*J. R. Statist. Soc. Ser. B* , **74**, 187-220.

[45] Cox, D. R. (1975). Partial likelihood. *Biometrika* , **62**, 269-276.

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

[47] Csörgö, M., and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. New York: John Wiley and Sons.

[48] Czado, C., and Munk, A. (2000). Noncanonical links in generalized linear models-when is the effort justified? *J.Stat.Plan.Infer.*, **87**, 317-345.

[49] Donoho, D.L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Presented at *AMS conference "mathematical challenges of the 21th century",* Aug 8.

[50] Donoho, D.L., Elad, M., and Temlyakov, V.N.(2006). Stable recovery of sparse overcomplete representations in the presence of noise.*IEEE. Trans. Inf. Theory.*, **52**, 6-18.

[51] Du,P., Ma, S., and Liang, H.(2010). Penalized variable selection procedure for cox models with semiparametric relative risk. *Ann. Statist.*, **38**, 2092-2117.

[52] Dvoretsky, A. (1972). Asymptotic normality for sums of dependent random variables. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability.* 513-535.

[53] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.(2004). Least angle regression (with discussion). *Ann.Statist.*, **32**, 407-499.

[54] Efroymson, M.A. (1966). Stepwise Regression-a backward and forward look. In *Eastern Regional Meetings of the Inst. of Math.Statistist.* Florham Park, New Jersey.

[55] Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010). SIS: Sure Independence Screening. R package version 0.6. http://cran.r-project.org/web/packages/SIS/index.html.

[56] Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screeing in sparse ultra-high dimensional additive models. *J. Am. Statist. Assoc.*, **116**, 544-557.

[57] Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model.*IMS Collections. Brorrowing Strength: Theory Powering Applications–A Festchrift for Lawrence D.Brown.***6**, 70-86.

[58] Fan, J., and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.*, **96**, 1348-1360.

[59] Fan, J., and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann.Statist.*, **30**, 74-99.

[60] Fan,J. , Li, G., and Li, R.(2005).An overview on Variable Selection for Survival Analysis. *Contemporary Multivariate Analysis and Experimental Design. The world scientific publisher.*

[61] Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Stat. Soc. Ser. B.*, **70**, 849-911.

[62] Fan, J., and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space.*Statistica Sinica* , **20**, 1-44.

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

[64] Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *J. Mach. Learn. Res.*, **10**, 2013-2038.

[65] Fan, J., and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality.*Ann. Statist.*, **38**, 3567-3604.

[66] Faraggi, D., and Simon, R. (1997). Large sample Bayesian inference on the parameters of the proportional hazard models. *Stat. Med.*, **16**, 2573-2585.

[67] Faraggi, D., and Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* , **54**, 1475-1485.

[68] Fill, J.A. (1983). Convergence rates related to the strong law of large numbers. *Ann. Probab.*, **11**, 123-142.

[69] Fleming, T.R., and Harrington, D.P.(1991).Counting Processes and Survival Analysis.*New York: Wiley.*

[70] Foygel, R., and Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. *Adv.Neural.Inf.Process.Syst.*, **23**, 2020-2028.

[71] Frank, I.E., and Friedman, J.H. (1993).A statistical view of some chemometrics regression tools.*Technometrics* , **35**, 109-135.

[72] Fraley, C., and Hesterberg, T.(2009). Least Angle Regression and LASSO for Large Datasets.*Stat.Anal.Data Min.*, **1**,251-259.

[73] Freedman, D.A. (1975). On tail probabilities for martingales. *Ann. probab.*, **3**, 100-118.

[74] Friedman, J, Hastie, T., and Tibshirani,R. (2010). glmnet: Lasso and elastic-net regularized generalized linear models R package version 1.7.1. http://cran.r-project.org/web/packages/glmnet/index.html.

[75] Friedman, J., Hastie, T., and Tibshirani, R.(2010). Regularization Paths for Generalized Linear Models via Coordinate Descent.*J.Stat.Softw.*, **33**, 1-22.

[76] George, E.I., and Mcculloch, R.E. (1993). Variable selection via Gibbs sampling. *J.Am.Statist.Assoc.*, **88**, 881-889.

[77] George, E.I., and Mcculloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339-373.

[78] Geyer, C.J. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist.*, **22**, 1993-2010.

[79] Giudici, P., Mezzetti, M., and Muliere, P. (2003). Mixtures of products of Dirichlet processes for variable selection in survival analysis. *J. Stat. Plan. Infer.*, **111**, 101-115.

[80] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., et al (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**, 531-537.

[81] Gui, J., and Li, H. Z. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001-3008.

[82] Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157-1182.

[83] Guyon, I., Elisseeff, A., and Aliferis,C. (2007). Causal feature selection. *Computational methods of feature selection*, **Series: Data mining and knowledge discovery.** 63-85.

[84] Haeusler, E. (1988). On the rate of convergence in the central limit theorm for martingales with discrite and continous time. *Ann. Probab.*, **16**, 275-299.

[85] Hastie, T. , Taylor, J. , Tibshirani, R. , and Walther, G. (2007). Forward stagewise regression and the monotone lasso. *Electron.J.Statist.*, **1**,1-29.

[86] Hesterberg, T. , Choi, N.H., Meier, L., and Fraley, C. (2008). Least Angles and l_1 penalized regression: a review. *Statist. Surv.* , **2**, 61-93.

[87] Hocking, R.R.(1976). The analysis and selection of variables in linear regression. *Biometrics* , **32**, 1-49.

[88] Hoerl, A.E., and Kennard, R.W.(1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* , **12**, 55-67.

[89] Huang, J. (1996). Efficient Estimation for the proportional hazards model with interval censoring. *Ann. Statist.*, **24**, 540-568.

[90] Huang, J. (1999). Efficient Estimation of the partly linear additive Cox model. *Ann.Statist.*, **27**, 1536-1563.

[91] Huang, J., Breheny, P. , Ma, S.G. , and Zhang, C.H. (2010). The Mnet method for variable selection.

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

[93] Huo, X., and Ni, X. (2007). When do stepwise algorithms meet subset selection criteria? *Ann. Statist.*, **35**, 870-887.

[94] Hwang, W.Y. , Zhang, H.H. , and Ghosal, S. (2009). FIRST: Combining forward iterative selection and shrinkage in high dimensional sparse linear regression. *Stat.Interface.*, **2**, 341-348.

[95] Ibrahim, J. G., and Chen, M. H. (2000). Power prior distributions for regression models. *Stat. Sci.*, **15**, 46-60.

