SIMULTANEOUS MODEL SELECTION AND ESTIMATION OF GENERALIZED LINEAR MODELS WITH HIGH DIMENSIONAL PREDICTORS

By

Alex Pijyan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2022

ABSTRACT

SIMULTANEOUS MODEL SELECTION AND ESTIMATION OF GENERALIZED LINEAR MODELS WITH HIGH DIMENSIONAL PREDICTORS

By

Alex Pijyan

In the past couple of decades the progressive use of technology made the enormous amount of data in different formats available and easily accessible. The size and volume of available data sets have grown rapidly and the technological capacity of the world to store information has almost doubled every 40 months since the 1980s [1]. As of 2020, every day 2.5 quintillion of data are generated. Based on an International Data Group (IDG) report, the global data volume was predicted to grow exponentially and by 2025, IDG predicts there will be 163 zettabytes of data [2].

This enormous amount of data is often characterized by its high dimensionality. Quite often, well-known statistical methods fail to manage such data due to their limitations (e.g., in high-dimensional settings they often encounter various issues such as no unique solution for the model parameters, inflated standard errors, overfitted models, multicollinearity). This resulted in resurging interest in the algorithms that are capable of handling massive quantities of data, extracting and analysing information from it, and uncovering key insights that subsequently will lead to decision making.

Techniques used by these algorithms are tend to speed up and improve the quality of predictive analysis, thus, they found their application in various fields. For instance, medicine becomes more and more individualized nowadays and drugs or treatments can be designed to target small groups, rather than big populations, based on characteristics such as medical history, genetic makeup etc. This kind of treatment is referred to as precision medicine.

In the era of precision medicine, constructing interpretable and accurate predictive models, based on patients' demographic characteristics, clinical conditions, and molecular biomarkers, has been crucial for disease prevention, early diagnosis and targeted therapy [3]. The models, for example, can be used to predict patients' susceptibility to disease [4], identify high risk groups [5], and guide behavioral changes [6]. Therefore, predictive models play a central role in decision making.

Several well-known approaches can be used to solve the problem mentioned above. Penalized regression approaches, such as least absolute shrinkage and selection operator (LASSO), have been widely used to construct predictive models and explain the impacts of the selected predictors, but the estimates are typically biased. Moreover, when data are ultrahigh-dimensional, penalized regression is usable only after applying variable screening methods to downsize variables.

In this dissertation, we would like to propose a procedure for fitting generalized linear models with ultrahigh-dimensional predictors. Our procedure can provide a final model, control both false negatives and false positives, and yield consistent estimates, which are useful to gauge the actual effect size of risk factors. In addition, under a sparsity assumption of the true model, the proposed approach can discover all of the relevant predictors within a finite number of steps.

The thesis work is organized as follows. Chapter 1 highlights an importance of predictive models and names several examples where these models can be implemented. The main focus of Chapter 2 is to describe all well-known and already existing in the theory methods that attempted to solve the aforementioned problems, along with their shortcomings and disadvantages. Chapter 3 proposes STEPWISE algorithm and introduces the model setup and its detailed description, followed by its theoretical properties and proof of theorems and lemmas used throughout the thesis. Additional lemmas used to construct the theory of the STEPWISE method are also stated.

Later it presents results obtained from various numerical studies such as simulations and real data analysis. Simulation studies comprise seven examples and are aimed to compare STEPWISE algorithm to other competing methods, and provide numerical evidence of its superiority. Real data analysis involves studies of gene regulation in the mammalian eye, esophageal squamous cell carcinoma, and neurobehavioral impairment from total sleep deprivation, and demonstrates the utility of the proposed method in real life scenarios.

Chapter 4 proposes a multi-stage hybrid machine learning ensemble method that is aimed to enhance STEPWISE's performance. It also introduces a web application that employs the method. Finally, Chapter 5 completes the thesis with final conclusion and discussions. Appendices include some tables and figures used throughout the thesis.

ACKNOWLEDGEMENTS

First and foremost, it is my genuine pleasure to express my sincere gratitude and the deepest appreciation to my academic advisor at Michigan State University, Dr. Hyokyoung Hong. This dissertation would not have been possible without her dedication, advice, continuous encouragement, priceless support, and persistent help. I have been extremely fortunate to have Dr. Hong as my advisor as her insistence on perfection and careful supervision have given me a lot of confidence and inspiration to achieve my goals.

I would like to thank our collaborators, Dr. Qi Zheng and Dr. Yi Li, for their significant contribution to the work presented in this thesis.

My sincere thanks also go to my committee members, Dr. Taps Maiti, Dr. Lyudmila Sakhanenko, and Dr. Chenxi Li, for generously giving their time to offer me valuable and insightful comments toward improving my work.

Finally, I would like to express special thanks to my friends and family for overwhelming love and support, for being by my side when I was going through hard times, and for celebrating each accomplishment. Through the struggles and challenges of writing this dissertation, you have been a constant source of joy and motivation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

1.1 Predictive Models

In biomedical research and clinical studies predictive model are utilized for several purposes such as risk management and prognosis. Consequently, the reliability of clinical data is directly related to the quality of predictive analysis. Nowadays, electronic health records became more available and contain rich information, which enables researchers to develop and deploy highly efficient clinical predictive methods. These methods have potential to be key components in making decisions related to patient treatments, drug development, and so on.

Over the last decade, the technological advances and explosion of information profounded the understanding of the molecular basis of tumor progression and identified numerous tumor biomarkers [7]. A certain type of biomarkers, which posses predictive power, are capable of assessing the benefit from clinical interventions and has a significant impact on clinical research. For instance, a cancer screening biomarker is a prognostic biomarker that can be used to predict the development of symptomatic cancer even in asymptomatic persons.

In practice, such screening biomarkers can be used as a cancer prediction model [8]. The main purpose of building these models is discovering new cancer screening biomarkers and assessing their effect on the disease. Insights and information obtained from these models can potentially lead to early detection of disease in patients, early intervention and prevention from its further development. Further, these predictive models can be feasibly used as cancer screening tests for patients, including ones with no symptoms [9].

Technological advances have also made possible detailed genetic characterization of biological specimens. High-throughput genomic technologies, including gene expression microarray, microRNA (micro Ribonucleic acid) array, RNA-seq, ChIP-seq (chromatin immunoprecipitation sequencing), and whole genome sequencing, have become powerful tools and dramatically changed

the landscape of biological research. For instance, a gene expression profile can be extracted for a specimen by simultaneously evaluating expression levels of thousands of genes on that single specimen using complementary DNA (cDNA) microarray technology [10].

Single nucleotide polymorphisms (SNPs) are one form of natural sequence variation common to all genomes [11]. These SNPs are highly abundant, and are estimated to occur at 1 out of every 1,000 bases in the human genome [12, 13]. SNPs are particularly useful as DNA markers for mapping susceptibility genes for complex diseases and population genetics since they demonstrate the high density and mutational stability [14, 15].

SNPs in the coding regions of genes that alter the function or structure of the encoded proteins can be a necessary and sufficient cause of most of the known recessively or dominantly inherited monogenic disorders [16], and are analyzed for diagnostic purposes. Moreover, SNPs can be analysed to assess the risk of an individual for a particular disease. For instance, the identification of SNPs made possible to screen somatic (non-tumor) DNA for mutations that alter treatment response or predispose to cancer [7].

In addition, a large number of profiles based on the abundance of micro-RNAs (miRNAs) have been used to predict prognosis or treatment response in cancer [7]. For example, Genome-wide association (GWA) have identified cancer-causing mutations in breast [17] and colon [18] tumors, somatic genetic screens can also identify predictors of radiation sensitivity [19] and the pharmacodynamics of anticancer drugs [20]. Moreover, sets of genes identified through mRNA profiling have been used to classify tumors into oncogenic subtypes of breast cancer [21, 22]; many individual miRNAs have been associated with patient survival and drug treatment response in a number of different cancers [23, 24].

Clearly, the enormous interest in genomic data is determined by the hope of finding candidate biomarkers and using them to identify genes that predispose individuals to common diseases. Although genome data analysis has already made a significant impact on biological and biomedical research, it is still accompanied by certain challenges that yet have to be overcome. Specifically, complex genomic data introduce substantial challenges for statistical data analysis as its

high-dimensionality makes the classical statistical model framework no longer implementable. As opposed to low-dimensional data when the number of observations is greater than the number of explanatory features (also known as predictors), high- and ultrahigh-dimensional settings are comprised of data in which number of predictors is greater than or is in the exponential order of the sample size, respectively.

Most of the traditional statistical methods are developed around the concept of low-dimensional data and are not aimed to accommodate high- or ultrahigh-dimensional data. Thus, high-dimensionality has significantly challenged traditional statistical theory. Applying these methods to high-dimensional data leads to unstable, unreliable, biased, and inconsistent results which demolishes the main purpose of predictive model development. The problems that arise while analyzing such data are typically referred to as the 'curse of dimensionality', a term introduced by mathematician Richard Bellman. Some aspects of it are discussed further.

First, classical statistical models applied to high-dimensional data have no unique solution for their parameters. In fact, these models will have infinitely many solutions. This is mainly induced by ill-defined, uninvertible, and singular matrices involved in the computation of parameter estimates, making the estimation process ill-posed. These models are also know as unidentifiable models. Consequently, effect size estimation in predictive models will become meaningless.

Second, as the number of predictors increases and surpasses the number of observations in the model, variances of the parameter estimates will become large (even infinitely large in some cases), resulting in inflated standard errors. In other words, a wide range of values of parameter estimates will be consistent with data, making the confidence intervals uncommonly wide. Hence, validating a significance of the predictors included in the predictive model will be nearly impossible.

Third, employment of classical statistical models in high-dimensional settings can provide inconsistent estimates as a small corruption of data can result in very different estimated parameters. Furthermore, these models tend to capture the artificial trends of measurement noise, also know as, overfitting. Overfitted models fit training data too closely and normally capture trends in data that are applicable to this particular data set only. This decays their ability to generalized results with

new unseen data and results in poor predictive capability.

Lastly, using classical statistical methods with high-dimensional data often introduces multicollinearity issues that violate the underlying assumption of independent predictors in the model. Multicollinearity implies the existence of highly correlated predictors among predictor features. These structures are commonly observed in genomic data. Multicollinearity can create inaccurate estimates of the model parameters; make insignificant predictors significant and vice versa, that is, imposing false positives and false negatives in the predictive model; and, finally, it can degrade the predictability of the model.

As it was shown, the traditional methods that perform well in low-dimensional settings run into severe problems in analyzing high- or ultrahigh-dimensional data. They cannot cope with the explosive growth of dimensionality of data. Therefore, in order to face the problem of high-dimensionality, we must reshape the classical statistical thinking. These problems create significant challenges, but, on the other hand, they create great opportunities for the development of new statistical methodologies.

It is worth mentioning that developing predictive models along with feature selection and estimation play crucial and fundamental role in knowledge discovery. As more amount of massive and complex data become available, there is no doubt that high-dimensional data analysis will be one of the most important and demanding research topics in our field.

In this thesis we propose a new method (introduced and described in Chapter 3) for model selection and estimation that will overcome aforementioned limitations in high-dimensional problems. The remaining sections of this chapter discuss a few problems from various research areas that will illustrate challenges of high-dimensional data and to which the proposed method could be applied.

1.1.1 Gene Regulation in the Mammalian Eye

Human genetics has sparked a revolution in medical research on the basis of the seemingly unthinkable notion that one can systematically discover the genes causing inherited diseases without any prior biological knowledge as to how they function [25]. Most characteristics of medical pertinence

do not follow simple Mendelian monogenic inheritance. Such complex traits include vulnerability to heart disease, hypertension, diabetes, cancer, and infection. The genetic dissection of complex traits is attracting many investigators with the promise of solving old problems and is generating a variety of analytical methods.

Recent advancement in microarray technology and bioinformatics made it possible to examine the expression of numerous genes in a large number of individuals and enabled researcher to identify genetic elements that cause the gene expression to vary among individuals [26, 27, 28, 29]. Discovering specific disease mechanics is a big challenge that biomedical researchers face nowadays. These mechanics might potentially underlie heritable disorders that reveal complex inheritance, for instance, Mendelian disorders [25, 30, 31]. In addition, these approaches can help identify genes related to development of Mendelian forms of complex diseases such as obesity [32, 33, 34], macular disease [35, 36, 37], hypertension [38], and glaucoma [39, 40].

Mutations that alter gene expression might play a significant role in complex disease. Transgenic animal studies revealed that gene dosage of mutant genes can have a keen effect on phenotype [41]. It was shown that the cause of disease can become an improper regulation of structurally normal genes and alterations in gene dosage [39]. For example, overexpression and haploinsufficiency of the FOXC1 gene can lead to developmental defects of the anterior chamber of the eye [39].

Scheetz *et al.* [41] used expression quantitative trait locus mapping in the laboratory rat to gain a broad perspective of gene regulation in the mammalian eye and to identify genetic variation relevant to human eye disease. They analyzed data obtained from Rat Genome Database by using analysis of variance (ANOVA) technique and identified significant genes based on their corresponding p-values. Certainly, Scheetz's results provide meaningful insights on how genetic variation can be associated with specific diseases, but they do not estimate the magnitude of effects these genes are having on the disease. In addition, they have not built a predictive model that will enable researchers to link genes and assess their contribution toward developing diseases, and have not evaluate its predictive power. We adopted their data and aimed to improve results achieved by Scheetz *et al.* [41]. Data contained 120 observation profiling 31042 probes of genes, but due to a

small variation in many of these probes, the number of probes was reduced to 5000.

A gene *TRIM32* that has been found to cause Bardet-Biedl syndrome [42] was treated a response variable, and the expression of 5000 genes as the predictors. Our predictive model has identified three probes of genes (*1376747_at*, *1381902_at*, *1382673_at*) that can be potentially linked to *TRIM32*. We achieved a high accuracy with the mean squared prediction error (MSPE) as low as 0.0012. Detailed description of the results can be found in Chapter 3.

1.1.2 An Esophageal Squamous Cell Carcinoma Study

Esophageal cancer is the 7th most common cancer among males and among both sexes combined in the world and ranks 6th in terms of mortality overall because of the poor survival rate it confers [43, 44]. Additionaly, incidence and mortality rates in males are 2- to 3-fold higher than the rates in females [43]. Compared with more developed geographic regions, overall incidence rates are 2-fold higher in less-developed countries, with the highest rates occurring in Asia [43].

Esophageal squamous cell carcinoma (ESCC) is the predominant histologic type with the highest incidence rate in populations within Southeastern and Central Asia [44]. There are two major histological types of esophageal carcinoma: esophageal squamous cell carcinoma (ESCC) and adenocarcinoma [45]. ESCC is the major type in China, where it accounts for more than 90% of cases of esophageal carcinoma; whereas adenocarcinoma is more common in the United States and in European countries [46]. ESCC is often diagnosed at a locally advanced stage and the outcomes for affected patients are poor [45].

With various treatment methods employed in clinical practice after extensive research, the diagnosis and treatment of ESCC have been greatly improved [47]. Esophagectomy, chemotherapy, and radiotherapy are currently the main treatments for ESCC [45]. However, the prognosis remains poor, with 5-year survival proportions of 21% and 14% (2005–2011) in the United States for whites and blacks, respectively, and 12% (2000–2007) in Europe [44], which is far below the estimated effectiveness of the therapy [47].

An accurate clinical staging and prognostic information is essential to direct appropriate treatment

strategies [45]. Accumulating evidence suggests that the prognosis is affected by several factors, including the delayed diagnosis, high recurrence, and metastasis rate [47, 48]. Thus, identifying the diagnostic and prognostic tumor markers and further elucidating their clinical implications are urgently needed. To develop new diagnostic methods and treatment strategies, investigators have focused on the potential of a particular class of microRNAs (miRNAs) to provide additional information about the characteristics and survival prospects of patients with ESCC.

miRNAs are small (22-24 nucleotides), noncoding RNA molecules that play important roles in regulating cell differentiation, proliferation, migration and apoptosis [44]. Altered miRNA expression in cancer tissue has been reported in most tumor types [49, 50]. There is increasing evidence that miRNA expression in cancer tissue is a useful prognostic marker [51, 52, 53]. In addition, the application of miRNA expression levels as a blood biomarker has been explored in various types of cancer, including gastric, hepatocellular, and non-small cell lung cancer [54, 55, 56]. However, whether miRNA levels in plasma are a useful biomarker for patients with ESCC remains largely unexplored [45].