[96] Ibrahim, J. G., Chen, M. H., and MacEachern, S. N. (1999). Bayesian variable selection for proportional hazards models. *Can. J. Stat.*, **27**, 701-717.

[97] Ing, C-K. , and Lai, T.L (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* , **21**, 1473-1513.

[98] James, G.M., Radchenko, P., and Lv, J. (2009).DASSO: connections between the Dantzig selector and lasso.*J. R. Statist. Soc. Ser.B.*, **71**,127-142.

[99] Jean-Baptiste Hiriart-Urruty, Claude Lemaréchal.(2001). Fundamentals of convex analysis. *New York: Springer.*

[100] Jiang, W. (2007). Bayesian Variable Selection for high dimensional generalized linear models: convergence rate of the fitted densities. *Ann. Statist.*, **35**, 1487-1511.

[101] Jia, J., and Yu, B. (2010). On model selection consistency of the Elastic Net when $P >> n$. *Statistica Sinica* , **20**, 595-611.

[102] John, G.H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and subset selection problem. *William, W.Cohen and Haym Hirsh, eds. Machine Learning: proceedings of the eleventh international conference,* 121-129.

[103] Johnson, B.A. (2009). On lasso for censored data. *Electron.J.Statist.*, **3**, 485-506.

[104] Karagrigorioua, A., Koukouvinosb, C., and Mylona, K.(2010). On the advantages of the non-concave penalized likelihood model selection method with minimum prediction errors in large-scale medical studies.*J. Appl. Stat.*, **1**, 13-24.

[105] Kass, R.E., and Wasserman, L. (1995). A reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *J.Am.Statist.Assoc.*, **90**, 928-934.

[106] Khalili, A. , Chen, J.H., and Lin, S.L.(2011). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics* , **12**, 156-172.

[107] Kim, Y., Choi, H., and Oh, H.(2008). Smoothly Clipped Absolute Deviation on High Dimensions.*J. Am. Statist. Assoc.* , **103**, 1665-1673.

[108] Klein, J. P., and Moeschberger, M. L. (2003). Survival Analysis: techniques for censored and truncated data. *New York: Springer.*

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

[110] Kraemer, N., and Schaefer, J. (2010). parcor: Regularized estimation of partial correlation matrices. R package version 0.2-2. http://cran.r-project.org/web/packages/parcor/index.html.

[111] Lee, K.E., Sha, N.J., Dougherty, E.R., Vannucci, M., and Mallick, B.K.(2003). Gene selection: a Bayesian variable selection approach. *Bioinformatices*, **19**, 90-97.

[112] Leng, C., Lin, Y., and Wahba, G. (2006). A note on Lasso and related procedures on model selection. *Statistica Sinica* , **16**, 1273-1284.

[113] Li, K.C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.*, **15**, 958-975.

[114] Li, R., and Liang, H.(2008). Variable selection in semiparametric regression modelling. *Ann. Statist.*, **36**, 261-286.

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

[116] Liu, Q., and Watbleda, F. (2009). Exponential inequalities for martingales and asymptotic properties of the free energy of directed polymers in a random environment. *Stoch. Proc. Appl.*, **119**, 3101-3132.

[117] Lukacs, P.M., Burnham, K.P., and Anderson, D.R.(2010). Model selection bias and Freedmans paradox. *Ann. Inst. Stat. Math.*, **62**, 117-125.

[118] Lv, J., and Fan, Y.Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares.*Ann.Statist.*, **37**, 3498-3528.

[119] Ma,Y., and Li, R.(2010). Variable selection in measurement error models. *Bernoulli* , **16**,274-300.

[120] Mallows, C.L. (1973). "Some Comments on C_p". *Technometrics* ,**15**, 661-675.

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

[122] Meinshausen, N., and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436-1462.

[123] Meinshausen, N. , and Yu, B. (2009). Lasso-type of sparse representations for high-dimensional data. *Ann.Statist.*, **37**,246-270.

[124] Mundry, R., and Nunn, C.L.(2009). Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution. *Am. Nat.*, **173**, 119-123.

[125] Mykland, P.A. (1993). Asymptotic expansions for martingales. *Ann. Probab.*, **21**, 800-818.

[126] Osborne, M.R., Presnell, B., and Turlach, B.A.(1998). Knot selection for regression splines via the Lasso. In Dimension Reduction, Computational Complexity, and Information, ed. S. Weisberg, Vol. 30. of Computing Science and Statistics, Fairfax Station, VA: Interface Foundation of North America, Inc., pp. 44-49.

[127] Osborne, M.R, Presnell, B., and Turlach, B.A. (2000).On the Lasso and its dual. *J. Comput. Graph. Stat.*, **9**, 319-337.

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

[129] Peïa Victor H., Lai, T., and Shao, Q.M. (2009). Self-normalized processes. *Limit theory and statistical applications. Chapter9. Martingale Inequalities and related tools.*

[130] Perkins, S., Lacker, K., and Theiler, J. (2003). Grafting: fast, incremental feature selection by gradient decrescent in function space. *J. Mach. Learn. Res.*, **3**, 1333-1356.

[131] Pinelis, I. (2006). Binomial upper bounds on generalized moments and tail probabilities of (super)martingales with differences bounded from above.*IMS Lecture Notes Monograph Series, High Dimensional Probability.*, **51**, 3352.

[132] Presnell, B., Turlach, B.A., and Osborne, M.R.(2000). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, **20**, 389-403.

[133] Radchenko, P., and James, G.M. (2011). Improved variable selection with forward-lasso adaptive shrinkage. *Ann. Appl. Stat.*, **5**, 427-448.

[134] Rao, C. R., and Wu, Y. H. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* , **76**, 369-374.

[135] Robbins, H. (1955). A Remark on Stirlings Formula.*Am. Math. Mon.*, **62**, 26-29.

[136] Rolando Rebolledo. (1980). Central Limit Theorem for Local Martingales. *Probab. Theory. Rel.*, **51**, 269-286.

[137] Rosenwald, A. et al .(2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New Engl J Med .*, **346**, 1937-1947.

[138] Rosset, S.(2004).Tracking curved regularized optimization solution paths. *Adv. Neural. Inf. Process. Syst.*

[139] Rosset, S. , and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann.Statist.*, **35**, 1012-1030.

[140] Sara A. Van de Geer. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann.Statist.*, **23**, 1779-1801.

[141] Sara A. Van de Geer. (2008). High-dimensional generalized linear models and the lasso.*Ann. Statist.*, **36,** 614-645.

[142] Sauerbrei, W., and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the cox regression model. *Stat. Med.*, **11**, 2093-2109.

[143] Scheetz, T.E. , Kim, K.-Y.A. , Swiderski, R.E., Philip1,A.R., Braun, T.A. , Knudtson, K.L. , Dorrance, A.M., DiBona, G.F., Huang, J., Casavant. T,L., Sheffield, V.C., and Stone, E.M. (2006). Regularization of gene expression in the mammalian eye and its relevance to eye disease. *Proc.Natl.Acad.Sci. U.S.A.*, **103**, 14429-14434.

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

[145] Scott, J.G., and Berger,J.O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable selection problem. *Ann.Statist.*, **38**, 2587-2619.

[146] Serfling, R. J.(1980). Approximation theorems of mathematical statistics, *New York: Wiley.* 33.

[147] Sha, N. J , Tadesse, M. G., and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, **22**, 2262-2268.

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

[149] Shao, J., and Chow, S-C. (2007). Variable screening in predicting clinical outcome with high-dimensional microarrays. *J. Multivariate. Anal.*, **98**, 1529-1538.

[150] Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Statist. Assoc.*, **62**, 626-633.

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

[152] Sinha, D., Chen, M. H., and Ghosh, S. K. (1999). Bayesian analysis and model selection for interval censored survival data. *Biometrics* , **55**, 585-590.

[153] Spirtes, P., Glymour, C., and Scheines, R. (1993). Causation, Prediction, and Search. *New York: Springer-Verlag.*

[154] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc. Ser. B.*, **39**, 111-147.

[155] Sun, T.N. , and Zhang, C.H. (2011). Scaled sparse linear regression.

[156] Tibshirani , R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. Ser. B.*, **58**, 267-288.

[157] Tibshirani , R. (1997). The Lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 385-395.

[158] Tropp, J. A. (2004). Greed is good: Algorithmic Results for Sparse Approximation. *IEEE Trans. Inf. Theory.*, **50**,1-21.

[159] Tropp, J. A., and Gilbert, A.C.(2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory.*, **53**, 4655-4666.

[160] Troyanskaya et al (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520-525.

[161] Tsiatis, A.A. (1981). A large sample study of Cox's regression model. *Ann.Statist.*, **9**, 93-108.

[162] van der Vaart, A.W. , and Wellner, J.A. (1996). Weak convergence and empirical processes. *New York: Springer.*

[163] Variyath, A.M. , Chen, J.H., and Abraham,B. (2009).Empirical likelihood based variable selection. *J. Stat. Plan. Inf.*, **140**, 971-981.

[164] Volinsky, C.T., and Raftery, A.E. (2000). Bayesian Information Criterion for Censored Survival Models. *Biometrics* , **56**, 256-262.

[165] Wainwright, M.J.(2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming(Lasso). *IEEE Trans. Inf. Theory.*, **55**, 2183-2202.

[166] Wang, H.S. (2009). Forward Regression for Ultra-High Dimensional Variable Screening. *J.Am. Statist. Assoc.*, **105**, 1512-1524.

[167] Wasserman, L. (2000). Bayesian Model Selection and Model Averaging.*J. Math. Psychol.*, **44**, 92-107.

[168] Wasserman, L., and Roeder, K. (2009). High-dimensional variable selection. *Ann.Statist.*, **37**, 2178-2201.

[169] Wedderburn, R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63**, 27-32.

[170] Weisberg, S. (1980). Applied Linear Regression. Wiley, New York.

[171] Wellner, J.A. (1992).Empirical Processes in Action: A Review. *Int. Stat. Rev.*, **60**, 247-269.

[172] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Statist.*, **9**, 60-62.

[173] Woodroofe, M. (1982).On model selection and the arc sine laws. *Ann.Statist.*, **10**, 1182-1194.

[174] Wu, Y., and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* , **19**, 801-817.

[175] Xie, H.L., and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Ann. Statist.*, **37**, 673-696.

[176] Yao, Y. C. (1988). Estimating the number of change-points via Schwartz' criterion. *Statist. Probab. Lett.*, **6**, 181-189.

[177] Yuan, M., and Lin, Y.(2007). On the non-negative garrote estimator. *J. R. Stat. Soc. Ser. B.*, **69**, 143-161.

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

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

[180] Zhang, Y.Y, Li, R.Z. , and Tsai, C.L. (2010). Regularization Parameter Selections via Generalized Information Criterion. *J.Am.Statist.Assoc.*, **105**, 312 - 323.

[181] Zhang, H. H. , and Lu, W. (2007). Adaptive lasso for coxs proportional hazards model.*Biometrika* , **94**, 691-703.

[182] Zhao, J. Y. , and Chen, Z. H. (2012). A Two-Stage Penalized Logistic Regression Approach to Case-Control Genome-Wide Association Studies .*J. Prob. Statist.*, **2012**, 1-15.

[183] Zhao, P., and Yu, B. (2006). On model selection consistency of LASSO. *J. Mach. Learn. Res.*, **7**, 2541-2567.

[184] Zhao, P., and Yu, B. (2007). Stagewise LASSO. (old title: Boosted Lasso). *J. Mach. Learn. Res.*, **8**, 2701-2726.

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

[186] Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* , **95**, 241-247.

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

[188] Zou, H. , and Li, R.Z. (2008). One-step sparse estimates in nonconcave penalized likelihood models.*Ann.Statist.*, **36**, 1509-1533.

[189] Zou, H., and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters.*Ann.Statist.*, **37**, 1733-1751.

Appendix A: The Verification of C6 in Section 4.1

In this Appendix, we will check Condition C6 in Section 4.1 by looking at the common GLMs with non-canonical link functions when σ_i^2 are assumed to be away from 0 and finite. For the ease of reference, condition C6 is given below:

C6 The quantities $|x_{ij}|, |h^{'}(X_i^\tau \boldsymbol{\beta}_0)|, |h^{''}(X_i^\tau \boldsymbol{\beta}_0)|, i = 1, \ldots, n; j = 1, \ldots, p_n$ are bounded from above, and $\sigma_i^2, i = 1, \ldots, n$ are bounded both from above and below away from zero. Furthermore,

$$\max_{1\le j\le p_n; 1\le i\le n} \frac{x_{ij}^2 [h^{'}(X_i^\tau \boldsymbol{\beta}_0)]^2}{\sum_{i=1}^n \sigma_i^2 x_{ij}^2 [h^{'}(X_i^\tau \boldsymbol{\beta}_0)]^2} = o(n^{-1/3}),$$
$$\max_{1\le i\le n} \frac{[h^{''}(X_i^\tau \boldsymbol{\beta}_0)]^2}{\sum_{i=1}^n \sigma_i^2 [h^{''}(X_i^\tau \boldsymbol{\beta}_0)]^2} = o(n^{-1/3}).$$

The common GLMs were considered in [169]. In particular, we consider the following exponential families and their corresponding link functions:

(1) Poisson Distribution: $\eta = \ln(\mu)$ or $\eta = \mu^\gamma$ where $0 < \gamma < 1$;

(2) Binomial Distribution: $\eta = \mu,\ \eta = \arcsin(\mu),\ \eta = \ln(\frac{\mu}{1-\mu}),\ \eta = \ln\left(-\ln(1-\mu)\right),\ \eta = \Phi^{-1}(\mu)$;

(3) Gamma Distribution ($G(1, \mu)$): $\eta = \ln \mu$ or $\eta = \mu^\gamma$ where $-1 \le \gamma < 0$.