Sudo *et al.* [57] explored ways of developing a detection model for ESCC based on large-scale miRNA profiling. For these purposes, they analyzed data submitted to the National Center for Biotechnology Gene Expression Omnibus (NCBI GEO) database, available under accession number GSE122497. To establish a diagnostic model, they developed a model based on the observations obtained from 566 patients (283 with ESCC and 283 healthy controls) profiling 2565 miRNAs by carrying out Fisher's linear discriminant analysis with a greedy algorithm.

Although their model has achieved high predictive accuracy, it has some drawbacks. Given the nature of the algorithm that has been employed, they developed a predictive model for the predetermined model size: they built models with model sizes ranging 2-8 and selected the one that achieved higher accuracy with fewer variables (model size = 6). The disadvantage of this method is that it might lead to false negatives and false positive in the final model. Moreover, the importance of the miRNAs included in the model will also be determined by the model size. This might lead to a wrong assessment of the effect sizes identified in the model.

We adopted this dataset and demonstrated the utility of our proposed method (introduced and described in Chapter 3) and its superiority over other methods. Our model achieved similar accuracy by recruiting fewer variables (3 miRNAs were selected: *miR* - 4783 - 3*p*, *miR* - 320*b*, *miR* - 1225 - 3*p*). It is worth mentioning that our model overcomes the issues associated with the model introduced by Sudo *et al.* as our model size was defined by scanning the entire feature space and selecting features based on their significance. Detailed description of our results and methodology is presented in Chapter 3.

1.1.3 Bladder Cancer Study

Bladder cancer is any of several types of cancer arising from the tissues of the urinary bladder and has high prevalence and recurrence rates [58, 59, 60]. According to American Society of Clinical Oncology, among men bladder cancer is the fourth most common cancer and men are 4 times more likely to be diagnosed with the disease. In addition, incidence in white men is twice more than that in black men.

The earlier bladder cancer is found, the better the chance for successful treatment and cure. Prognosis varies inversely with higher tumor stage and lymph node involvement [61]. Typically, the 5- and 10-year survival rates for patients with lymph node involvement are 31% and 23%, respectively [62]. Combination platinum-based chemotherapy is an potion for patients with metastatic disease, but the survival is only 15 months, with a 5-year survival rate of 15% [63]. Since there is not yet an accurate test to screen the general population for bladder cancer, most people are diagnosed with bladder cancer once they have developed symptoms. As a result, some people have more advanced (later stage) disease when the cancer is found.

In the year 2000, the total expenditure for lower tract urothelial cancers in the United States surpassed one billion dollars [64]. Bladder cancer affected about 1.6 million people globally in 2020 with 549,000 new cases and 200,000 deaths, and the late stage disease is associated with poor survival. The cost of bladder cancer per patient from diagnosis to death is the highest of all cancers [65].

Identifying the related biomarkers and predicting the disease at its early stage is crucial for better prognosis. Discovery of diagnostic, prognostic, and predictive biomarkers in bladder cancer made molecular markers an area of research. Potential biomarkers include the overexpression of mutated genes, whole genome-wide array signatures, and microRNAs. For instance, microarray gene expression profiling is studied in the blood of cancer patients in order to detect gene expression patterns representing the cancer itself or a host's reaction to the tumor [66, 67].

Recent studies have suggested that 70% of bladder cancer involve a specific mutation in genes [68], therefore it can be potentially used as a biomarker in early detections of the disease and preventing it from further development. Gene changes can also assist doctors in choosing the best treatment possible or be useful in finding bladder cancers that can potentially come back after treatment. Although recent progress made by scientists is significant, classification of bladder cancer patients using gene expression data with regular statistical tools can become complicated and sometimes be even impossible due to incapability of these methods to process data with a large scale, also known as, high-dimensional data.

Usuba *et al.* [69] attempted to develop a predictive model for an early detection in bladder cancer. They applied similar technique as described in Sudo *et al.* [57] via Fisher's linear discriminant analysis and recruited seven miRNAs in their final model. They achieved high accuracy in prediction, but their model suffered from the same issues mentioned in Sudo *et al.* [57]. Specifically, the method they employed might lead to false negatives and false positives in the final model, and won't be able to assess the effect sizes correctly due to pre-determined model size.

We aimed to improve given results and adopted data utilized in Usuba's model. These data were submitted to the NCBI GEO under accession number GSE113486. The predictive model was built based on observations obtained from 768 patients (310 patients with bladder cancer and 468 healthy controls) profiling 2565 miRNAs.

We demonstrated that our proposed multi-stage hybrid machine learning method (introduced and described in Chapter 4) has achieved high prediction accuracy with sensitivity, specificity, and area under the receiver operating curve (AUC) of 0.98, 0.98, and 0.99, respectively, and outperformed

Usuba's model. Detailed description of our results can be found in Chapter 4.

1.1.4 Neurobehavioral Impairment from Total Sleep Deprivation

Sleep plays a key role in health, performance, and cognition [70]. Sleep deprivation is commonplace in modern society and its effects on neurobehavioral function (e.g., vigilance and cognition) are well studied and documented. Sleep deprivation can induce giddiness, child-like behaviors, and silliness [71], as well as more widely recognized negative effects including dysphoria, increased irritability, and lowered frustration tolerance.

The increased irritability that often accompanies sleep deprivation hints that sleep-deprived individuals are highly reactive to emotional signals. These effects on mood can lead to negative consequences and impact functioning abilities [72]. For example, sleep duration is inversely associated with interpersonal difficulties and even violence has been observed in medical residents [73], and sleeping less than 8 hours is associated with increased risk for adolescent suicidal behavior [74].

Interestingly, alertness and vigilance also appear to be the cognitive capacities most consistently and dramatically impacted by insufficient sleep [75]. When the envelope of continuous wakefulness is pushed beyond about 16 hours, most individuals begin to show a substantial slowing of reaction time (RT) and worsening of performance accuracy on tests of psychomotor vigilance [76]. Moreover, neurobehavioral tests have revealed assorted forms of performance deficits from sleep loss, including impairment of learning and of responses to feedback in decision making [77]. The Psychomotor Vigilance Test (PVT) is one of the most commonly applied neurobehavioral assays of performance impairment due to sleep loss [75]. This test assays stimulus-response time, with failure to respond within 500 ms recorded as a lapse. Sleep deprivation is associated with increased variability in stimulus-response times, and more lapses, on the PVT [78].

Besides neurobehavioral testing, efforts have been made to identify molecular biomarkers such as differentially expressed genes or metabolites affected by sleep loss [79, 80]. A biomarker has been defined as "a characteristic that is objectively measured and evaluated as an indicator of normal

biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [81]. Many biomarkers such as differentially expressed genes can provide meaningful insights including identification of a process or response. Humans are known to differ in their sensitivity to sleep loss [82], and recent work has sought to identify biomarkers distinguishing individuals as susceptible or resistant to sleep deprivation [83]. Yet surprisingly little effort has been made to research molecular biomarkers with results from neurobehavioral assays.

Microarrays and bioinformatics analyses can be employed to explore candidate gene expression biomarkers associated with total sleep deprivation (TSD), and more specifically, the phenotype of neurobehavioral impairment from TSD. Uyhelji *et al.* [70] explored gene expression biomarker candidates for neurobehavioral impairment from total sleep deprivation. They employed Weighted Gene Co-expression Network Analysis (WGCNA) using data obtained from the NCBI GEO under accession number GSE98582. Data contain 555 samples profiling 8284 gene features. In the treatment effect analysis, they identified 212 genes that exhibited a significant difference between TSD and control group, and 91 of them passed human blood biomarker filter.

Although Uyhelji *et al.* have done a great job in identifying important gene biomarkers associated with TSD, effect sizes of these genes have not been estimated. Moreover, neither predictive power of the diagnostic model was assessed. Thus, we employed our proposed method (discussed in Chapter 3) to overcome the issues introduced in Uyhelji's model. We have built a model based on 389 observations profiling 8284 gene features. Our model recruited five genes (*PF4V1*, *USP32P1*, *EMR1*, *NBR2*, and *DUSP23*) and achieved high accuracy with sensitivity, specificity, and AUC of 0.99, 0.97, and 0.99, respectively.

CHAPTER 2

LITERATURE REVIEW

When the number of predictors is moderate, penalized regression approaches such as least absolute shrinkage and selection operator (LASSO) by Tibshirani [84] have been used to construct predictive models and explain the impacts of the selected predictors. LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. However, in ultrahigh-dimensional settings where a number of predictors p is in the exponential order of the sample size n, penalized methods may incur computational challenges [85], may not reach globally optimal solutions, and often generate biased estimates [86].

Sure independence screening (SIS) proposed by Fan and Lv [87] has emerged as a powerful tool for modeling ultrahigh dimensional data. This method is based on correlation learning, which filters out the features that have weak correlation with the response and reduces dimensionality from high to a moderate that is below the sample size. Specifically, such correlation learning ranks the importance of features according to their marginal correlation with the response variable and eliminates the ones with weak marginal correlations.

However, the method relies on a partial faithfulness assumption, which stipulates that jointly important variables must be marginally important, an assumption that may not be always realistic. To relieve this condition, some iterative procedures, such as ISIS [87], have been adopted to repeatedly screen variables based on the residuals from the previous iterations, but with heavy computation and unclear theoretical properties. Conditional screening approaches (see, e.g. [88]) have, to some extent, addressed the challenge. However, screening methods do not directly generate a final model, and post-screening regularization methods, such as LASSO, are recommended by Fan and Lv [87] to produce a final model.

Closely related to forward selection is least angle regression (LARS) by Efron *et al.* [89], a widely

used model selection algorithm for high-dimensional models. In the LARS method, a multivariate solution path is defined by using the geometrical theory of the linear regression model. The resulting method defines a continuous solution path for Generalized Linear Models (GLMs), with on the extreme of the path the maximum likelihood estimate of the coefficient vector and on the other side the intercept-only estimate. The LARS method is based on a recursive procedure selecting, at each step, the covariates having largest absolute correlation with the response feature [89].

It is worth mentioning that a simple modification of the LARS algorithm implements the LASSO method and calculates all possible LASSO estimates for a given problem. In addition, an approximation for the degrees of freedom of a LARS estimate is available, from which Mallows's C_p estimate of prediction error can be derived; this allows a principled choice among the range of possible LARS estimates. Though LARS achieves impressive results in its performance, some researches raised concerns in the following regards.

Ishwaran (see discussion section in Efron *et al.* [89]) suggests that the use of C_p coupled with LARS forward optimization procedure might raise some potential flags. Specifically, the use of C_p will encourage large models in LARS, especially in high-dimensional orthogonal problems, and will have a negative impact on variable selection performance. The claim was supported with the high-dimensional simulation examples. Moreover, Weisberg (see discussion section in Efron *et al.* [89]) believes that multicollinearity problem among independent features and presence of noise in the dependent variable will affect the performance of LARS in regards of variable selection, specifically, reducing chances of selecting significant variables in the model. Examples supporting the claim were provided.

In the GLM setting, Augugliaro *et al.* [90] and Pazira *et al.* [91] proposed differential geometrical LARS (dgLARS) based on a differential geometrical extension of LARS. The dgLARS estimator follows naturally from a differential geometric interpretation of a GLM, generalizing the LARS method.

The subsequent section discusses sequential model selection techniques known to the literature. Sequential model selection assumes including features into the final model sequentially with the

entry order determined by their relative importance based on certain criteria. Although the methods described in this section have achieved significant results and have been implemented in different scenarios, lack of certain properties make them less reliable and more vulnerable against challenges introduced by the size and volume of data available nowadays.

2.1 Sequential Model Selection

For generating a final predictive model in ultrahigh-dimensional settings, recent years have seen a surging interest of performing forward regression, an old technique for model selection that has been widely used for model building when the number of covariates is relatively low. But due to its complicated computations and unknown theoretical properties, forward regression technique is rarely used in high-dimensional settings.

Under some regularity conditions and with some proper stopping criteria, forward regression can achieve screening consistency and sequentially select variables according to metrics such as Akaike information criterion (AIC), Bayesian information criterion (BIC), or R^2. Below are listed methods that try to overcome limitations introduced by forward regression and utilize it for models with high-dimensional predictors.

2.1.1 Feature Selection using BIC criteria

The problem of variable selection with an ultrahigh-dimensional predictor becomes a problem of fundamental importance. The traditional method of best subset selection is computationally infeasible for high dimensional data. As a result, various shrinkage methods have gain a lot of popularity. All those methods are very useful and can be formulated as penalized optimization problems, which could be selection consistent, if the sample size is much larger than the predictor dimension. However, if the predictor dimension is much larger than the sample size, the story changes drastically.

One frequently used assumption is the so-called sparsity condition which assumes that the effective

contribution to a dependent variable rests on a much small number of regressors than the sample size. The challenge then is to find those 'true' regressors from a much larger number of candidate variables. This leads to a surging interest in new methods and theory for regression model selection with $p \gg n$.

An *et al.* [92] revisited the classical forward and backward stepwise regression methods for model selection and adapted them to the cases with the number of candidate variables p greater than the number of observations n. In the noiseless case, they gave definite upper bounds for the number of forward search steps to recover all relevant variables, given each step of the forward search is approximately optimal in reduction of residual sum of squares, up to a fraction.

In the presence of noise, they proposed two information criteria BICP (BIC modified for a case with large number of predictors) and BICC (BIC with an added constant) that overcome the difficulties related to employing regular BIC and AIC. These criteria serve as a stopping rule in the stepwise search: the BICP increases the penalty to overcome overfitting and the BICC controls the residuals in the sense that it will stop the search before the residuals diminish to 0 as the number of selected variables increases to n.

In addition, they proved that the BICP stops the forward search as soon as it recovers all relevant variables and removes all extra variables in the backward deletion, which lead to the selection consistency of the estimated models. The algorithm can be summarized as follows. Consider a linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is an n-vector of random responses, $\mathbf{X} = (x_1, \ldots, x_p)$ is a $n \times p$ design matrix, $\boldsymbol{\beta}$ is a p-vector of regression coefficients, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)' \sim N(0, \sigma^2 \boldsymbol{I})$, where $\sigma^2 > 0$ is an unknown but fixed constant and $\boldsymbol{I}$ denotes an identity matrix. Let $I_n = \{1 \leq i \leq p : \beta_{n,i} \neq 0\}$ and $d_n = |I_n|$ denote the number of elements in I_n.