Since for poisson distribution, binomial distribution and gamma distribution, $(\theta, b(\theta)) = \left(\ln(\mu), e^\theta\right),\ \left(\ln(\frac{\mu}{1-\mu}), \ln(1+e^\theta)\right),\ \left(-\frac{1}{\mu}, -\ln(-\theta)\right)$ respectively, the above can be rewritten as follows:

(1) Poisson Distribution: $\theta = \eta$ or $\theta = \frac{1}{\gamma}\ln\eta$ where $0 < \gamma < 1$;

(2) Binomial Distribution: $\theta = \ln\frac{\eta}{1-\eta}$, $\theta = \ln\frac{\sin(\eta)}{1-\sin(\eta)}$, $\theta = \eta$, $\theta = \ln\left(\exp(e^{\eta}) - 1\right)$, $\theta = \ln\left(\frac{\Phi(\eta)}{1-\Phi(\eta)}\right)$.

(3) Gamma Distribution: $\theta = -e^{-\eta}$ or $\theta = -\eta^{-\frac{1}{\gamma}}$.

Poisson Distribution

$\eta = \mu^{\gamma}$ where $0 < \gamma < 1$: assume $\mu_i \in [a, b]$ for all i. Under this situation,

$$h^{'}(\eta) = \frac{1}{\gamma\eta},\ h^{''}(\eta) = -\frac{1}{\gamma\eta^2},\ \sigma^2 = \eta^{\frac{1}{\gamma}}.$$

Hence under the assumption, $\forall 1 \le i \le n$,

$$|h^{'}(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0)| \in [\frac{1}{\gamma b^{\gamma}}, \frac{1}{\gamma a^{\gamma}}], \sigma_i^2 \in [a, b], |h^{''}(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0)| \in [\frac{1}{\gamma b^{2\gamma}}, \frac{1}{\gamma a^{2\gamma}}],$$

$$\frac{x_{i,j}^2\left(h^{'}(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0)\right)^2}{\sum_{i=1}^{n}\sigma_i^2 x_{i,j}^2\left(h^{'}(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0)\right)^2} = \frac{b^{2\gamma}}{a^{2\gamma+1}}O\left(\frac{x_{i,j}^2}{\sum_{i=1}^{n}x_{i,j}^2}\right),$$

$$\frac{\left(h^{''}(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0)\right)^2}{\sum_{i=1}^{n}\sigma_i^2\left(h^{''}(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0)\right)^2} = \frac{b^{4\gamma}}{a^{4\gamma+1}}O(n^{-1}).$$

When $0 < a < b < +\infty$, $C6$ is true if $\max_{1\le j\le p_n}\max_{1\le i\le n}\left\{\frac{x_{i,j}^2}{\sum_{i=1}^{n}x_{i,j}^2}\right\} = o(n^{-1/3})$.

Binomial Distribution

For binomial distribution, $\sigma_i^2 = \mu_i(1-\mu_i) = \dfrac{e^{\theta_i}}{(1+e^{\theta_i})^2}$. Here we assume

$$\min_{1\le i\le n}\left(\mu_i \wedge (1-\mu_i)\right) \ge c \text{ where } 0 < c \le 1/2. \tag{A.1.1}$$

This implies, $c^2 \le \min_{1\le i\le n}\sigma_i^2 \le \max_{1\le i\le n}\sigma_i^2 \le 1/4$. Therefore,

$$\frac{x_{i,j}^2\left(h^{'}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2}{\sum_{i=1}^n \sigma_i^2 x_{i,j}^2\left(h^{'}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2} = O\left(\frac{x_{i,j}^2\left(h^{'}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2}{\sum_{i=1}^n x_{i,j}^2\left(h^{'}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2}\right)$$

$$\frac{\left(h^{''}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2}{\sum_{i=1}^n \sigma_i^2\left(h^{''}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2} = O\left(\frac{\left(h^{''}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2}{\sum_{i=1}^n \left(h^{''}(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0)\right)^2}\right).$$

(1) $\mu = \eta, 0 < \eta < 1$:

$$h^{'}(\eta) = \frac{1}{\eta(1-\eta)},\ h^{''}(\eta) = \frac{2\eta-1}{\eta^2(1-\eta)^2},\ \sigma^2 = \eta(1-\eta).$$

Under assumption (A.1.1),

$$4 \le h^{'}\left(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0\right) \le \frac{1}{c^2};\ \left|h^{''}\left(\boldsymbol{x}_i^\tau\boldsymbol{\beta}_0\right)\right| \le \frac{1-2c}{c^4}$$

for all $1 \le i \le n$. C6 holds when

$$\max_{1\le j\le p_n}\max_{1\le i\le n}\left\{\frac{x_{i,j}^2}{\sum_{i=1}^n x_{i,j}^2}\right\} = o(n^{-1/3}).$$

(2) $\eta = \arcsin\mu$:

$$h^{'}(\eta) = \frac{\cos\eta}{\sin\eta(1-\sin\eta)},\ h^{''}(\eta) = \frac{\sin\eta}{1-\sin\eta} - \frac{\cos^2\eta}{\sin^2\eta},\ \sigma^2 = \sin\eta(1-\sin\eta).$$

Under assumption (A.1.1),

$$4\sqrt{2c-c^2} \leq \left|h^{'}\left(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0\right)\right| \leq \frac{\sqrt{1-c^2}}{c^2};$$
$$\frac{3c-c^2-1}{(1-c)^2c} \leq \left|h^{''}\left(\boldsymbol{x}_i^{\tau}\boldsymbol{\beta}_0\right)\right| \leq \frac{1-c^2-c}{c^2(1-c)}$$

for all $1 \leq i \leq n$. C6 holds when

$$\max_{1\leq j\leq p_n}\max_{1\leq i\leq n}\left\{\frac{x_{i,j}^2}{\sum_{i=1}^{n} x_{i,j}^2}\right\} = o(n^{-1/3}).$$

(3) $\eta = g(\mu) = \ln\{-\ln(1-\mu)\}$ or $\eta = g(\mu) = \ln\{-\ln(\mu)\}$.

For the first link function, complementary log-log link, we have

$$\theta = \ln(\frac{\mu}{1-\mu}) = h(\eta) = \ln\{\exp(e^{\eta})-1\},\ \sigma^2 = \frac{\exp(e^{\eta})-1}{\exp(2e^{\eta})}. \qquad (A.1.2)$$

Therefore, the first and second order derivatives of $h(\cdot)$ are

$$h^{'}(\eta) = \frac{e^{\eta+e^{\eta}}}{e^{e^{\eta}}-1};\ h^{''}(\eta) = \frac{e^{\eta+e^{\eta}}[e^{e^{\eta}}-e^{\eta}-1]}{\{e^{e^{\eta}}-1\}^2}. \qquad (A.1.3)$$

It is easy to see that $e^{\eta} \leq h^{'}(\eta) \leq e^{e^{\eta}}$. Now let us look at $h^{''}(\eta)$. It is straightforward that $|h^{''}(\eta)| \leq |h^{'}(\eta)| \leq e^{e^{\eta}}$. Consider the function $f(x) = \frac{e^x(e^x - x - 1)}{(e^x - 1)^2}$ on $(0, +\infty)$. Since

$$\lim_{x\to 0} f(x) = \lim_{x\to 1} \frac{x^2/2}{x^2} = \frac{1}{2}; \quad \lim_{x\to+\infty} f(x) = \lim_{x\to+\infty} \frac{1 - \frac{x}{e^x} - \frac{1}{e^x}}{(1 - \frac{1}{e^x})^2} = 1, \quad \text{(A.1.4)}$$

there exists a positive constant C_1, C_2 independent of x such that $C_1 \leq f(x) \leq C_2$. That is, $C_1 e^{\eta} \leq h^{''}(\eta) \leq C_2 e^{\eta}$. When $\sigma_i^2 \in [a, b]$ for some $0 < a \leq b \leq 1/4$, for $1 \leq i \leq n$, we have

$$\frac{1 + \sqrt{1 - 4b}}{2b} \leq \exp(e^{\eta_i}) \leq \frac{1 + \sqrt{1 - 4a}}{2a} \text{ or } \frac{1 - \sqrt{1 - 4a}}{2a} \leq \exp(e^{\eta_i}) \leq \frac{1 - \sqrt{1 - 4b}}{2b}.$$

That is, $|h^{'}(\eta_i)|$ and $|h^{''}(\eta_i)|$ are both bounded away from 0 and finite. C6 holds when

$$\max_{1\leq j\leq p_n} \max_{1\leq i\leq n} \left\{ \frac{x_{i,j}^2}{\sum_{i=1}^{n} x_{i,j}^2} \right\} = o(n^{-1/3}).$$

The same argument applies to the second link function by changing η to $-\eta$.

(4) $\eta = \Phi^{-1}(\mu)$:

$$h^{'}(\eta) = \frac{f(\eta)}{\Phi(\eta)(1-\Phi(\eta))},\ h^{''}(\eta) = \frac{f^{'}(\eta)}{\Phi(\eta)(1-\Phi(\eta))} + f^2(\eta)[\frac{1}{(1-\Phi(\eta))^2} - \frac{1}{\Phi^2(\eta)}]$$

$$\sigma^2 = \Phi(\eta)(1-\Phi(\eta)).$$

Under assumption (A.1.1), $\Phi^{-1}(c) \leq |\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0| \leq \Phi^{-1}(1-c)$. Note that

$$1 - \Phi(t) \leq \frac{f(t)}{t}, \forall t > 0,$$

therefore, we have

$$4c\Phi^{-1}(c) \leq 4f\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right) \leq \left|h^{'}\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right)\right| \leq \frac{1}{c^2} f\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right) \leq \frac{1}{\sqrt{2\pi}c^2};$$

$$4f^{'}\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right) \leq \left|\frac{f^{'}(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0)}{\Phi(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0)(1-\Phi(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0))}\right| \leq \frac{1}{c^2} f^{'}\left(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0\right) \leq \frac{\Phi^{-1}(1-c)}{\sqrt{2\pi}c^2};$$

$$\left|f^2(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0)[\frac{1}{(1-\Phi(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0))^2} - \frac{1}{\Phi^2(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0)}]\right| \leq \frac{|2c-1|}{c^2(1-c)^2} f^2(\boldsymbol{x}_i^\tau \boldsymbol{\beta}_0) \leq \frac{|2c-1|}{2\pi c^2(1-c)^2}$$

for all $1 \leq i \leq n$. C6 holds when

$$\max_{1\leq j\leq p_n} \max_{1\leq i\leq n} \left\{ \frac{x_{i,j}^2}{\sum_{i=1}^{n} x_{i,j}^2} \right\} = o(n^{-1/3}).$$

Gamma Distribution

(1) $\eta = \ln(\mu) : h^{'}(\eta) = e^{-\eta},\ h^{''}(\eta) = -e^{-\eta},\ \sigma^2 = e^{2\eta}$. When σ_i^2 is away from 0

and finite, $|h^{'}|, |h^{''}|$ are bounded. C6 holds when

$$\max_{1\leq j\leq p_n}\max_{1\leq i\leq n}\left\{\frac{x_{i,j}^2}{\sum_{i=1}^{n}x_{i,j}^2}\right\}=o(n^{-1/3}).$$

(2) $\eta=\mu^{\gamma}$ where $-1\leq\gamma<0$. Let $\tilde{\gamma}=-\frac{1}{\gamma}$, then $0<\tilde{\gamma}\leq 1$. Then

$$h^{'}(\eta)=-\tilde{\gamma}\eta^{\tilde{\gamma}-1},\ h^{''}(\eta)=\tilde{\gamma}(1-\tilde{\gamma})\eta^{\tilde{\gamma}-2},\ \sigma^2=\eta^{2\tilde{\gamma}}.$$

When σ_i^2 is away from 0 and finite, $|h^{'}|, |h^{''}|$ are bounded. C6 holds when

$$\max_{1\leq j\leq p_n}\max_{1\leq i\leq n}\left\{\frac{x_{i,j}^2}{\sum_{i=1}^{n}x_{i,j}^2}\right\}=o(n^{-1/3}).$$

Appendix B: Proofs of Equations (7.3.5) and (7.3.7)

In this section, we provide proofs of Equations (7.3.5) and (7.3.7). Let $s_{\star k}$ be the set of selected features at the kth step of sequential LASSO, $\triangle^{\mu}(s_{\star k})=\|[\boldsymbol{I}-H_0(s_{\star k})]\boldsymbol{y}\|_2^2$ and $\boldsymbol{\beta}$ be the true coefficient vector in the linear model. The contents of these inequalities are as follows:

Equation (7.3.5): $\dfrac{\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1})}{\ln n}\to+\infty.$

Equation (7.3.7): $\forall 0 \leq k < \tilde{p}_0 - 1, \triangle^{\mu}(s_{\star k}) - \triangle^{\mu}(s_{\star k+1}) \geq \frac{\triangle^{\mu}(s_{\star k+1})^2}{(1+\lambda_0)^2 n \|\boldsymbol{\beta}_{(s^-_{\star k+1})}\|_1^2}$.