Further, for any subject $J \subset \{1, \ldots, p\}$, let $\mathbf{X}_J$ denote the $n \times |J|$ matrix consisting of the columns of $\mathbf{X}$ corresponding to the indices in J, and β_J the $|J|$-vector consisting of the components β

corresponding to the indices of J. Put

$$\boldsymbol{P}_J = \mathbf{X}_J(\mathbf{X}_J^{\mathrm{T}}\mathbf{X}_J)^{-}\mathbf{X}_J^{\mathrm{T}}, \quad \boldsymbol{P}_J^{\perp} = I_n - \boldsymbol{P}_J, \quad L_{u,v}(J) = u^{\mathrm{T}}\boldsymbol{P}_J^{\perp}v, \quad u, v \in \boldsymbol{R}^n \tag{2.2}$$

$\boldsymbol{P}_J$ is a projection matrix onto the linear space spanned by the columns of $\mathbf{X}_J$, $L_{y,y}(J)$ is the sum of squared residuals resulted from the least square fitting $\hat{\boldsymbol{y}} = \mathbf{X}_J\hat{\boldsymbol{\beta}}_J = \boldsymbol{P}_J\boldsymbol{y}$. The algorithm concerned is based on a combined use of the standard stepwise addition and deletion with some adjusted information criteria and can be described in the following steps:

Stage I - Forward Addition:

1. Let $J_1 = \{j_1\}$, where $j_1 = \arg\min_{1\le i\le p} L_{y,y}(\{i\})$. Put

$$\mathrm{BICP}_1 = \log\{L_{y,y}(J_1)/n\} + 2\log\{p/n\}$$

2. Continue with $k = 1, 2, 3, \ldots,$ provided $\mathrm{BICP}_k < \mathrm{BICP}_{k-1}$, where

$$\mathrm{BICP}_k = \log\{L_{y,y}(J_k)/n\} + 2k\log\{p/n\}$$

   In the above expression, $J_k = J_{k-1} \cup \{j_k\}$

3. For $\mathrm{BICP}_k \ge \mathrm{BICP}_{k-1}$, let $\tilde{k} = k - 1$, and $\hat{I}_{n,1} = J_{\tilde{k}}$

Stage II - Backward deletion:

1. Let $\mathrm{BICP}^*_{\tilde{k}} = \mathrm{BICP}_{\tilde{k}}$ and $J^*_{\tilde{k}} = \hat{I}_{n,1}$

2. Continue with $k = \tilde{k} - 1, \tilde{k} - 2, \ldots,$ providing $\mathrm{BICP}^*_k \le \mathrm{BICP}^*_{k+1}$, where

$$\mathrm{BICP}^*_k = \log\{L_{y,y}(J^*_k)/n\} + 2k\log\{p/n\}$$

   In the above expression, $J^*_k = J_{k+1}\backslash\{j_k\}$

3. $\mathrm{BICP}^*_k > \mathrm{BICP}^*_{k+1}$, $\tilde{k} = k + 1$, and $\hat{I}_{n,2} = J^*_{\tilde{k}}$

The drawback of the method is that it is unclear whether the results are applicable to high-dimensional GLMs.

2.1.2 Forward Regression with High-Dimensional Predictors

Consider, for example, those useful methods with non-convex objective functions (e.g., bridge regression, the SCAD, etc). With the predictor dimension much larger than the sample size, computationally how to optimize those non-convex objective functions remains a nontrivial task. Efficient algorithms (e.g., LARS) do exist for LASSO-type methods, where the objective functions are strictly convex. However, those methods are not selection consistent under a general design condition. Another reasonable solution can be variable screening, such as very popular yet classical method Forward Regression.

Motivated by SIS method [87], Wang [93] proposed a Forward Regression (FR) method for ultrahigh-dimensional variable selection. It was showed that FR method can identify all relevant predictors consistently, even if the number of predictors is significantly larger than the sample size. Particularly, FR is capable of discovering all relevant predictors within a finite number of steps, given that the dimension of the true model is finite. To select the final model from the set of candidate models, Wang [93] makes use of BIC criteria introduced by Chen and Chen [94]. The resulting model can then serve as an excellent starting point, from where many existing variable selection methods can be applied directly.

FR algorithms can be summarized as follows. Suppose $(\mathbf{X}_i, \boldsymbol{Y}_i)$ are observation from the ith subject ($1 \leq i \leq n$), $\boldsymbol{Y}_i \in \boldsymbol{R}^1$ is the response variable, and $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}} \in \boldsymbol{R}^d$ is ultrahigh-dimensional predictor with $d \gg n$. The response and predictor features are linked as $\boldsymbol{Y}_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\sigma}_i$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}} \in \boldsymbol{R}^d$ and $\boldsymbol{\sigma}_i$ is a random noise. Let $\boldsymbol{M} = \{j_1, \ldots, j_{d^*}\}$ denote an arbitrary model with $X_{j_1}, \ldots, X_{j_{d^*}}$ as relevant predictors. Then the full model is defined as $\boldsymbol{\Gamma} = \{1, \ldots, d\}$ and the true model as $\boldsymbol{\tau} = \{j : \beta_j \neq 0\}$. Additionally, $\mathbf{X}_{i(M)} = \{X_{ij} : j \in M\}$ denotes the subvector of $\mathbf{X}_{i(M)}$ corresponding to $\boldsymbol{M}$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \in \boldsymbol{R}^n$ is the response vector, and $\boldsymbol{\xi}_M = (X_1, \ldots, X_n) \in \boldsymbol{R}^{n \times d}$ is the sub-design matrix corresponding to $\boldsymbol{M}$.

FR algorithm is implemented in three major steps:

1. (Initialization) It starts with initialization, that is, setting $S^0 = \emptyset$.

2. (Forward Regression) $S^{(k-1)}$ is given at the kth step. For every $j \in \mathbf{\Gamma} \backslash S^{(k-1)}$, it constructs a candidate model $\boldsymbol{M}_j^{(k-1)} = S^{(k-1)} \cup \{j\}$. Then it computes $\mathrm{RSS}_j^{(k-1)} = \mathbf{Y}^{\mathrm{T}}\{I^{\mathrm{T}} - \hat{\boldsymbol{H}}_j^{(k-1)}\}\mathbf{Y}$, where

$$\hat{\boldsymbol{H}}_j^{(k-1)} = \boldsymbol{\xi}_{\boldsymbol{M}_j^{(k-1)}} \left\{ \boldsymbol{\xi}_{\boldsymbol{M}_j^{(k-1)}}^{\mathrm{T}} \boldsymbol{\xi}_{\boldsymbol{M}_j^{(k-1)}} \right\}^{-1} \boldsymbol{\xi}_{\boldsymbol{M}_j^{(k-1)}}^{\mathrm{T}}$$

is a projection matrix. It finds $a_k = \arg\max_{j \in \mathbf{\Gamma} \backslash S^{(k-1)}} \mathrm{RSS}_j^{(k-1)}$ and updates $S^k = S^{(k-1)} \cup \{a_k\}$ accordingly.

3. (Solution Path) Then FR algorithms iterates step 2 n times and generates total of n nested candidate models and collects these models by a solution path $\mathbf{S} = \{S^k : 1 \leq k \leq n\}$ with $S^k = \{a_1, \ldots, a_k\}$.

Authors showed both theoretically and numerically that FR can discover all relevant predictors consistently, even if the predictor dimension is substantially larger that the sample size. However, the proposed method is limited to linear regression models in high-dimensional settings only .

2.1.3 A Stepwise Regression Algorithm for High-Dimensional Variable Selection

Hwang *et al.* [95] proposed a stepwise regression algorithm with a simple stopping rule for the identification of significant predictors and interactions among a huge number of variables in various statistical models. It improves the results of the Forward Regression method called paring-down variation (SPV) algorithm, proposed by Hwang and Hu [96], which was limited to the analysis of the variation model for continuous responses, and required independence between factor predictors. The new stepwise regression algorithm, like ordinary stepwise regression, at each forward selection step includes a variable in the current model if the test statistic of the enlarged model with the predictor against the current model has the minimum p-value among all the candidates and is smaller than a predetermined threshold. Instead of using conventional information types of criteria, the threshold is determined by a lower percentile of the beta distribution. The proposed stopping rule is based on the well-known theoretical properties that (1) the p-values of the test statistics are

Unif(0, 1) distributed if the predictors are irrelevant to the responses and (2) the minimum of m independent *Unif*(0, 1) random variables can be assumed to be beta distributed with parameters 1 and m approximately [97, 98]. The algorithm can be summarized as follows.

Suppose $\boldsymbol{Y}$ is an n-vector of responses and $\mathbf{X} = (X_1, \ldots, X_p)$ is a $n \times p$ design matix of variables with $p \gg n$. Let S be a subset of $\{0, 1, \ldots, p\}$ and denote $\mathbf{X}_S$ as the sub-matrix of $\mathbf{X}$ obtained by extracting its columns corresponding to the indices in S. In addition let $\boldsymbol{M}_S$ be the model relating the distribution of $\boldsymbol{Y}$ to the predictors $\mathbf{X}_S$ through a function of the linear predictor $\mathbf{X}_S\boldsymbol{\beta}_S$ with parameter vector $\boldsymbol{\beta}_S$ of size |S|. Finally, let denote the vector of residuals from the fitted model $\boldsymbol{M}_S$ is denoted by $\boldsymbol{R}_S$. Then the algorithm can be expressed in the following steps.

Forward Selection

Step 1: Start with the null model $\boldsymbol{M}_S$, where $S = \varnothing$

Step 2: Let the step count be $l = |S| + 1$

Step 3: Calculate the correlation between $\boldsymbol{R}_S$ and $\mathbf{X}_j$, denoted as r_j, for $j = 1, \ldots, p$. Set $D = \{1 \leq j \leq p : |r_j| \text{ is among the first } d \text{ largest of all}\}$, where $d = [n/\log\{n\}]$

Step 4: Test the difference in the goodness-of-fit between each $\boldsymbol{M}_{S\cup\{j\}}$ against $\boldsymbol{M}_S$ for all $j \in D\backslash S$

Step 5: Replace S with $S \cup \{j\}$ when the test statistic of $M_{S\cup\{j\}}$ against $\boldsymbol{M}_S$ has the minimum p-value, denote as p_l, among the $|D\backslash S|$ competing models

Step 6: Stop forward selection and go to Step 7 when $p_h >$ the 10th percentile of Beta$(1, p - h + 1)$ for $h = l, l - 1, \ldots, l - 9$; otherwise, go to Step 2

Backward Selection:

Step 7: Set $\pi = exp(\mu - z \times v)$ where μ and v are the sample mean and standard deviation of $\log\{p_h\}$

Step 8: Test the difference in the goodness-of-fit between $\boldsymbol{M}_S$ and $M_{S\backslash\{j\}}$, for each $j \in S$

Step 9: Replace S with $S\backslash\{j_S\}$ and go to Step 8 if the test statistic of $\boldsymbol{M}_S$ against $\boldsymbol{M}_{S\backslash\{j_S\}}$ has the largest p-value e among all the reduced models and is larger than π; otherwise, stop and report the set of remaining predictors $\{\mathbf{X}_j, j \in S\}$ as final influential predictors

The main drawback of this method is that it was not supported with theoretical properties on model selection.

2.1.4 Generalized Linear Models with High-Dimensional Predictors via Forward Regression: offset approach

As the dimension of predictors defies any existing modeling approaches, feature screening has been commonly used for dimension reduction. The most popular screening approach is marginal screening [87], which selects variables based on their marginal associations with the response. However, marginal screening may miss signals that are marginally unimportant but conditionally important [88], resulting in biased predictive results.

Conditional screening methods have been known as an alternative to well-known marginal screening, as they identify marginally weak but conditionally important variables. Nevertheless, the initial conditioning set need to be fixed for the most of existing conditional screening methods and if not chosen properly, may produce false positives and false negatives, and the selected variable might depend on the conditioning set. Moreover, screening approaches typically need to involve tuning parameters and extra modeling steps in order to reach a final model.

Zheng *et al.* [99] proposed a sequential conditioning (SC) approach, wherein variables sequentially enter the conditioning set according to the increment of likelihood. The procedure updates the conditioning set at each iteration based on the extended Bayesian information criterion (EBIC), and constructs an offset term based on the variables in this set. In essence, this offset summarizes the information contained in the updated conditioning set, and it searches for a new variable that maximizes the likelihood given the offset term. The authors emphasize that the proposed SC approach deviates fundamentally from the variable screening or selection approaches as it naturally leads to a final model when the procedure terminates. The SC approach can be summarized as follows.

Suppose $(\mathbf{X}_i, Y_i)$ are observations from the ith subject $(1 \leq i \leq n)$, $Y_i \in \boldsymbol{R}^1$ is the response variable, and $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})^T$ is a collection of p+1 predictors for the ith sample and $X_{i0} = 1$ corresponds to the intercept. SC modeling focuses on GLMs by assuming that the conditional

density of Y_i given X_i belongs to the linear exponential family

$$\pi(Y \mid \mathbf{X}) = \exp\{Y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - b(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) + \mathcal{A}(Y)\}, \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of coefficients, β_0 is the intercept, and $\mathcal{A}(\cdot)$ and $b(\cdot)$ are known functions. This model with a canonical link function and a unit dispersion parameter, belongs to a larger exponential family [100]. It also assumes that $b(\cdot)$ is twice continuously differentiable with a non-negative second derivative $b''(\cdot)$. In addition, it uses $\mu(\cdot)$ and $\sigma(\cdot)$ to denote $b'(\cdot)$ and $b''(\cdot)$, i.e. the mean and variance functions, respectively. For example, $b(\theta) = \log(1 + \exp(\theta))$ in a logistic distribution and $b(\theta) = \exp(\theta)$ in a Poisson distribution. Let $\mathbb{E}_n\{f(\xi)\} = n^{-1}\sum_{i=1}^{n} f(\xi_i)$ denote the mean of $\{f(\xi_i)\}_{i=1}^{n}$ for a sequence of i.i.d. random variables ξ_i $(i = 1, \ldots, n)$ and a non-random function $f(\cdot)$.

The loglikelihood function, apart from an additive constant is

$$n^{-1}\sum_{i=1}^{n} L(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, Y_i) = \mathbb{E}_n\{L(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, Y)\} \tag{2.4}$$

It uses $\boldsymbol{\beta}_* = (\beta_{*0}, \beta_{*1}, \ldots, \beta_{*p})^{\mathrm{T}}$ to denote the true values of $\boldsymbol{\beta}$. Then the true model is $\mathcal{M} = \{j : \beta_{*j} \neq 0, j \geq 1\} \cup \{0\}$, which consists of the intercept and all variables with nonzero effects. The estimate of $\mathcal{M}$ is denoted as $\hat{\mathcal{M}}$. It elaborate on the idea of building model with the proposed SC approach. The key is to include an offset term which summarizes the information acquired from the previous selection steps and to search for a new candidate variable that maximizes the likelihood with such an offset.

An SC approach algorithm starts with initial index set, S_0, and initial offset, O_0. Having $S_0 = \{0\}$, $O_0 = \hat{\beta}_{S_0}$, where $\hat{\beta}_{S_0}$ is estimated intercept without any covariates. First, with such O_0, it computes $\hat{\beta}_j^{(1)} = \arg\max_{\boldsymbol{\beta}} l_{O_0,j}(\beta)$ for $\mathrm{j} \in \{0, 1, \ldots, p\}$, where $l_{O_0,j}(\boldsymbol{\beta})$ is the average log-likelihood of the regression model. Then, $j_1 = \arg\max_{j\in\{0,1,\ldots,p\}} l_{O_0,j}(\hat{\beta}_j^{(1)})$, $S_1 = \{0, j_1\}$, and $O_1 = \mathbf{X}_{S_1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{S_1}$. Iteratively, for $k \geq 1$, given S_k and O_k, it computes $\hat{\beta}_j^{(k+1)} = \arg\max_{\beta} l_{O_k,j}(\boldsymbol{\beta})$ for $j \in S_k^c$. Then, $j_{k+1} = \arg\max_{j\in S_k^c} l_{O_k,j}(\hat{\boldsymbol{\beta}}_j^{(k+1)})$, $S_{k+1} = S_k \cup \{j_{k+1}\}$, and $O_{k+1} = \mathbf{X}_{S_{k+1}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{S_{k+1}}$. To decide whether it recruits another variable j_{k+1} or stops procedure at kth step, it computed EBIC on set S_{k+1}, where

$$\mathrm{EBIC}(S_{k+1}) = -2\ell_{S_{k+1}}(\hat{\boldsymbol{\beta}}_{S_{k+1}}) + (k+1)(\log n + 2\eta \log p)/n$$

If EBIC(S_{k+1})≥EBIC(S_k), it stops and declares S_k the final model.

The SC approach is computationally efficient as it maximizes the likelihood with respect to only one covariate at each step given the offset, the property that was not observed in other methods. The main drawback of SC approach is that it may be suboptimal compared to a full scale forward optimization approach. Additionally, the consistency of the estimated model parameters has not been addressed in related literature.

2.1.5 Cox Models with High-Dimensional Predictors via Forward Regression

As mentioned, forward regression can consistently identify all relevant predictors in high-dimensional linear regression settings by using EBIC stopping rule. However, existing results from recent works are based on the sum of residual squares from linear models and it is not certain whether forward regression can be applied to more general regression settings, such as Cox proportional hazards models since the results are based on the sum of residual squares from linear models.

There has been active research in developing high-dimensional screening tools for survival data. These works include principled sure screening [101], feature aberration at survival times screening [102] and conditional screening [103], quantile adaptive sure independence screening [104], a censored rank independence screening procedure [105], and integrated powered density screening [106]. However, the screening methods require a threshold to dictate how many variables to retain, for which unfortunately there are no clear rules.

Zhao and Li [101] did tie the threshold with false discoveries, but it still needs to prefix the number of false positives that users are willing to tolerate. Recently, Li *et al.* [107] designed a model-free measure, namely the survival impact index, that sensibly captures the overall influence of a covariate on the survival outcome and can help guide selecting important variables. However, even this method, like the other screening methods, does not directly lead to a final model, for which extra modeling steps have to be implemented.

Hong *et al.* [108] introduced a new forward variable selection procedure for survival data based on partial likelihood. It selects important variables sequentially according to the increment of partial

likelihood, with a stopping rule based on EBIC. The algorithm for the proposed method is the following.

Suppose n objects with p covariates are observed, where $p \gg n$. For subject i, denote by X_{ij} the jth covariate for subject i, write $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, and let T_i and C_i be the underlying survival and censoring times. We only observe $Y_i = \min(T_i, C_i)$, and the event indicator $\delta_i = I(T_i \leq C_i)$, where I is the indicator function. We assume random censoring such that C_i and T_i are independent given $\mathbf{X}$. To link T_i to $\mathbf{X}_i$, for each $i \in \{1, \ldots, n\}$, we consider the Cox proportional hazards model

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp\{\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{X}_i\}, \tag{2.5}$$

where λ_0 is the unspecified baseline hazard function and $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^{\mathrm{T}}$ is the vector of regression coefficients. Additionally, let $S \subset \{1, 2, \ldots, p\}$ be an index set and $|S|$ cardinality of S. First, we initialize $S_0 = \emptyset$ and sequentially select the sets of covariates such that $S_0 \subset S_1 \subset \cdots \subset S_k$. At the $(k+1)$th step the algorithm selects a new covariate not observed in S_k and then decides whether it includes the new variable into selection and proceeds to the next step or stops at the kth step. The selection criteria is based on the partial likelihood. Given S_k, for every $j \in S_k^c$, it fits a Cox model on the variables indexed by $S_{k,j}$, where $S_{k,j} = S_k \cup j$. Then it computes an increment of log partial likelihood for each $j \in S_k^c$, that is, $l_{S_{k,j}}(\hat{\boldsymbol{\beta}}_{S_{k,j}}) - l_{S_k}(\hat{\boldsymbol{\beta}}_{S_k})$, where $l_S(\hat{\boldsymbol{\beta}}_S)$ is log partial likelihood function given $\mathbf{X}_S$:

$$l_S(\hat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} \int_0^{\tau} \left[\boldsymbol{\beta}_S^{\mathrm{T}} \mathbf{X}_{iS} - ln\left\{ \sum_{l=1}^{n} \bar{Y}_l(t) \exp \boldsymbol{\beta}_S^{\mathrm{T}} \mathbf{X}_{lS} \right\} \right] dN_i(t), \tag{2.6}$$

where $N_i(t) = I(Y_i \leq t, \delta_i = 1)$ is the counting process, $\bar{Y}_l(t) = I(Y_i \geq t)$ is the at-risk process, and $\tau > 0$ is the study duration such that $P(Y \geq \tau) > 0$. The candidate index is chosen as $j^* = \arg\max_{j \notin S_k} l_{S_{k,j}}(\hat{\boldsymbol{\beta}}_{S_{k,j}}) - l_{S_k}(\hat{\boldsymbol{\beta}}_{S_k})$ and the index set is updated $S_{k+1} = S_k \cup \{j^*\}$.

To decide whether to stop at the kth step or to include j^* in the selection and proceed to the next step, the algorithm makes its decision based on EBIC criteria, which is defined as follows:

$$\text{EBIC}(S_{k+1}) = -2l_{S_{k+1}}(\hat{\beta}_{S_{k+1}}) + (k+1)\big(\ln\{d\} + 2\eta\ln\{p\}\big), \tag{2.7}$$

where $d = \delta_1 + \cdots + \delta_n$ is the number of events and η some positive constant. If $\text{EBIC}(S_{k+1}) > \text{EBIC}(S_k)$, the algorithm stops and declares S_k the final model, otherwise it proceeds to the next stop. They showed that if the dimension of the true model is finite, within a finite number of steps forward regression can discover all relevant predictors, with the entry order determined by the size of the likelihood increment.

The proposed model could potentially be the first work that investigated the partial likelihood-based forward regression in survival models with high-dimensional predictors. Moreover, it represents technical advances and a broadened scope compared to the existing forward regression (e.g., [109], [110], [93]), and it improves the partial likelihood-based variable selection developed by [111], [112] for survival data in low dimensional settings. The disadvantage of the proposed work is that it does not address parameter estimation, which limits its usage in building predictive models.

2.1.6 A Stepwise Regression Method and Consistent Model Selection for High-Dimensional Sparse Linear Models

Stepwise least squares regression is widely used in applied regression analysis to handle a large number of input variables, which consists of forward selection of input variables in a "greedy" manner so that the selected variable at each step minimizes the residual sum of squares after least squares regression is performed on it together with previously selected variables, a stopping rule to terminate forward inclusion of variables, and stepwise backward elimination of variables according to some criterion.

Ing *et al.* [113] developed an asymptotic theory for a version of stepwise regression in the context of high-dimensional regression under certain sparsity assumptions. They introduced a fast stepwise regression method, called the orthogonal greedy algorithm (OGA), that selects input variables to enter a p-dimensional linear regression model (with $p \gg n$, the sample size) sequentially so that the selected variable at each step minimizes the residual sum squares. They derived the convergence

rate of OGA and developed a consistent model selection procedure along the OGA path that can adjust for potential spuriousness of the greedily chosen regressors among a large number of candidate variables. The resultant regression estimate is shown to have the oracle property of being equivalent to least squares regression on an asymptotically minimal set of relevant regressors under a strong sparsity condition.

The forward stepwise component of the procedure is compressed sensing and approximation theory, which focuses on approximations in noiseless models. They also developed a fast iterative procedure for updating OGA that uses componentwise linear regression similar to the L_2-boosting procedure of Buhlmann and Yu [114] and does not require matrix inversion. Consider the linear regression model

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.8}$$

with p predictors and $x_{i1}, x_{i2}, \ldots, x_{ip}$ that are uncorrelated with the mean-zero random disturbances ϵ_i. As mentioned, L_2-boosting is an iterative procedure that generates a sequence of linear approximations $\hat{y}_k(x)$ of the regression function (with $\alpha = 0$), by applying componentwise linear least squares to the residuals obtained at each iteration.

Initializing with $\hat{y}_0(\cdot) = 0$, it computes the residuals $U_i^k := y_i - \hat{y}_k(x_i)$, $1 \le i \le n$, at the end of the kth iteration and chooses $x_{i,\hat{j}_{k+1}}$ on which the pseudo-responses $U_i^{(k)}$ are regressed, such that

$$\hat{j}_{k+1} = \underset{1 \le j \le p}{\arg\min} \sum_{i=1}^{n} (U_i^{(k)} - \tilde{\beta}_j^{(k)} x_{ij})^2, \tag{2.9}$$

where

$$\tilde{\beta}_j^{(k)} = \sum_{i=1}^{n} U_i^{(k)} x_{ij} \Bigg/ \sum_{i=1}^{n} x_{ij}^2.$$

This yields the update

$$\hat{y}_{k+1}(x) = \hat{y}_k(x) + \tilde{\beta}_{\hat{j}_{k+1}}^{(k)} x_{\hat{j}_{k+1}}. \tag{2.10}$$

The procedure is then repeated until a pre-specified upper bound m on the number of iterations is reached. When the procedure stops at the mth iteration, $y(x)$ is approximated by $\hat{y}_m(x)$.

OGA uses the variable selector (2.9). Since $\sum_{i=1}^{n}(U_i^{(k)} - \tilde{\beta}_j^{(k)} x_{ij})^2 / \sum_{i=1}^{n}(U_i^{(k)})^2 = 1 - r_j^2$, where r_j is the correlation coefficient between x_{tj} and $U_i^{(k)}$, (2.9) chooses the predictor that is most correlated with $U_i^{(k)}$ at the kth stage. However,the implementation of OGA updates (2.10) in another way and also carries out an additional linear transformation of the vector $\mathbf{X}_{\hat{j}_{k+1}}$ to form $\mathbf{X}^{\perp}_{\hat{j}_{k+1}}$, where $\mathbf{X}_j = (x_{ij}, \ldots, x_{nj})^{\mathrm{T}}$. The idea is to orthogonalize the predictor variables sequentially so that OLS can be computed by componentwise linear regression, thereby circumventing difficulties with inverting high-dimensional matrices in the usual implementation of OLS.

With the orthogonal vectors $\mathbf{X}_{\hat{j}_1}, \mathbf{X}^{\perp}_{\hat{j}_2}, \ldots, \mathbf{X}^{\perp}_{\hat{j}_k}$ already computed in the previous stages, it can compute the projection $\hat{\mathbf{X}}_{\hat{j}_{k+1}}$ of $\mathbf{X}_{\hat{j}_{k+1}}$ into the linear space by adding the k projections into the respective one-dimensional linear spaces. This also yields the residual vector $\mathbf{X}^{\perp}_{\hat{j}_{k+1}} = \mathbf{X}_{\hat{j}_{k+1}} - \hat{\mathbf{X}}_{\hat{j}_{k+1}}$. OGA uses the following updates in lieu of (2.10):

$$\hat{y}_{k+1}(x) = \hat{y}_k(x) + \hat{\beta}^{(k)}_{\hat{j}_{k+1}} x^{\perp}_{\hat{j}_{k+1}}, \tag{2.11}$$

where

$$\hat{\beta}^{(k)}_{\hat{j}_{k+1}} = \sum_{i=1}^{n} U_i^{(k)} x^{\perp}_{i,\hat{j}_{k+1}} \Bigg/ \sum_{i=1}^{n} (x^{\perp}_{i,\hat{j}_{k+1}})^2.$$

By sequentially orthogonalizing the input variables, OGA preserves the attractive computational features of componentwise linear regression in PGA. However, unlike PGA for which the same predictor variable can be entered repeatedly, OGA excludes variables that are already precluded from further consideration in (2.9).

In addition, they developed a consistent model selection procedure along an OGA path under a "strong sparsity" condition that the nonzero regression coefficients satisfying the weak sparsity condition are not too small. Applying the convergence rate of OGA, they proved that, with probability approaching 1 as $n \to \infty$, the OGA path includes all relevant regressors when the

number of iterations is large enough. They modified the model selection criteria like BIC and called it high-dimensional information criterion (HDIC), which is defined as:

$$\text{HDIC}(J) = n\log\{\hat{\sigma}_J^2\} + |J|w_n\log\{p\}, \tag{2.12}$$

where J is a non-empty subset of $\{1, \ldots, p\}$ and $\hat{\sigma}_J^2 = n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_{i;J})^2$. OGA+HDIC is shown to select the smallest set of all relevant variables along the OGA path with probability approaching 1 (and is therefore variable-selection consistent).

2.2 Our Contribution

Although methods described in the previous sections brought novelty to the research area and made a significant contribution to it, they still have some drawbacks. First, once a variable is identified by the forward selection, it is not removable from the list of selected variables. Hence, false positives are unavoidable without a systematic elimination procedure. Second, most of the existing works focus on variable selection and are silent with respect to estimation accuracy.

To address the first issue, some works have been proposed to add backward elimination steps once forward selection is accomplished, as backward elimination may further eliminate false positives from the variables selected by forward selection. For example, An *et al.* [92] and Ing *et al.* [113] proposed a stepwise selection for linear regression models in high-dimensional settings and proved model selection consistency. However, it is unclear whether the results hold for high-dimensional GLMs; Hwang *et al.* [95] proposed a similar stepwise algorithm in high-dimensional GLM settings, but with no theoretical properties on model selection. Moreover, none of the relevant works have touched upon the accuracy of estimation.

We extend a stepwise regression method to accommodate GLMs with high-dimensional predictors. Our method, termed STEPWISE hereafter and introduced in Chapter 3, embraces both model selection and estimation. It starts with an empty model or pre-specified predictors, scans all features and sequentially selects features, and conducts backward elimination once forward selection is completed. Our proposal controls both false negatives and false positives in high dimensional

settings: the forward selection steps recruit variables in an inclusive way by allowing some false positives for the sake of avoiding false negatives, while the backward selection steps eliminate the potential false positives from the recruited variables.

We use different stopping criteria in the forward and backward selection steps, to control the numbers of false positives and false negatives. It is achieved by adding flexibility with two tuning parameters in the stopping criteria, which strengthens our algorithm. Values of these parameters define how many parameters will likely be included in the model. For instance, a small value of the tuning parameter in the forward selection will include more variables, and a large value will recruit too few features. In similar fashion, in the backward elimination step, a large value of the parameter would eliminate more variables and vice versa for a small value. It is worth mentioning that our method includes forward selection as a special case when the tuning parameter is equal to 0, making the algorithm more flexible.

Moreover, we prove that, under a sparsity assumption of the true model, the proposed approach can discover all of the relevant predictors within a finite number of steps, and the estimated coefficients are consistent, a property still unknown to the literature. Finally, our GLM framework enables our work to accommodate a wide range of data types, such as binary, categorical and count data.

Extensive numerical studies have been conducted to compare STEPWISE procedure with the other competing methods mentioned in previous subsections. Specifically, we compared our algorithm with LASSO, dgLARS, forward regression (FR), the SC approach, and the screening methods such as SIS. Our numerical studies included simulations: we compared the proposed method with the other methods over comprehensive simulated studies covering different model structures and variable dependencies. All these examples were tested over various model types, including Normal, Binomial, and Poisson models. The obtained results have indicated that the STEPWISE algorithm was able to detect all the true signals with nearly zero false positive rate. In addition, it outperformed the other methods by selecting more true positives with fewer false positives.

Moreover, our numerical studies included real data analysis, which aimed to demonstrate the utility of our method with real-life scenarios. We analyzed data obtained from studies about

gene regulation in the mammalian eye, esophageal squamous cell carcinoma, and neurobehavioral impairment from total sleep deprivation. We demonstrated that STEPWISE achieved comparable prediction accuracy, specificity, sensitivity, and AUC by selecting fewer variables that the other variable selection methods.

Finally, to enhance the predictive power of the proposed method, we developed a multi-stage hybrid machine learning method. It incorporates a stacking technique and includes both model-free and model-based methods. Specifically, it comprises Random Forest (RF), Extreme Gradient Boosting Machine (XGBoost), Support Vector Machine (SVM), Artificial Neural Network (ANN), and LASSO models along with the STEPWISE procedure. The numerical study has shown an improvement in the predictive power of our method. Furthermore, we developed a web application that enables users to utilize the aforementioned method in practice.

To recap, our proposed method distinguishes from the existing stepwise approaches in high-dimensional settings. For example, it improves An *et al.* [92] and Ing *et al.* [113] by extending the work to a more broad GLM setting and Hwang *et al.* [95] by establishing the theoretical properties. Compared with the other variable selection and screening works, our method produces a final model in ultrahigh-dimensional settings, without applying a pre-screening step which may produce unintended false negatives.

Under some regularity conditions, the method identifies or includes the true model with probability going to 1. Moreover, unlike the penalized approaches such as LASSO, the coefficients estimated by our STEPWISE selection procedure in the final model will be consistent, which are useful for gauging the real effect sizes of risk factors.