For $k \geq 0$, let $\mathscr{A}_k$ be the index set of the variables with bounded size (or the only variable) being added at the $(k+1)$th step of sequential LASSO, we assume that there exists constants L, λ_0 such that

$$\max_{0\leq k<\tilde{p}_0} |\mathscr{A}_k| \leq L, \ \max_{0\leq k<\tilde{p}_0} \frac{\lambda_{\max}\left(X^{\tau}_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right)}{\lambda_{\min}\left(X^{\tau}_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right)} \leq \lambda_0.$$

Lemma A.2.1. *Use the notation ∂ in Proposition 6.2.3, then there exists a vector $\widehat{\boldsymbol{\beta}}^{(k+1)}_{\mathscr{A}_k}$ with componentwise nonzero elements such that*

$$|\partial(\widehat{\boldsymbol{\beta}}^{(k+1)}_j)| \leq 1, \ \ \forall j \in s^c_{\star k+1}, \qquad \textit{where} \tag{A.2.1}$$

$$\begin{aligned}
\partial\left(\widehat{\boldsymbol{\beta}}^{(k+1)}_j\right) =& 2(X^{\tau}_j\left[\boldsymbol{I} - H_0(s_{\star k+1})\right]\boldsymbol{y}_n)(\lambda^{\star}_{i+1})^{-1} \\
&+ X^{\tau}_j\left[\boldsymbol{I} - H_0(s_{\star k}\right] X_{\mathscr{A}_k}\left\{X^{\tau}_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k}\right] X_{\mathscr{A}_k}\right\}^{-1}\partial(\widehat{\boldsymbol{\beta}}^{(k+1)}_{\mathscr{A}_k}) \\
=& 2(X^{\tau}_j\left[\boldsymbol{I} - H_0(s_{\star k+1})\right]\boldsymbol{\epsilon}_n)(\lambda^{\star}_{i+1})^{-1} \\
&+ 2(X^{\tau}_j\left[\boldsymbol{I} - H_0(s_{\star k+1})\right]\boldsymbol{\mu}_n)(\lambda^{\star}_{i+1})^{-1} \\
&+ X^{\tau}_j\left[\boldsymbol{I} - H_0(s_{\star k}\right] X_{\mathscr{A}_k}\left\{X^{\tau}_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k}\right] X_{\mathscr{A}_k}\right\}^{-1}\partial(\widehat{\boldsymbol{\beta}}^{(k+1)}_{\mathscr{A}_k}) \\
\overset{def}{=}& N_1 + N_2 + N_3; \\
\partial\left(\widehat{\boldsymbol{\beta}}^{(k+1)}_{\mathscr{A}_k}\right) =& \partial\left(X^{\tau}_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right]\boldsymbol{y}_n\right).
\end{aligned} \tag{A.2.2}$$

Proof of Lemma A.2.1. Denote by $Q(\boldsymbol{\beta})$ the objective function at the $(k+1)$th step in sequential LASSO. From Proposition 6.2.2 and the Karush-Kuhn-Tucker(KKT)

conditions in Proposition 6.2.3, we know that $Q(\boldsymbol{\beta})$ can reach its minimum at $\widehat{\boldsymbol{\beta}}$ if and only if

$$2X^{\tau}(s^{c}_{\star k})\left[\boldsymbol{I}-H_0(s_{\star k})\right](\boldsymbol{y}_n - X(s^{c}_{\star k})\widehat{\boldsymbol{\beta}}) = \lambda\partial(\widehat{\boldsymbol{\beta}}). \tag{A.2.3}$$

Note that $\partial(\widehat{\boldsymbol{\beta}})^{\tau}\widehat{\boldsymbol{\beta}} = \|\widehat{\boldsymbol{\beta}}\|_1$, the tuning parameter λ can be solved as follows immediately,

$$\lambda = \frac{2(\boldsymbol{y}_n - X(s^{c}_{\star k})\widehat{\boldsymbol{\beta}})^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])X(s^{c}_{\star k})\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_1}. \tag{A.2.4}$$

Plugging λ back into equation (A.2.3), we have

$$\frac{(\boldsymbol{y} - X(s^{c}_{\star k})\widehat{\boldsymbol{\beta}})^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])X(s^{c}_{\star k})\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_1}\partial(\widehat{\boldsymbol{\beta}}) = X(s^{c}_{\star k})^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right](\boldsymbol{y}-X(s^{c}_{\star k})\widehat{\boldsymbol{\beta}}). \tag{A.2.5}$$

Specially, if $X_{\mathscr{A}_k}$ is the set of variables (or the unique variable) being added in the $(k+1)$th step, that is, equation (A.2.3) holds for some $\widehat{\boldsymbol{\beta}} = (0,\cdots,0,\widehat{\boldsymbol{\beta}}^{(k+1)}_{\mathscr{A}_k},0,\cdots,0)^{\tau}$ and $|\partial(\widehat{\boldsymbol{\beta}}^{(k+1)}_{j})| \leq 1, \quad \forall j \in s^{c}_{\star k+1}$. Let $\lambda^{\star}_{k+1}$ be the corresponding λ in (A.2.4) for this $\widehat{\boldsymbol{\beta}}$. The following equation can be derived directly from (A.2.5) after plugging this particular $\widehat{\boldsymbol{\beta}}$ in (A.2.4),

$$\widehat{\boldsymbol{\beta}}^{(k+1)}_{\mathscr{A}_k} = \left\{X^{\tau}_{\mathscr{A}_k}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}\left\{X^{\tau}_{\mathscr{A}_k}\left[\boldsymbol{I}-H_0(s_{\star k})\right]\boldsymbol{y} - \frac{\lambda^{\star}_{k+1}}{2}\partial(\widehat{\boldsymbol{\beta}}^{(k+1)}_{\mathscr{A}_k})\right\}. \tag{A.2.6}$$

Similarly, by plugging into (A.2.5), we have the following equation for $j \in s^c_{\star k+1}$,

$$\begin{aligned}\partial(\widehat{\boldsymbol{\beta}}_j^{(k+1)}) &= \frac{2}{\lambda^\star_{k+1}} X_j^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] (\boldsymbol{y} - X_{\mathscr{A}_k} \widehat{\boldsymbol{\beta}}_{\mathscr{A}_k}^{(k+1)}) \\ = &- 2\frac{X_j^\tau Q_i^{\star k+1}}{\lambda^\star_{k+1}} \boldsymbol{y} + X_j^\tau \left[\boldsymbol{I} - H_0(s_{\star k}\right] X_{\mathscr{A}_k} \left\{X_{\mathscr{A}_k}^\tau \left[\boldsymbol{I} - H_0(s_{\star k}\right] X_{\mathscr{A}_k}\right\}^{-1} \partial(\widehat{\boldsymbol{\beta}}_{\mathscr{A}_k}^{(k+1)}).\end{aligned} \tag{A.2.7}$$

where

$$Q_i^{\star k+1} = \left[\boldsymbol{I} - H_0(s_{\star k})\right] \left(X_{\mathscr{A}_k} \left\{X_{\mathscr{A}_k}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right]) X_{\mathscr{A}_k}\right\}^{-1} X_{\mathscr{A}_k}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] - \boldsymbol{I}\right).$$

By applying the identity (4.17) in [97], it follows that $Q_i^{\star k+1} = -\left[\boldsymbol{I} - H_0(s_{\star k+1})\right]$. Then plug back into equation (A.2.7), we can obtain the desired result. □

Proof of (7.3.5). We prove this conclusion by contradiction. Assume there exists some $k \leq \tilde{p}_0$ such that $\triangle^\mu(s_{\star k}) - \triangle^\mu(s_{\star k+1}) = O(\ln n)$. Let $s^- = s_{0n} \cap s^c$ for any set s. When $k < \tilde{p}_0 - 1, s^c_{\star k+1} \neq \emptyset$, define

$$X_{k+1\star} = \underset{m \in s^-_{\star k+1}}{\operatorname{argmax}} \left|\gamma_n\left(m, s_{\star k+1}, \boldsymbol{\beta}\right)\right|,$$

where the definition of $\gamma_n(\cdot)$ is given in Theorem 7.1.1. We remark here that if this $X_{k+1\star}$ is not unique, we let it be one of the representatives. This will not influence our results. Also, $X_{k+1\star}$ could be a candidate of features being selected

at the $(k+2)$th step. Denote by N_1, N_2, N_3 the three terms of $\partial(\widehat{\beta}^{(k+1)}_{k+1\star})$ in equation (A.2.2), we have the following convergence rate from (ii) in the proof of Theorem 7.1.1,

$$\frac{\sqrt{n}|\gamma_n\left(k+1\star, s_{\star k+1}, \boldsymbol{\beta}\right)|}{\ln p_n} \geq C_n \to +\infty. \tag{A.2.8}$$

To see more clearly about the convergence rate of the terms in N_1, N_2, N_3, we need the following relationships, which can be easily verified through calculation:

$$\begin{aligned}
&\boldsymbol{y}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k} \left\{X^\tau_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right\}^{-1} X^\tau_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right])\boldsymbol{y} \\
=&[\triangle^\mu(s_{\star k}) - \triangle^\mu(s_{\star k+1})] - 2[\triangle^{\mu,\epsilon}(s_{\star k}) - \triangle^{\mu,\epsilon}(s_{\star k+1})] + [\triangle^\epsilon(s_{\star k}) - \triangle^\epsilon(s_{\star k+1})] \\
=&[(\triangle^\mu(s_{\star k}) - \triangle^\mu(s_{\star k+1}))] + O_p(\|(H_0(s_{\star k}) - H_0(s_{\star k+1}))\boldsymbol{\mu}_n\|_2) + O_p(\ln n) \\
\geq&\|\boldsymbol{y}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\|_2^2 \lambda^{-1}_{\max}\left(X^\tau_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right); \\
&\boldsymbol{\mu}_n^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k} \left\{X^\tau_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right\}^{-1} X^\tau_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right])\boldsymbol{\mu}_n \\
=&\triangle^\mu\left(s_{\star k}\right) - \triangle^\mu(s_{\star k+1}) \\
\geq&\|\boldsymbol{\mu}_n^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\|_2^2 \lambda^{-1}_{\max}\left(X^\tau_{\mathscr{A}_k}\left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right); \\
&\|(H_0(s_{\star k}) - H_0(s_{\star k+1}))\boldsymbol{\mu}_n\|_2^2 = \triangle^\mu(s_{\star k}) - \triangle^\mu(s_{\star k+1}).
\end{aligned} \tag{A.2.9}$$

Multiplying $\partial(\widehat{\beta}^{(k+1)}_{\mathscr{A}_k})$ by both sides of equation(A.2.6), from the third equation in

Lemma A.2.1, we know that the positivity on the left hand side implies

$$\begin{aligned}\lambda_{k+1}^{\star} <& \frac{2\boldsymbol{y}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}}{\partial(\widehat{\beta}_{\mathscr{A}_k}^{(k+1)})^{\tau}\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}\partial(\widehat{\beta}_{\mathscr{A}_k}^{(k+1)})\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_1} \\ =&2\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_1.\end{aligned} \tag{A.2.10}$$

If $(\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1}))=O(\ln n)$, (A.2.9) implies

$$\begin{aligned}\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|^2 =&\lambda_{\max}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)O_p(\ln n)=O_p(n\ln n), \\ \|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{\mu}_n\|^2 =&\lambda_{\max}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)O_p(\ln n)=O_p(n\ln n).\end{aligned} \tag{A.2.11}$$

Because of the bounded size of $X_{\mathscr{A}_k}$, we know that $\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_2$ and $\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_1$ have exactly the same order, this also applies to $X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{\mu}_n$ and $X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}$. Hence, the following inequalities can be obtained from (A.2.8),(A.2.9),(A.2.10),(A.2.11) for $\partial\left(\widehat{\boldsymbol{\beta}}_{k+1\star}^{(k+1)}\right)$:

$$\begin{aligned}|N_2| >& \frac{n\gamma_n\left(k+1\star,s_{\star k+1},\boldsymbol{\beta}\right)}{\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_1}=+\infty; \\ \frac{|N_1|}{|N_2|} =&O_p(\frac{\|\left[\boldsymbol{I}-H_0(s_{\star k+1})\right])X_{k+1\star}\|_2}{n\gamma_n\left(k+1\star,s_{\star k+1},\boldsymbol{\beta}\right)})\leq O_p(\frac{1}{C_n\ln p_n})=o_p(1).\end{aligned} \tag{A.2.12}$$