CHAPTER 3

STEPWISE METHOD: THEORY AND APPLICATIONS

3.1 Model Setup

Let $(\mathbf{X}_i, Y_i),\ i = 1, \ldots, n$, denote n independent and identically distributed (i.i.d.) copies of $(\mathbf{X}, Y)$. Here, $\mathbf{X} = (1, X_1, \ldots, X_p)^{\mathrm{T}}$ is a $(p+1)$-dimensional predictor vector with $X_0 = 1$ corresponding to the intercept term, and Y is an outcome. Suppose that the conditional density of Y, given $\mathbf{X}$, belongs to a linear exponential family:

$$\pi(Y \mid \mathbf{X}) = \exp\{Y\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} - b(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) + \mathcal{A}(Y)\}, \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of coefficients, β_0 is the intercept, and $\mathcal{A}(\cdot)$ and $b(\cdot)$ are known functions. Model (3.1), with a canonical link function and a unit dispersion parameter, belongs to a larger exponential family [100]. Further, $b(\cdot)$ is assumed twice continuously differentiable with a non-negative second derivative $b''(\cdot)$. We use $\mu(\cdot)$ and $\sigma(\cdot)$ to denote $b'(\cdot)$ and $b''(\cdot)$, i.e. the mean and variance functions, respectively. For example, $b(\theta) = \log(1+\exp(\theta))$ in a logistic distribution and $b(\theta) = \exp(\theta)$ in a Poisson distribution.

Let $L(u, v) = uv - b(u)$ and $\mathbb{E}_n\{f(\xi)\} = n^{-1}\sum_{i=1}^{n} f(\xi_i)$ denote the mean of $\{f(\xi_i)\}_{i=1}^{n}$ for a sequence of i.i.d. random variables ξ_i $(i = 1, \ldots, n)$ and a non-random function $f(\cdot)$. Based on the i.i.d. observations, the log-likelihood function is

$$\ell(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} L(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}, Y_i) = \mathbb{E}_n\{L(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, Y)\}. \tag{3.2}$$

We use $\boldsymbol{\beta}_* = (\beta_{*0}, \beta_{*1}, \ldots, \beta_{*p})^{\mathrm{T}}$ to denote the true values of $\boldsymbol{\beta}$. Then the true model is $\mathcal{M} = \{j : \beta_{*j} \neq 0, j \geq 1\} \cup \{0\}$, which consists of the intercept and all variables with nonzero effects. Overarching goals of ultrahigh-dimensional data analysis are to identify $\mathcal{M}$ and estimate β_{*j} for $j \in \mathcal{M}$. While most of the relevant literature [99, 108] is on estimating $\mathcal{M}$, this work is to accomplish both identification of $\mathcal{M}$ and estimation of β_{*j}.

When p is in the exponential order of n, we aim to generate a predictive model that contains the true model with high probability, and provide consistent estimates of regression coefficients. We further introduce the following notation. For a generic index set $S \subset \{0, 1, \ldots, p\}$ and a $(p+1)$-dimensional vector $\mathbf{A}$, we use S^c to denote the complement of a set S and $\mathbf{A}_S = \{A_j : j \in S\}$ to denote the subvector of $\mathbf{A}$ corresponding to S. For instance, if $S = \{2, 3, 4\}$, then $\mathbf{X}_{iS} = (X_{i2}, X_{i3}, X_{i4})^{\mathrm{T}}$. Moreover, denote by $\ell_S(\boldsymbol{\beta}_S) = \mathbb{E}_n\{L(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S, Y)\}$ the log-likelihood of the regression model of Y on $\mathbf{X}_S$ and denote by $\hat{\boldsymbol{\beta}}_S$ the maximizer of $\ell_S(\boldsymbol{\beta}_S)$. Under model (3.1), we elaborate on the idea of stepwise selection, consisting of the forward and backward stages.

Forward stage: We start with F_0, a set of variables that need to be included according to some *a priori* knowledge, such as clinically important factors and conditions. If no such information is available, F_0 is set to be $\{0\}$, corresponding to a null model. We sequentially add covariates as follows:

$$F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_k,$$

where $F_k \subset \{0, 1, \ldots, p\}$ is the index set of the selected covariates upon completion of the kth step, with $k \geq 0$. At the $(k+1)$th step, we append new variables to F_k one at a time and refit GLMs: for every $j \in F_k^c$, we let $F_{k,j} = F_k \cup \{j\}$, obtain $\hat{\boldsymbol{\beta}}_{F_{k,j}}$ by maximizing $\ell_{F_{k,j}}(\boldsymbol{\beta}_{F_{k,j}})$, and compute the increment of log-likelihood,

$$\ell_{F_{k,j}}(\hat{\boldsymbol{\beta}}_{F_{k,j}}) - \ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}).$$

Then the index of a new candidate variable is determined to be

$$j_{k+1} = \arg\max_{j \in F_k^c} \ell_{F_{k,j}}(\hat{\boldsymbol{\beta}}_{F_{k,j}}) - \ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}).$$

And we update $F_{k+1} = F_k \cup \{j_{k+1}\}$. We then need to decide whether to stop at the kth step or move on to the $(k+1)$th step with F_{k+1}. To do so, we use the following EBIC criterion:

$$\mathrm{EBIC}(F_{k+1}) = -2\ell_{F_{k+1}}(\hat{\boldsymbol{\beta}}_{F_{k+1}}) + |F_{k+1}|n^{-1}(\log n + 2\eta_1 \log p), \tag{3.3}$$

where the second term is motivated by Chen and Chen [115] and $|F|$ denotes the cardinality of a set F.

The forward selection stops if $\text{EBIC}(F_{k+1}) > \text{EBIC}(F_k)$. We denote the stopping step by k^* and the set of variables selected so far by F_{k^*}.

Backward stage: Upon the completion of forward stage, backward elimination, starting with $B_0 = F_{k^*}$, sequentially drops covariates as follows:

$$B_0 \supset B_1 \supset B_2 \supset \cdots \supset B_k,$$

where B_k is the index set of the remaining covariates upon the completion of the kth step of the backward stage, with $k \geq 0$. At the $(k+1)$th backward step and for every $j \in B_k$, we let $B_{k/j} = B_k \setminus \{j\}$, obtain $\hat{\boldsymbol{\beta}}_{B_{k/j}}$ by maximizing $\ell(\boldsymbol{\beta}_{B_{k/j}})$, and calculate the difference of the log-likelihoods between these two nested models,

$$\ell_{B_k}(\hat{\boldsymbol{\beta}}_{B_k}) - \ell_{B_{k/j}}(\hat{\boldsymbol{\beta}}_{B_{k/j}}).$$

The variable that can be removed from the current set of variables is indexed by

$$j_{k+1} = \underset{j \in B_k}{\arg\min}\, \ell_{B_k}(\hat{\boldsymbol{\beta}}_{B_k}) - \ell_{B_{k/j}}(\hat{\boldsymbol{\beta}}_{B_{k/j}}).$$

Let $B_{k+1} = B_k \setminus \{j_{k+1}\}$. We determine whether to stop at the kth step or move on to the $(k+1)$th step of the backward stage according to the following BIC criterion:

$$\text{BIC}(B_{k+1}) = -2\ell_{B_{k+1}}(\hat{\boldsymbol{\beta}}_{B_{k+1}}) + \eta_2 n^{-1} |B_{k+1}| \log n. \tag{3.4}$$

If $\text{BIC}(B_{k+1}) > \text{BIC}(B_k)$, we end the backward stage at the kth step. Let k^{**} denote the stopping step and we declare the selected model $B_{k^{**}}$ to be the final model. Thus, $\hat{\mathcal{M}} = B_{k^{**}}$ is the estimate of $\mathcal{M}$. As the backward stage starts with the k^* variables selected by forward selection, k^{**} cannot exceed k^*.

A strength of our algorithm is the added flexibility with η_1 and η_2 in the stopping criteria for controlling the false negatives and positives. For example, a smaller value of η_1 close to zero in the forward selection step will likely include more variables, thus incur more false positives and less false negatives, whereas a larger value of η_1 will recruit too few variables and cause too many false negatives. Similarly, in the backward selection step, a large η_2 would eliminate more variables and

therefore further reduce more false positives, and vice versa for a small η_2. While finding optimal η_1 and η_2 is not trivial, our numerical experiences suggest a small η_1 and a large η_2 may well balance the false negatives and positives. When $\eta_2 = 0$, no variables can be dropped after forward selection; hence, our proposal includes forward selection as a special case.

Moreover, Zheng *et al.* [99] proposed a sequentially conditioning approach based on offset terms that absorb the prior information. However, our numerical experiments indicate that the offset approach may be suboptimal compared to our full stepwise optimization approach, which will be demonstrated in the simulation studies.

3.2 Theoretical Properties

With a column vector $\mathbf{v}$, let $\|\mathbf{v}\|_q$ denote the L_q-norm for any $q \geq 1$. For simplicity, we denote the L_2-norm of $\mathbf{v}$ by $\|\mathbf{v}\|$, and denote $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ by $\mathbf{v}^{\otimes 2}$. We use $C_1, C_2, \ldots,$ to denote some generic constants that do not depend on n and may change from line to line. The following regularity conditions are set.

1. There exist a positive integer q satisfying $|\mathcal{M}| \leq q$ and $q \log p = o(n^{1/3})$ and a constant $K > 0$ such that $\sup_{|S| \leq q} \|\boldsymbol{\beta}_S^*\|_1 \leq K$, where $\boldsymbol{\beta}_S^* = \arg\max_{\boldsymbol{\beta}_S} E\left[\ell_S(\boldsymbol{\beta}_S)\right]$ is termed the least false value of model S.

2. $\|\mathbf{X}\|_\infty \leq K$. In addition, $E(X_j) = 0$ and $E(X_j^2) = 1$ for $j \geq 1$.

3. Let $\epsilon_i = Y_i - \mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)$. There exists a positive constant M such that the Cramer condition holds, i.e., $E[|\epsilon_i|^m] \leq m! M^m$ for all $m \geq 1$.

4. $|\sigma(a) - \sigma(b)| \leq K|a - b|$ and $\sigma_{\min} := \inf_{|t| \leq K^3} |b''(t)|$ is bounded below.

5. There exist two positive constants, $\kappa_{\min}$ and $\kappa_{\max}$ such that $0 < \kappa_{\min} < \Lambda\left(E\left(\mathbf{X}_S^{\otimes 2}\right)\right) < \kappa_{\max} < \infty$, uniformly in $S \subset \{0, 1, \ldots, p\}$ satisfying $|S| \leq q$, where $\Lambda(\mathbf{A})$ is the collection of all eigenvalues of a square matrix $\mathbf{A}$.

6. $\min_{S:\mathcal{M}\not\subseteq S,|S|\leq q} D_S > Cn^{-\alpha}$ for some constants $C > 0$ and $\alpha > 0$ that satisfies $qn^{-1+4\alpha}\log p \to 0$, where $D_S = \max_{j\in\mathcal{S}^c\cap\mathcal{M}} \left|E\left[\left(\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}) - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)\right)X_j\right]\right|$.

Condition (1), as assumed in Buhlmann *et al.* [116] and Zheng *et al.* [99], is an alternative to the Lipschitz assumption [117, 87]. The bound of the model size allowed in the selection procedure or q is often required in model-based screening methods (see, e.g. [94, 118, 109, 99]). The bound should be large enough so that the correct model can be included, but not too large; otherwise, excessive noise variables would be included, leading to unstable and inconsistent estimates. Indeed, Conditions (1) and (6) reveal that the range of q depends on the true model size $|\mathcal{M}|$, the minimum signal strength, $n^{-\alpha}$, and the total number of covariates, p. The upper bound of q is $o((n^{1-4\alpha}/\log p) \wedge (n^{1/3}/\log p))$, ensuring the consistency of EBIC [115].

Condition (1) also implies that the parameter space under consideration can be restricted to $\mathbb{B} := \{\boldsymbol{\beta} \in \mathbb{R}^{p+1} : \|\boldsymbol{\beta}\|_1 \leq K^2\}$, for any model S with $|S| \leq q$. Condition (2), as assumed in Zhao *et al.* [101] and Kwemou *et al.* [119], reflects that data are often standardized at the pre-processing stage. Condition (3) ensures that Y has a light tail, and is satisfied by Gaussian and discrete data, such as binary and count data [120]. Condition (4) is satisfied by common GLM models, such as Gaussian, Binomial, Poisson and Gamma distributions.

Condition (5) represents the sparse Riesz condition [121] and Condition (6) is a strong "irrepresentable" condition, suggesting that $\mathcal{M}$ cannot be represented by a set of variables that does not include the true model. It further implies that adding a signal variable to a mis-specified model will increase the log-likelihood by a certain lower bound [99]. The signal rate is comparable to the conditions required by the other sequential methods (see, e.g. [93, 109]).

Theorem 3.2.1 develops a lower bound of the increment of the log-likelihood if the true model $\mathcal{M}$ is not yet included in a selected model S.

Theorem 3.2.1. *Suppose Conditions (1) – (6) hold. There exists some constant C_1 such that with probability at least* $1 - 6\exp(-6q\log p)$*,*

$$\min_{S:\mathcal{M}\not\subseteq S,|S|<q}\left\{\max_{j\in S^c}\ell_{S\cup\{j\}}(\hat{\boldsymbol{\beta}}_{S\cup\{j\}}) - \ell_S(\hat{\boldsymbol{\beta}}_S)\right\} \geq C_1 n^{-2\alpha}.$$

Theorem 3.2.1 shows that, before the true model is included in the selected model, we can append a variable which will increase the log-likelihood by at least $C_1 n^{-2\alpha}$ with probability tending to 1. This ensures that in the forward stage, our proposed STEPWISE approach will keep searching for signal variables until the true model is contained. To see this, suppose at the kth step of the forward stage that F_k satisfies $M \nsubseteq F_k$ and $|F_k| < q$, and let r be the index selected by Stepwise. By Theorem 3.2.1, we obtain that, for any $\eta_1 > 0$, when n is sufficiently large,

$$\begin{aligned}\mathrm{EBIC}(F_{k,r}) - \mathrm{EBIC}(F_k) &= -2\ell_{F_{k,r}}(\hat{\boldsymbol{\beta}}_{F_{k,r}}) + (|F_k|+1)n^{-1}(\log n + 2\eta_1 \log p) \\ &\quad - \left[-2\ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}) + |F_k| n^{-1}(\log n + 2\eta_1 \log p)\right] \\ &\leq -2C_1 n^{-2\alpha} + n^{-1}(\log n + 2\eta_1 \log p) < 0,\end{aligned}$$

with probability at least $1 - 6\exp(-6q \log p)$, where the last inequality is due to Condition (6). Therefore, with high probability the forward stage of STEPWISE continues as long as $\mathcal{M} \nsubseteq F_k$ and $|F_k| < q$. We next establish an upper bound of the number of steps in the forward stage needed to include the true model.

Theorem 3.2.2. *Under the same conditions as in Theorem 3.2.1 and if*

$$\max_{S:|S|\leq q} \left\{ \max_{j \in \mathcal{S}^c \cap \mathcal{M}^c} \left| E\left[\left\{Y - \mu(\boldsymbol{\beta}_S^{*\mathrm{T}} \mathbf{X}_S)\right\} X_j\right] \right| \right\} = o(n^{-\alpha}),$$

then there exists some constant $C_2 > 2$ such that $\mathcal{M} \subset F_k$, for some F_k in the forward stage of Stepwise and $k \leq C_2|\mathcal{M}|$, with probability at least $1 - 18\exp(-4q \log p)$.

The "max" condition, as assumed in Section 5.3 of Fan *et al.* [122], relaxes the partial orthogonality assumption that $\mathbf{X}_{\mathcal{M}^c}$ are independent of $\mathbf{X}_{\mathcal{M}}$, and ensures that with probability tending to 1, appending a signal variable increases log-likelihood more than adding a noise variable does, uniformly over all possible models S satisfying $\mathcal{M} \nsubseteq S, |S| < q$. This entails that the proposed procedure is much more likely to select a signal variable, in lieu of a noise variable, at each step. Since EBIC is a consistent model selection criterion [110, 123], the following theorem guarantees termination of the proposed procedure with $\mathcal{M} \subset F_k$ for some k.