Note that

$$
\begin{aligned}
&X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{k+1\star}\\
=&\left[\|\boldsymbol{I}-H_0(s_{\star k})\right]X_{k+1\star}\|_2^2-\left[\|\boldsymbol{I}-H_0(s_{\star k+1})\right]X_{k+1\star}\|_2^2\\
\geq&\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_2^2\lambda_{\max}^{-1}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)\\
\geq&n^{-1}|\mathscr{A}_k|^{-1}\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_2^2
\end{aligned}
\tag{A.2.13}
$$

Hence, $\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_2=O(n)$. Moreover, for a matrix A with order $m_1\times m_2$, denote $\|A\|_r=\sup_{x\neq 0}\frac{\|Ax\|_r}{\|x\|_r}(r=1,2)$, then $\frac{1}{\sqrt{m_1}}\|A\|_1\leq\|A\|_2\leq\sqrt{m_2}\|A\|_1$, this leads to

$$
\begin{aligned}
|N_3|\leq&\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}\|_1\\
\leq&\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_1\|\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}\|_1\\
\leq&|\mathscr{A}_k|\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_2\|\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}\|_2\\
\leq&|\mathscr{A}_k|\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_2\|\lambda_{\min}^{-1}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)\\
\leq&Cn^{-1}\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_2=O(1).
\end{aligned}
\tag{A.2.14}
$$

Therefore, we have $\partial(\widehat{\boldsymbol{\beta}}_{k+1\star}^{(k+1)})=|N_1+N_2+N_3|=+\infty$, which is a contradiction according to Lemma A.2.1. That is, (7.3.5) is proved. □

Proof of (7.3.7). Use the same notations as in the Proof of (7.3.5), again, we focus

on the three terms in $\partial\left(\widehat{\boldsymbol{\beta}}_{k+1\star}^{(k+1)}\right)$: (i) and (ii) in the Proof of Theorem 7.1.1 lead to

$$|N_1 + N_2| = |N_2|(1 + o_p(1)).$$

From equations (A.2.9) and (7.3.5), we have

$$\begin{aligned}
&\boldsymbol{y}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k} \left\{X_{\mathscr{A}_k}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right\}^{-1} X_{\mathscr{A}_k}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right])\boldsymbol{y} \\
=&\left[\triangle^\mu(s_{\star k}) - \triangle^\mu(s_{\mathscr{A}_k})\right](1 + o_p(1)) \\
\geq&\|\boldsymbol{y}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\|_2^2 \lambda_{\max}^{-1}\left(X_{\mathscr{A}_k}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\right) \\
\geq&n^{-1}|\mathscr{A}_k|^{-1}\|\boldsymbol{y}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\|_2^2.
\end{aligned} \tag{A.2.15}$$

Therefore, by expanding μ as summation from parts $s_{\star k+1}$ and $s_{\star k+1}^c$, after a direct comparison, we have, with probability tending to 1,

$$\begin{aligned}
|N_2| \geq&\frac{n|\gamma_n\left(k+1\star, s_{\star k+1}, \boldsymbol{\beta}\right)|}{\|\boldsymbol{y}^\tau \left[\boldsymbol{I} - H_0(s_{\star k})\right] X_{\mathscr{A}_k}\|_1} \geq \frac{n|\gamma_n\left(k+1\star, s_{\star k+1}, \boldsymbol{\beta}\right)|}{|\mathscr{A}_k|\sqrt{n(\triangle^\mu(s_{\star k}) - \triangle^\mu(s_{\star k+1}))}} \\
>&\frac{\triangle(s_{\star k+1})}{|\mathscr{A}_k|\|\boldsymbol{\beta}(s_{\star k+1}^-)\|_1\sqrt{n(\triangle^\mu(s_{\star k}) - \triangle^\mu(s_{\star k+1}))}} = r_{k,n}.
\end{aligned} \tag{A.2.16}$$

From the middle two terms of (A.2.13) and (A.2.15),

$$\frac{\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_1\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}\|_1}{|\mathscr{A}_k|}$$

$$\leq\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_2\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\left\{X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right\}^{-1}\|_2$$

$$\leq\|X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right])\boldsymbol{y}\|_2\|X_{k+1\star}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}^{\tau}\|_2\lambda_{\min}^{-1}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)$$

$$\leq\sqrt{n(\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1}))}\frac{\lambda_{\max}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)}{\lambda_{\min}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)}\stackrel{def}{=}\lambda_{0,k}\sqrt{n(\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1}))},$$

where

$$C\geq\lambda_{0,k}=\frac{\lambda_{\max}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)}{\lambda_{\min}\left(X_{\mathscr{A}_k}^{\tau}\left[\boldsymbol{I}-H_0(s_{\star k})\right]X_{\mathscr{A}_k}\right)}\geq 1.$$

Consequently, we have

$$\frac{|N_2|}{|N_3|}\geq\frac{n|\gamma_n\left(k+1\star,s_{\star k+1},\boldsymbol{\beta}\right)|}{\lambda_{0,k}\sqrt{n(\triangle^{\mu}(s_{\star k})-\triangle^{\mu}(s_{\star k+1}))}}\geq\frac{r_{k,n}}{\lambda_{0,k}}\tag{A.2.17}$$

by combining with the first inequality in (A.2.14) together with the first inequality in (A.2.16). If $r_{k,n}\to+\infty$, we have $\partial(\widehat{\boldsymbol{\beta}}_{k+1\star}^{(k+1)})\to+\infty$; if $\lambda_{0,k}<r_{k,n}<+\infty$, we have $|N_3|\leq\dfrac{\lambda_{0,k}}{r_{k,n}}|N_2|\leq|N_2|$, therefore,

$$1\geq|\partial(\widehat{\boldsymbol{\beta}}_{k+1\star}^{(k+1)})|\geq|N_2|-|N_3|\geq(1-\frac{\lambda_{0,k}}{r_{k,n}}+o_p(1))|N_2|\geq r_{k,n}-\lambda_{0,k},$$

and $0<r_{k,n}<\lambda_{0,k}+1$. Plugging into the definition of $r_{k,n}$ in (A.2.16), then we have our desired result (7.3.7). □