Theorem 3.2.3. *Under the same conditions as in Theorem 3.2.2 and if* $\mathcal{M} \not\subset F_{k-1}$ *and* $\mathcal{M} \subset F_k$*, the forward stage stops at the* k*th step with probability going to* $1 - \exp(-3q \log p)$.

Theorem 3.2.3 ensures that the forward stage of STEPWISE will stop within a finite number of steps and will cover the true model with probability at least $1 - q\exp(-3q \log p) \geq 1 - \exp(-2q \log p)$. We next consider the backward stage and provide a probability bound of removing a signal from a set in which the set of true signals $\mathcal{M}$ is contained.

Theorem 3.2.4. *Under the same conditions as in Theorem 3.2.2,* $BIC(S\backslash\{r\}) - BIC(S) > 0$ *uniformly over* $r \in \mathcal{M}$ *and* S *satisfying* $\mathcal{M} \subset S$ *and* $|S| \leq q$*, with probability at least* $1 - 6\exp(-6q \log p)$.

Theorem 3.2.4 indicates that with probability at $1 - 6\exp(-6q \log p)$, BIC would decrease when removing a signal variable from a model that contains the true model. That is, with high probability, back elimination is to reduce false positives.

Recall that F_{k^*} denotes the model selected at the end of the forward selection stage. By Theorem 3.2.2, $\mathcal{M} \subset F_{k^*}$ with probability at least $1 - 18\exp(-4q \log p)$. Then Theorem 3.2.4 implies that at each step of the backward stage, a signal variable will not be removed from the model with probability at least $1 - 6\exp(-6q \log p)$. By Theorem 3.2.2, $|F_{k^*}| \leq C_2|\mathcal{M}|$. Thus, the backward elimination will carry out at most $(C_2 - 1)|\mathcal{M}|$ steps. Combining results from Theorems 3.2.2 and 3.2.3 yields that $\mathcal{M} \subset \hat{\mathcal{M}}$ with probability at least $1 - 18\exp(-4q \log p) - 6(C_2 - 1)|\mathcal{M}|\exp(-6q \log p)$. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}_*$ in model (3.1) at the termination of STEPWISE. By convention, the estimates of the coefficients of the unselected covariates are 0.

Theorem 3.2.5. *Under the same conditions as in Theorem 3.2.2, we have that* $\mathcal{M} \subseteq \hat{\mathcal{M}}$ *and*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\| \to 0$$

in probability.

The theorem warrants that the proposed STEPWISE yields consistent estimates, a property not shared by many regularized methods, including LASSO. Our later simulations verified this. Proof of main theorems and lemmas are provided in the following Chapter.

3.3 Proof of the Theorems

Since $b(\cdot)$ is twice continuously differentiable with a nonnegative second derivative $b''(\cdot)$, $b_{\max} := \max_{|t|\le K^3} |b(t)|$, $\mu_{\max} := \max_{|t|\le K^3} |b'(t)|$, and $\sigma_{\max} := \sup_{|t|\le K^3} |b''(t)|$ are bounded above, where L and K are some constants from Conditions (1) and (2), respectively. Let $\mathbb{G}_n\{f(\xi)\} = n^{-1/2}\sum_{i=1}^n (f(\xi_i) - E[f(\xi_i)])$ for a sequence of i.i.d. random variables ξ_i $(i = 1, \ldots, n)$ and a non-random function $f(\cdot)$.

Given any $\boldsymbol{\beta}_S$, when a variable $X_r, r \in S^c$ is added into the model S, we define the augmented log-likelihood as

$$\ell_{S\cup\{r\}}(\boldsymbol{\beta}_{S+r}) := \mathbb{E}_n \left\{ L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S + \beta_r X_r, Y\right)\right\}. \tag{3.1}$$

We use $\hat{\boldsymbol{\beta}}_{S+r}$ to denote the maximizer of (3.1). Thus, $\hat{\boldsymbol{\beta}}_{S+r} = \hat{\boldsymbol{\beta}}_{S\cup\{r\}}$. In addition, denote the maximizer of $E[\ell_{S\cup\{r\}}(\boldsymbol{\beta}_{S+r})]$ by $\boldsymbol{\beta}^*_{S+r}$. Due to the concavity of the log-likelihood in GLMs with the canonical link, $\boldsymbol{\beta}^*_{S+r}$ is unique.

Proof of Theorem 3.2.1. Given an index set S and $r \in S^c$, let $\mathcal{B}_S^0(d) = \{\boldsymbol{\beta}_S : \|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \le d/(K\sqrt{|S|})\}$ where $d = A_2\sqrt{q^3 \log p/n}$ with A_2 defined in Lemma 3.4.6.

Let Ω be the event that

$$\left\{ \sup_{|S|\le q, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)} \left| \mathbb{G}_n\left[L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]\right| \le 20A_1 d\sqrt{q\log p} \quad \text{and} \right.$$

$$\left. \max_{|S|\le q} |\mathbb{G}_n\left[L(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right]| \le 10(A_1K^2 + b_{\max})\sqrt{q\log p}\right\},$$

where A_1 is some constant defined in Lemma 3.4.4. By Lemma 3.4.4, $P(\Omega) \ge 1 - 6\exp(-6q\log p)$. Thus in the rest of the proof, we only consider the sample points in Ω.

In the proof of Lemma 3.4.6, we show that $\max_{|S|\le q} \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\| \le A_2K^{-1}(q^2\log p/n)^{1/2}$ under Ω. Then given an index set S and $\boldsymbol{\beta}_S$ such that $|S| < q$, $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \le A_2K^{-1}(q^2\log p/n)^{1/2}$, and for

any $j \in S^c$,

$$\ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \geq \inf_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leq A_2 K^{-1}(q^2 \log p/n)^{1/2}} \ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j}) - \ell_S(\boldsymbol{\beta}_S) =$$

$$n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}}, Y)\right] - n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S, Y)\right] -$$

$$\sup_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leq A_2 K^{-1}(q^2 \log p/n)^{1/2}} \left| n^{-1/2}\mathbb{G}_n\left[L(\boldsymbol{\beta}^{\mathrm{T}}_S\mathbf{X}_S, Y) - L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S, Y)\right]\right|$$

$$+ E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}}, Y)\right] - E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S, Y)\right] \geq$$

$$- 20(A_1K^2 + b_{\max})\sqrt{q\log p/n} - 20A_1A_2q^2\log p/n +$$

$$\frac{\sigma_{\min}\kappa_{\min}}{2}\|\boldsymbol{\beta}^*_{S+j} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}\|^2,$$

where the second inequality follows from the event Ω and Lemma 3.4.5.

By Lemma 3.4.1, if $\mathcal{M} \not\subseteq S$, there exists $r \in S^c \cap \mathcal{M}$, such that

$$\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)\| \geq C\sigma^{-1}_{\max}\kappa^{-1}_{\max}n^{-\alpha}.$$

Thus, there exists some constant C_1 that does not depend on n such that

$$\max_{j\in S^c} \ell_{S\cup\{j\}}(\hat{\boldsymbol{\beta}}_{S+j}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \geq \max_{j\in S^c} \ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \geq$$

$$\ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \geq -20(A_1K^2 + b_{\max})\sqrt{q\log p/n} -$$

$$20A_1A_2q^2\log p/n + \frac{C^2\sigma_{\min}\kappa_{\min}n^{-2\alpha}}{2\sigma^2_{\max}\kappa^2_{\max}} \geq C_1n^{-2\alpha}, \tag{3.2}$$

where the first inequality follows from $\hat{\boldsymbol{\beta}}_{S+j}$ being the maximizer of (3.1) and the second inequality follows from Conditions (1) and (6).

Withdrawing the restriction to Ω, we obtain that

$$P\left(\min_{|S|<q, \mathcal{M}\not\subseteq S} \max_{j\in S^c} \ell_{S\cup\{j\}}(\hat{\boldsymbol{\beta}}_{S\cup\{j\}}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \geq C_1n^{-2\alpha}\right) \geq 1 - 6\exp(-6q\log p).$$

□

Proof of Theorem 3.2.2. We have shown that our forward stage will not stop when $\mathcal{M} \nsubseteq S$ and $|S| < q$ with probability converging to 1.

For any $r \in S^c \cap \mathcal{M}^c$, $\boldsymbol{\beta}^*_{S+r}$ is the unique solution to the equation $E\left[\left\{Y - \mu\left(\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}}\right)\right\}\mathbf{X}_{S\cup\{r\}}\right] = \mathbf{0}$. By the mean value theorem,

$$E\left[\left\{Y - \mu\left(\boldsymbol{\beta}^{*\mathrm{T}}_{S}\mathbf{X}_S\right)\right\}X_r\right] = E\left[\left\{\mu\left(\boldsymbol{\beta}^{\mathrm{T}}_{*}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}^{*\mathrm{T}}_{S}\mathbf{X}_S\right)\right\}X_r\right]$$

$$= E\left[\left\{\mu\left(\boldsymbol{\beta}^{\mathrm{T}}_{*}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}^{*\mathrm{T}}_{S}\mathbf{X}_S\right)\right\}X_r\right] - E\left[\left\{\mu\left(\boldsymbol{\beta}^{\mathrm{T}}_{*}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}}\right)\right\}X_r\right]$$

$$= \left(\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_{S}, 0)\right)E\left[\sigma\left(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}}\right)\mathbf{X}^{\otimes 2}_{S\cup\{r\}}\right]\mathbf{e}_r,$$

where $\tilde{\boldsymbol{\beta}}_{S+r}$ is some point between $\boldsymbol{\beta}_{S+r}$ and $(\boldsymbol{\beta}^{*\mathrm{T}}_{S}, 0)^{\mathrm{T}}$ and $\mathbf{e}_r$ is a vector of length $(|S|+1)$ with the rth element being 1.

Since $|\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}}| \le |\boldsymbol{\beta}^{*\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}}| + |(\boldsymbol{\beta}^{*\mathrm{T}}_{S}, 0)\mathbf{X}_{S\cup\{r\}}| \le 2K^2$ by Conditions (1) and (2),

$$|\sigma(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}_{S+r}\mathbf{X}_{S\cup\{r\}})| \ge \sigma_{\min} \quad \text{and}$$

$$o(n^{-\alpha}) = \left|E\left[\left\{Y - \mu\left(\boldsymbol{\beta}^{*\mathrm{T}}_{S}\mathbf{X}_S\right)\right\}X_r\right]\right| \ge \sigma_{\min}\kappa_{\min}\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_{S}, 0)\|.$$

Therefore,

$$\max_{S:|S|\le q, r\in S^c\cap\mathcal{M}^c} \|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_{S}, 0)\| = o(n^{-\alpha}).$$

Under Ω that is defined in Theorem 3.2.1, $\max_{|S|\le q} \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\| \le A_2K^{-1}(q^2\log p/n)^{1/2}$.

For any $j \in S^c$,

$$\ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \leq \sup_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leq A_2 K^{-1}(q^2 \log p/n)^{1/2}} \ell_{S\cup\{j\}}(\boldsymbol{\beta}^*_{S+j}) - \ell_S(\boldsymbol{\beta}_S)$$

$$\leq \left| n^{-1/2}\mathbb{G}_n \left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}}, Y) \right] \right| + \left| n^{-1/2}\mathbb{G}_n \left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S \mathbf{X}_S, Y) \right] \right|$$

$$+ \sup_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\| \leq A_2 K^{-1}(q^2 \log p/n)^{1/2}} \left| n^{-1/2}\mathbb{G}_n \left[L(\boldsymbol{\beta}^{\mathrm{T}}_S \mathbf{X}_S, Y) - L(\boldsymbol{\beta}^{*\mathrm{T}}_S \mathbf{X}_S, Y) \right] \right|$$

$$+ \left| E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_{S+j}\mathbf{X}_{S\cup\{j\}}, Y) \right] - E\left[L(\boldsymbol{\beta}^{*\mathrm{T}}_S \mathbf{X}_S, Y) \right] \right| \leq$$

$$20(A_1 K^2 + b_{\max})\sqrt{qn^{-1}\log p} + 20A_1 A_2 q^2 n^{-1}\log p +$$

$$\sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}^*_{S+j} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}\|^2/2,$$

where the second inequality follows from the event Ω and Lemma 3.4.5. Since

$$\max_{S:|S|<q, r\in S^c\cap\mathcal{M}^c} \|\boldsymbol{\beta}^*_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}\| = o(n^{-\alpha}) \quad \text{and} \quad qn^{-1+4\alpha}\log p \to 0,$$

we have

$$\max_{S:|S|<q, r\in S^c\cap\mathcal{M}^c} \ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \leq 20(A_1 K^2 + b_{\max})\sqrt{qn^{-1}\log p} +$$

$$20A_1 A_2 q^2 n^{-1}\log p \quad + \sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}^*_{S+j} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}\|^2/2 = o(n^{-2\alpha}),$$

with probability at least $1 - 6\exp(-6q\log p)$. Then by Lemma 3.4.6,

$$\max_{S:|S|<q, r\in S^c\cap\mathcal{M}^c} \ell_{S\cup\{r\}}(\hat{\boldsymbol{\beta}}_{S+r}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \leq \max_{S:|S|<q, r\in S^c\cap\mathcal{M}^c} |\ell_{S\cup\{r\}}(\hat{\boldsymbol{\beta}}_{S+r}) -$$

$$\ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r})| + \max_{S:|S|<q, r\in S^c\cap\mathcal{M}^c} \left| \ell_{S\cup\{r\}}(\boldsymbol{\beta}^*_{S+r}) - \ell_S(\hat{\boldsymbol{\beta}}_S) \right| \leq$$

$$A_3 q^2 n^{-1}\log p + o(n^{-2\alpha}) = o(n^{-2\alpha}), \tag{3.3}$$

with probability at least $1 - 12\exp(-6q\log p)$. By Theorem 3.2.1, if $\mathcal{M} \nsubseteq S$, the forward stage would select a noise variable with probability less than $18\exp(-6q\log p)$.

For $k > |\mathcal{M}|$, $\mathcal{M} \nsubseteq S_k$ implies that at least $k - |\mathcal{M}|$ noise variables are selected within the k steps. Then for $k = C_2|\mathcal{M}|$ with $C_2 > 2$,

$$P(\mathcal{M} \nsubseteq S_k) \leq \sum_{j=k-|\mathcal{M}|}^{k} \binom{k}{j} \left\{18\exp(-6q\log p)\right\}^j \leq |\mathcal{M}|k^{|\mathcal{M}|}\left\{18\exp(-6q\log p)\right\}^{k-|\mathcal{M}|}$$

$$\leq 18\exp(-6q\log p + \log|\mathcal{M}| + |\mathcal{M}|\log k) \leq 18\exp(-4q\log p).$$

Therefore, $\mathcal{M} \subset S_{C_2|\mathcal{M}|}$ with probability at least $1 - 18\exp(-4q\log p)$. □

Proof of Theorem 3.2.3. By Theorem 3.2.2, $\mathcal{M}$ will be included in F_k for some $k < q$ with probability going to 1. Therefore, the forward stage stops at the kth step if $\mathrm{EBIC}(F_{k+1}) > \mathrm{EBIC}(F_k)$. On the other hand, that $\mathrm{EBIC}(F_{k+1}) < \mathrm{EBIC}(F_k)$ if and only if $2\ell_{F_{k+1}}(\hat{\boldsymbol{\beta}}_{F_{k+1}}) - 2\ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}) \geq (\log n + 2\eta_1\log p)/n$. Thus, to show the forward stage stops at the kth step, we only need to show that with probability tending to 1,

$$2\ell_{F_{k+1}}(\hat{\boldsymbol{\beta}}_{F_{k+1}}) - 2\ell_{F_k}(\hat{\boldsymbol{\beta}}_{F_k}) < (\log n + 2\eta_1\log p)/n, \tag{3.4}$$

for all $\eta_1 > 0$.

To prove (3.4), we first verify the conditions (A4) and (A5) in Chen and Chen [115]. Given any index S such that $\mathcal{M} \subseteq S$ and $|S| \leq q$, let $\boldsymbol{\beta}_{*S}$ be the subvector of $\boldsymbol{\beta}_*$ corresponding to S. We obtain that

$$E\left[(Y - \mu(\boldsymbol{\beta}_{*S}^{\mathrm{T}}\mathbf{X}_S))\mathbf{X}_S\right] = E\left[E\left[(Y - \mu(\boldsymbol{\beta}_{*\mathcal{M}}^{T}\mathbf{X}_{\mathcal{M}}))|\mathbf{X}_S\right]\mathbf{X}_S\right] = 0.$$

This implies $\boldsymbol{\beta}_S^* = \boldsymbol{\beta}_{*S}$.

Given any $\boldsymbol{\pi} \in \mathbb{R}^{|S|}$, let $\mathcal{H}_S := \left\{h(\boldsymbol{\pi}, \boldsymbol{\beta}_S) = (\sigma_{\max}K^2|S|)^{-1}\sigma\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2, \|\boldsymbol{\pi}\| = 1, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)\right\}$. By Conditions (1) and (2), $h(\boldsymbol{\pi}, \boldsymbol{\beta}_S)$ is bounded between -1 and 1 uniformly over $\|\boldsymbol{\pi}\| = 1$ and $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$.

By Lemma 2.6.15 in van der Vaart *et al.* [124], the VC indices of $\mathcal{W} := \{(K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S, \|\boldsymbol{\pi}\| = 1\}$ and $\mathcal{V} := \{\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)\}$ are bounded by $|S| + 2$. For the definitions of the VC index

and covering numbers, we refer to pages 83 and 85 in van der Vaart *et al.* [124]. The VC index of the class $\mathcal{U} := \{(K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2, \|\boldsymbol{\pi}\| = 1\}$ is the VC index of the class of sets $\{(\mathbf{X}_S, t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t, \|\boldsymbol{\pi}\| = 1, t \in \mathbb{R}\}$.

Since $\{(\mathbf{X}_S, t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t\} = \{(\mathbf{X}_S, t) : 0 < (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S \leq \sqrt{t}\} \cup \{(\mathbf{X}_S, t) : -\sqrt{t} < (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S \leq 0\}$, each set of $\{(\mathbf{X}_S, t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t, \|\boldsymbol{\pi}\| = 1, t \in \mathbb{R}\}$ is created by taking finite unions, intersections, and complements of of the basic sets $\{(\mathbf{X}_S, t) : (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S < t\}$. Therefore, the VC index of $\{(\mathbf{X}_S, t) : (K^2|S|)^{-1}(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S)^2 \leq t, \|\boldsymbol{\pi}\| = 1, t \in \mathbb{R}\}$ is of the same order as the VC index of $\{(\mathbf{X}_S, t) : (K\sqrt{|S|})^{-1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S < t\}$, by Lemma 2.6.17 in [124].

Then by Theorem 2.6.7 in van der Vaart *et al.* [124], for any probability measure Q, there exists some universal constant C_3 such that $N(\epsilon, \mathcal{U}, L_2(Q)) \leq (C_3/\epsilon)^{2(|S|+1)}$. Likewise, $N(d\epsilon, \mathcal{V}, L_2(Q)) \leq (C_3/\epsilon)^{2(|S|+1)}$. Given a $\boldsymbol{\beta}_{S,0} \in \mathcal{B}_S^0(d)$, for any $\boldsymbol{\beta}_S$ in the ball $\{\boldsymbol{\beta}_S : \sup_{\mathbf{x}} |\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x} - \boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x}| < d\epsilon\}$, we have $\sup_{\mathbf{x}} |\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x}) - \sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})| < Kd\epsilon$ by Condition (4).Let $\mathcal{V}' := \{\sigma_{\max}^{-1}\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S), \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)\}$. By the definition of covering number, $N(Kd\epsilon, \mathcal{V}', L_2(Q)) \leq (C_3/\epsilon)^{2(|S|+1)}$Given a $\sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})$ and $\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{x}$, for any $\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x})$ in the ball $\{\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x}) : \sup_{\mathbf{x}} |\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x}) - \sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})| \leq Kd\epsilon\}$ and $\boldsymbol{\pi}$ in the ball $\{\boldsymbol{\pi} : \sup_{\mathbf{x}} |(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{x})^2 - (\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{x})^2| < \epsilon\}$, $(\sigma_{\max}K^2|S|)^{-1} \sup_{\mathbf{x}} |\sigma(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{x})(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{x})^2 - \sigma(\boldsymbol{\beta}_{S,0}^{\mathrm{T}}\mathbf{x})(\boldsymbol{\pi}_0^{\mathrm{T}}\mathbf{x})^2| \leq (\sigma_{\max}^{-1}Kd + (K^2|S|)^{-1})\epsilon$. Thus, $N((\sigma_{\max}^{-1}Kd + (K^2|S|)^{-1})\epsilon, \mathcal{H}_S, L_2(Q)) \leq (C_3/\epsilon)^{4(|S|+1)}$, and consequently $N(\epsilon, \mathcal{H}_S, L_2(Q)) \leq (C_4/\epsilon)^{4(|S|+1)}$ for some constant C_4.

By Theorem 1.1 in Talagrand [125] and $|S| \leq q$, we can find some constant C_5 such that

$$P\left(\sup_{\|\boldsymbol{\pi}\|=1, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)} |\mathbb{G}_n\left[h(\boldsymbol{\pi}, \boldsymbol{\beta}_S)\right]| \geq C_5\sqrt{q\log p}\right)$$

$$\leq \frac{C_4'}{C_5\sqrt{q\log p}}\left(\frac{C_4'C_5^2 q\log p}{4(|S|+1)}\right)^{4(|S|+1)} \exp(-2C_5^2 q\log p)$$

$$\leq \exp\left(4(|S|+1)\log(C_4'C_5^2 q\log p) - 2C_5^2 q\log p\right)$$

$$\leq \exp\left(-5q\log p\right),$$

where C_4' is some constant that depends on C_4 only. Thus,

$$P\Big(\sup_{|S|\le q, \|\boldsymbol{\pi}\|=1, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)} \Big| \mathbb{E}_n \left\{ \sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2 \right\} - E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2 \right] \Big|$$

$$\ge C_5 K^2 \sqrt{q^3 \log p/n} \Big) \sum_{s=|\mathcal{M}|}^{q} \left(\tfrac{ep}{s}\right)^s \exp\left(-5q\log p\right) \le \exp(-3q\log p). \tag{3.5}$$

By Condition (5), $\sigma_{\min}\kappa_{\min} \le \Lambda\left(E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\mathbf{X}_S^{\otimes 2}\right]\right) \le \sigma_{\max}\kappa_{\max}$, for all $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$ and $S : \mathcal{M} \subseteq S, |S| < q$. Then, by (3.5),

$$\sigma_{\min}\kappa_{\min}/2 \le \Lambda\left(\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\mathbf{X}_S^{\otimes 2}\right\}\right) \le 2\sigma_{\max}\kappa_{\max}$$

uniformly over all S satisfying $\mathcal{M} \subseteq S$ and $|S| \le q$, with probability at least $1 - \exp(-3q\log p)$. Hence, the condition (A4) in [115] is satisfied with probability at least $1 - \exp(-3q\log p)$.

Also for any $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$,

$$\left|\mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\} - \mathbb{E}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right|$$

$$\le \left|n^{-1/2}\mathbb{G}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right| + \left|n^{-1/2}\mathbb{G}_n\left\{\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right\}\right|$$

$$+ \left|E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right] - E\left[\sigma\left(\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_{*S}\right)\left(\boldsymbol{\pi}^{\mathrm{T}}\mathbf{X}_S\right)^2\right]\right|$$

$$\le 2C_5K^2\sqrt{q^3\log p/n} + \mu_{\max}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{*S}\|\sqrt{|S|}K\lambda_{\max}.$$

Hence, the condition (A5) in Chen and Chen [115] is satisfied uniformly over all S such that $\mathcal{M} \subseteq S$ and $|S| \le q$, with probability at least $1 - \exp(-3q\log p)$.

Then (3.4) can be shown by following the proof of Equation (3.2) in Chen and Chen [115]. Thus, our forward stage stops at the kth step with probability at least $1 - \exp(-3q\log p)$. □

Proof of Theorem 3.2.4. Suppose that a covariate X_r is removed from S. For any $r \in \mathcal{M}$, since $\mathcal{M} \not\subseteq S\backslash\{r\}$ and r is the only element that is in $(S\backslash\{r\})^c \cap \mathcal{M}$, by Lemma 3.4.1 and (3.2)

$$\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_{S\backslash\{r\}}(\hat{\boldsymbol{\beta}}_{S\backslash\{r\}}) \ge \ell_S(\boldsymbol{\beta}_S^*) - \ell_{S\backslash\{r\}}(\hat{\boldsymbol{\beta}}_{S\backslash\{r\}})$$

$$= \ell_{S\backslash\{r\}\cup\{r\}}(\boldsymbol{\beta}^*_{S\backslash\{r\}+r}) - \ell_{S\backslash\{r\}}(\hat{\boldsymbol{\beta}}_{S\backslash\{r\}}) \ge C_1 n^{-2\alpha},$$

with probability at least $1 - 6\exp(-6q\log p)$. From the proof of Theorem 3.2.1, we have for any $\eta_2 > 0$, $\mathrm{BIC}(S) - \mathrm{BIC}(S\backslash\{r\}) \leq -2C_1 n^{-2\alpha} + \eta_2 n^{-1}\log n < 0$, uniformly over $r \in \mathcal{M}$ and S satisfying $\mathcal{M} \subset S$ and $|S| \leq q$, with probability at least $1 - 6\exp(-6q\log p)$. □

Proof of Theorem 3.2.5. By Theorems 3.2.1, 3.2.2, and 3.2.3, we have that the event $\Omega_1 := \{|\hat{\mathcal{M}}| \leq q \text{ and } \mathcal{M} \subseteq \hat{\mathcal{M}}\}$ holds with probability at least $1 - 25\exp(-2q\log p)$. Thus, in the rest of the proof, we restrict our attention on Ω_1.

As shown in the proof of Theorem 3.2.3, we obtain that $\boldsymbol{\beta}^*_{\hat{\mathcal{M}}} = \boldsymbol{\beta}_{*\hat{\mathcal{M}}}$. Then by Lemma 3.4.6, we have $\|\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}} - \boldsymbol{\beta}^*_{\hat{\mathcal{M}}}\| \leq A_2 K^{-1}\sqrt{q^2\log p/n}$ with probability at least $1 - 6\exp(-6q\log p)$. Withdrawing the attention on Ω_1, we obtain that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\| = \|\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}} - \boldsymbol{\beta}_{*\hat{\mathcal{M}}}\| = \|\hat{\boldsymbol{\beta}}_{\hat{\mathcal{M}}} - \boldsymbol{\beta}^*_{\hat{\mathcal{M}}}\| \leq A_2 K^{-1}\sqrt{q^2\log p/n},$$

with probability at least $1 - 31\exp(-2q\log p)$. □

3.4 Additional Lemmas

Lemma 3.4.1. *Given a model S such that $|S| < q$, $\mathcal{M} \nsubseteq S$, under Condition (6),*

(i): $\exists r \in S^c \cap \mathcal{M}$, such that $\boldsymbol{\beta}^*_{S+r} \neq (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}$.

(ii): Suppose Conditions (1), (2), and (6') hold. $\exists r \in S^c \cap \mathcal{M}$, such that $\|\boldsymbol{\beta}^{*\mathrm{T}}_{S+r} - (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)\| \geq C\sigma^{-1}_{\max}\kappa^{-1}_{\max}n^{-\alpha}$.

Proof. As $\boldsymbol{\beta}^*_{S+j}$ is the maximizer of $E\left[\ell_{S\cup\{j\}}(\boldsymbol{\beta}_{S+j})\right]$, by the concavity of $E\left[\ell_{S\cup\{j\}}(\boldsymbol{\beta}_{S+j})\right]$, $\boldsymbol{\beta}^*_{S+j}$ is the solution to the equation $E\left[\left(Y - \mu\left(\boldsymbol{\beta}^{*T}_S\mathbf{X}_S + \beta_j X_j\right)\right)\mathbf{X}_{S\cup\{j\}}\right] = \mathbf{0}$,

(i): Suppose that $\boldsymbol{\beta}^*_{S+j} = (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}, \forall j \in S^c \cap \mathcal{M}$. Then,

$$\begin{aligned} 0 &= E\left[\left(Y - \mu\left(\boldsymbol{\beta}^{*\mathrm{T}}_S\mathbf{X}_S\right)\right)X_j\right] = E\left[\left(\mu\left(\boldsymbol{\beta}^{\mathrm{T}}_*\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}^{*T}_S\mathbf{X}_S\right)\right)X_j\right] \\ &\Rightarrow \max_{j\in S^c\cap\mathcal{M}}\left|E\left[\left(\mu\left(\boldsymbol{\beta}^{\mathrm{T}}_*\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}^{*T}_S\mathbf{X}_S\right)\right)X_j\right]\right| = 0, \end{aligned}$$

which violates the Condition (6). Therefore, we can find a $r \in S^c \cap \mathcal{M}$, such that $\boldsymbol{\beta}^*_{S+r} \neq (\boldsymbol{\beta}^{*\mathrm{T}}_S, 0)^{\mathrm{T}}$.

(ii): Let $r \in S^c \cap \mathcal{M}$ satisfy that $\left|E\left[\left(\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right)X_r\right]\right| > Cn^{-\alpha}$. Without loss of generality, we assume that X_r is the last element of $\mathbf{X}_{S\cup\{r\}}$. By the mean value theorem,

$$E\left[\left(\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right)X_r\right]$$

$$= E\left[\left(\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right)X_r\right] - E\left[\left(\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_{S+r}^{*\mathrm{T}}\mathbf{X}_{S\cup\{r\}}\right)\right)X_r\right]$$

$$= E\left[\left(\mu\left(\boldsymbol{\beta}_{S+r}^{*\mathrm{T}}\mathbf{X}_{S\cup\{r\}}\right) - \mu\left((\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)\mathbf{X}_{S\cup\{r\}}\right)\right)X_r\right]$$

$$= \left(\boldsymbol{\beta}_{S+r}^{*\mathrm{T}} - (\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)\right)E\left[\sigma\left(\tilde{\boldsymbol{\beta}}_{S+r}^{\mathrm{T}}\mathbf{X}_{S\cup\{r\}}\right)\mathbf{X}_{S\cup\{r\}}^{\otimes 2}\right]\mathbf{e}_r, \tag{3.1}$$

where $\tilde{\boldsymbol{\beta}}_{S+r}$ is some point between $\boldsymbol{\beta}_{S+r}^*$ and $(\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)^{\mathrm{T}}$ and $\mathbf{e}_r$ is a vector of length $(|S|+1)$ with the rth element being 1.

As $\tilde{\boldsymbol{\beta}}_{S+r}$ is some point between $\boldsymbol{\beta}_{S+r}^*$ and $(\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)^{\mathrm{T}}$,

$$|\tilde{\boldsymbol{\beta}}_{S+r}^{\mathrm{T}}\mathbf{X}_{S\cup\{r\}}| \le |\boldsymbol{\beta}_{S+r}^{*\mathrm{T}}\mathbf{X}_{S\cup\{r\}}| + |(\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)\mathbf{X}_{S\cup\{r\}}| \le 2K^2,$$

by Conditions (1) and (2). Thus, $|\sigma(\tilde{\boldsymbol{\beta}}_{S+r}^{\mathrm{T}}\mathbf{X}_{S\cup\{r\}})| \le \sigma_{\max}$. By (3.1) and Condition (5),

$$Cn^{-\alpha} \le \left|E\left[\left(\mu\left(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}\right) - \mu\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right)X_r\right]\right|$$

$$\le \|\boldsymbol{\beta}_{S+r}^{*\mathrm{T}} - (\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)\|\sigma_{\max}\lambda_{\max}\left(E\left[\mathbf{X}_{S\cup\{r\}}^{\otimes 2}\right]\right)\|\mathbf{e}_r\|$$

$$\le \sigma_{\max}\kappa_{\max}\|\boldsymbol{\beta}_{S+r}^{*\mathrm{T}} - (\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)\|.$$

Therefore, $\|\boldsymbol{\beta}_{S+r}^{*\mathrm{T}} - (\boldsymbol{\beta}_S^{*\mathrm{T}}, 0)\| \ge C\sigma_{\max}^{-1}\kappa_{\max}^{-1}n^{-\alpha}$. □

Lemma 3.4.2. *Let $\xi_i, i = 1, \ldots, n$ be n i.i.d random variables such that $|\xi_i| \le B$ for a constant $B > 0$. Under Conditions (1) – (3), we have $E\left[|Y_i\xi_i - E\left[Y_i\xi_i\right]|^m\right] \le m!(2B(\sqrt{2}M + \mu_{\max}))^m$, for every $m \ge 1$.*

Proof. By Conditions (1) and (2), $|\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i| \le KL, \forall i \ge 1$ and consequently $\left|\mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)\right| \le \mu_{\max}$. Then

by Condition (3),

$$E\left[|Y_i|^m\right] = E\left[|\epsilon_i + \mu(\boldsymbol{\beta}_*^{\mathrm{T}}\mathbf{X}_i)|^m\right] \leq \sum_{t=0}^{m} \binom{m}{t} E\left[|\epsilon_i|^t\right] \mu_{\max}^{m-t}$$

$$\leq \sum_{t=0}^{m} t! \binom{m}{t} M^t \mu_{\max}^{m-t} \leq m!(M + \mu_{\max})^m,$$

for every $m \geq 1$. By the same arguments, it can be shown that, for every $m \geq 1$,

$$E\left[|Y_i\xi_i - E\left[Y_i\xi_i\right]|^m\right] \leq E\left[\left(|Y_i\xi_i| + |E\left[Y_i\xi_i\right]|\right)^m\right] \leq m!(2B(M + \mu_{\max}))^m.$$

□

Lemma 3.4.3. *Under Conditions (1) – (3), when n is sufficiently large such that* $28\sqrt{q \log p/n} < 1$, *we have* $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\mathbb{E}_n\left\{L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\right\}\right| \leq 2(M+\mu_{\max})K^3 + b_{\max}$, *with probability* $1 - 2\exp(-10q \log p)$.

Proof. By Conditions (2), $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right| \leq K^3$. Thus,

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\mathbb{E}_n\left\{L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\right\}\right| \leq \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left|\mathbb{E}_n\left\{\left|Y\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right|\right\}\right| + b_{\max}$$

$$\leq \left(\left|\mathbb{E}_n\left\{|Y| - E\left[|Y|\right]\right\}\right| + E\left[|Y|\right]\right) K^3 + b_{\max}$$

$$\leq \left(\left|\mathbb{E}_n\left\{|Y| - E\left[|Y|\right]\right\}\right|\right) K^3 + (M + \mu_{\max})K^3 + b_{\max},$$

where the last inequality follows from that $E[|Y|] \leq M + \mu_{\max}$ as shown in the proof of Lemma 3.4.2.

Let $\xi_i = 1\{Y_i > 0\} - 1\{Y_i < 0\}$. Thus $|\xi_i| \leq 1$. By Lemma 3.4.2, we have $E\left[\left||Y_i| - E\left[|Y_i|\right]\right|^m\right] \leq m!(2(M + \mu_{\max}))^m$. Applying Bernstein's inequality (e.g., Lemma 2.2.11 in van der Vaart *et al.* [124]) yields that

$$P\left(\left|\mathbb{E}_n\left\{|Y| - E\left[|Y|\right]\right\}\right| > 10(M + \mu_{\max})\sqrt{q \log p/n}\right)$$

$$\leq 2\exp\left(-\frac{1}{2}\frac{196q \log p}{4+20\sqrt{q \log p/n}}\right) \leq 2\exp(-10q \log p), \tag{3.2}$$

when n is sufficiently large such that $20\sqrt{q\log p/n} < 1$. Since $10(M+\mu_{\max})\sqrt{q\log p/n} = o(1)$, then

$$P\left(\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\mathbb{E}_n\left\{L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},Y)\right\}\right| \geq 2(M+\mu_{\max})K^3 + b_{\max}\right) \leq 2\exp(-10q\log p).$$

□

Lemma 3.4.4. *Given an index set S and $r \in S^c$, let $\mathcal{B}_S^0(d) = \{\boldsymbol{\beta}_S : \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leq d/(K\sqrt{|S|})\}$ and $A_1 := (M + 2\mu_{\max})$. Under Conditions (1) – (3), when n is sufficiently large such that $10\sqrt{q\log p/n} < 1$, we have*

1. $|\mathbb{G}_n\left[L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]| \leq 20A_1 d\sqrt{q\log p}$, *uniformly over $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$ and $|S| \leq q$, with probability at least $1 - 4\exp(-6q\log p)$.*

2. $|\mathbb{G}_n\left[L(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right]| \leq 10(A_1K^2 + b_{\max})\sqrt{q\log p}$, *uniformly over $|S| \leq q$, with probability at least $1 - 2\exp(-8q\log p)$.*

Proof. : (1): Let $\mathcal{R}_{|S|}(d)$ be a $|S|$-dimensional ball with center at 0 and radius $d/(K\sqrt{|S|})$. Then $\mathcal{B}_S^0(d) = \mathcal{R}_{|S|}(d) + \boldsymbol{\beta}_S^*$. Let $C_{|S|} := \{C(\boldsymbol{\xi}_k)\}$ be a collection of cubes that cover the ball $R_{|S|}(d)$, where $C(\boldsymbol{\xi}_k)$ is a cube containing $\boldsymbol{\xi}_k$ with sides of length $d/(K\sqrt{|S|}n^2)$ and $\boldsymbol{\xi}_k$ is some point in $\mathcal{R}_{|S|}(d)$. As the volume of $C(\boldsymbol{\xi}_k)$ is $\left(d/(K\sqrt{|S|}n^2)\right)^{|S|}$ and the volume of $\mathcal{R}_{|S|}(d)$ is less than $(2d/(K\sqrt{|S|}))^{|S|}$, we can select $\boldsymbol{\xi}_k$'s so that no more than $(4n^2)^{|S|}$ cubes are needed to cover $\mathcal{R}_{|S|}(d)$. We thus assume $|C_{|S|}| \leq (4n^2)^{|S|}$. For any $\boldsymbol{\xi} \in C(\boldsymbol{\xi}_k)$, $\|\boldsymbol{\xi} - \boldsymbol{\xi}_k\| \leq d/(Kn^2)$. In addition, let $T_{1S}(\boldsymbol{\xi}) := \mathbb{E}_n\left[Y\boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}_S\right]$, $T_{2S}(\boldsymbol{\xi}) := \mathbb{E}_n\left[b\left(\left(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}\right)^{\mathrm{T}}\mathbf{X}_S\right) - b\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S\right)\right]$, and $T_S(\boldsymbol{\xi}) := T_{1S}(\boldsymbol{\xi}) - T_{2S}(\boldsymbol{\xi})$.
Given any $\boldsymbol{\xi} \in \mathcal{R}_{|S|}(d)$, there exists $C(\boldsymbol{\xi}_k) \in C_{|S|}$ such that $\boldsymbol{\xi} \in C(\boldsymbol{\xi}_k)$. Then

$$\begin{aligned}|T_S(\boldsymbol{\xi}) - E\left[T_S(\boldsymbol{\xi})\right]| &\leq |T_S(\boldsymbol{\xi}) - T_S(\boldsymbol{\xi}_k)|\,|T_S(\boldsymbol{\xi}_k) - E\left[T_S(\boldsymbol{\xi}_k)\right]| + |E\left[T_S(\boldsymbol{\xi})\right] - E\left[T_S(\boldsymbol{\xi}_k)\right]| \\ &=: I + II + III.\end{aligned}$$

We deal with III first. By the mean value theorem, there exists a $\tilde{\boldsymbol{\xi}}$ between $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_k$ such that

$$|E\left[T_S(\boldsymbol{\xi}_k)\right] - E\left[T_S(\boldsymbol{\xi})\right]| = \left|E\left[Y(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right] + E\left[\mu\left(\left(\boldsymbol{\beta}_S^* + \tilde{\boldsymbol{\xi}}\right)^{\mathrm{T}}\mathbf{X}_S\right)(\boldsymbol{\xi}_k - \boldsymbol{\xi})^{\mathrm{T}}\mathbf{X}_S\right]\right|$$

$$\leq E[|Y|]\|\boldsymbol{\xi}_k - \boldsymbol{\xi}\|\|\mathbf{X}_S\| + \mu_{\max}\|\boldsymbol{\xi}_k - \boldsymbol{\xi}\|\|\mathbf{X}_S\| \leq (M + 2\mu_{\max})d\sqrt{|S|}n^{-2} = A_1 d\sqrt{|S|}n^{-2} \quad (3.3)$$

where the last inequality follows from Lemma 3.4.2 and $A_1 = M + 2\mu_{\max}$.

Next, we evaluate II. By Condition (2), $|\mathbf{X}_{iS}^{\mathrm{T}}\boldsymbol{\xi}| \le \|\mathbf{X}_{iS}\|\|\boldsymbol{\xi}\| \le d/(K\sqrt{|S|})\sqrt{|S|}K = d$, for all $\boldsymbol{\xi} \in \mathcal{R}_{|S|}(d)$. Then by Lemma 3.4.2,

$$E\left[\left|Y\boldsymbol{\xi}_k^{\mathrm{T}}\mathbf{X}_S - E\left[Y\boldsymbol{\xi}_k^{\mathrm{T}}\mathbf{X}_S\right]\right|^m\right] \le m!(2(M+\mu_{\max})d)^m.$$

By Bernstein's inequality, when n is sufficiently large such that $10\sqrt{q\log p/n} \le 1$.

$$\begin{aligned} &P\left(|T_{1S}(\boldsymbol{\xi}_k) - E\left[T_{1S}(\boldsymbol{\xi}_k)\right]| > 10(M+\mu_{\max})d\sqrt{qn^{-1}\log p}\right) \\ &\le 2\exp\left(-\frac{1}{2}\frac{100q\log p}{4+20\sqrt{q\log p/n}}\right) \le 2\exp(-10q\log p). \end{aligned} \tag{3.4}$$

Since $|b(\left(\boldsymbol{\beta}_S^* + \boldsymbol{\xi}_k\right)^{\mathrm{T}}\mathbf{X}_S) - b(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S)| \le \mu_{\max}d$, by the same arguments used for (3.4), we have

$$P\left(|T_{2S}(\boldsymbol{\xi}_k) - E\left[T_{2S}(\boldsymbol{\xi}_k)\right]| > 10\mu_{\max}d\sqrt{qn^{-1}\log p}\right) \le 2\exp(-10q\log p). \tag{3.5}$$

Combining (3.4) and (3.5) yields that uniformly over $\boldsymbol{\xi}_k$

$$|T_S(\boldsymbol{\xi}_k) - E\left[T_S(\boldsymbol{\xi}_k)\right]| \le 10A_1d\sqrt{qn^{-1}\log p}, \tag{3.6}$$

with probability at least $1 - 2(4n^2)^{|S|}\exp(-10q\log p)$.

We now assess I. Following the same arguments as in Lemma 3.4.3,

$$P\Big(\sup_{\boldsymbol{\xi}\in C(\boldsymbol{\xi}_k)}|T_S(\boldsymbol{\xi}) - T_S(\boldsymbol{\xi}_k)| > (2M+3\mu_{\max})d\sqrt{|S|}n^{-2}\Big) \le 2\exp(-8q\log p). \tag{3.7}$$

Since $\sqrt{|S|}n^{-2} = o(\sqrt{qn^{-1}\log p})$, combining (3.3), (3.6), and (3.7) together yields that

$$\begin{aligned} &P\Big(\sup_{\boldsymbol{\xi}\in\mathcal{R}_{|S|}(d)}|T_S(\boldsymbol{\xi}) - E\left[T_S(\boldsymbol{\xi})\right]| \ge 20A_1d\sqrt{qn^{-1}\log p}\Big) \\ &\le 2(4n^2)^{|S|}\exp(-10q\log p) + 2\exp(-8q\log p) \le 4\exp(-8q\log p). \end{aligned}$$

By the combinatoric inequality $\binom{p}{s} \le (ep/s)^s$, we obtain that

$$\begin{aligned} &P\Big(\sup_{|S|\le q,\boldsymbol{\beta}_S\in\mathcal{B}_S^0(d_1)}\left|\mathbb{G}_n\left[L\left(\boldsymbol{\beta}_S^{\mathrm{T}}\mathbf{X}_S, Y\right) - L\left(\boldsymbol{\beta}_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]\right| \ge 20A_1d\sqrt{q\log p}\Big) \\ &\le \sum_{s=1}^{q}(ep/s)^s 4\exp(-8q\log p) \le 4\exp(-6q\log p). \end{aligned}$$

(2): We evaluate the mth moment of $L(\beta_S^* \mathbf{X}_S, Y)$.

$$E\left[\left(Y\beta_S^*\mathbf{X}_S - b(\beta_S^*\mathbf{X}_S)\right)^m\right] \le E\left[\sum_{t=0}^{m}\binom{m}{t}|Y|^t K^{2t} b_{\max}^{m-t}\right]$$
$$\le \sum_{t=0}^{m}\binom{m}{t} t!\left((M+\mu_{\max})K^2\right)^t b_{\max}^{m-t} \le m!((M+\mu_{\max})K^2 + b_{\max})^m.$$

Then, by Bernstein's inequality,

$$P\left(|\mathbb{G}_n\left[L(\beta_S^{*\mathrm{T}}\mathbf{X}_S, Y)\right]| > 10(A_1K^2 + b_{\max})\sqrt{q\log p}\right) \le 2\exp(-10q\log p).$$

By the same arguments used in (i), we obtain that

$$P\Big(\sup_{|S|\le q}\left|\mathbb{G}_n\left[L\left(\beta_S^{*\mathrm{T}}\mathbf{X}_S, Y\right)\right]\right| \ge 10(A_1K^2 + b_{\max})\sqrt{q\log p}\Big)$$
$$\le \sum_{s=1}^{q}(ep/s)^s 2\exp(-10q\log p) \le 2\exp(-8q\log p).$$

□

Lemma 3.4.5. *Given a model S and $r \in S^c$, under Conditions (1), (2), and (5), for any $\|\beta_S - \beta_S^*\| \le K/\sqrt{|S|}$,*

$$\sigma_{\min}\kappa_{\min}\|\beta_S - \beta_S^*\|^2/2 \le E\left[\ell_S(\beta_S^*)\right] - E\left[\ell_S(\beta_S)\right] \le \sigma_{\max}\kappa_{\max}\|\beta_S - \beta_S^*\|^2/2.$$

Proof. Due to the concavity of the log-likelihood in GLMs with the canonical link, $E\left[Y\mathbf{X}_S - \mu(\beta_S^{*\mathrm{T}}\mathbf{X}_S)\mathbf{X}_S\right] = \mathbf{0}$. Then for any $\|\beta_S - \beta_S^*\| \le K/\sqrt{|S|}$,

$$|\beta^{\mathrm{T}}\mathbf{X}_S| \le |\beta^{*\mathrm{T}}\mathbf{X}_S| + |(\beta - \beta^*)^{\mathrm{T}}\mathbf{X}_S| \le K^2 + K/\sqrt{|S|} \times K\sqrt{|S|} = 2KL.$$

Thus, by Taylor's expansion,

$$E\left[\ell_S(\beta_S)\right] - E\left[\ell_S(\beta_S^*)\right] = -\frac{1}{2}(\beta_S - \beta_S^*)^{\mathrm{T}} E\left[\sigma\left(\tilde{\beta}_S^{\mathrm{T}}\mathbf{X}_S\right)\mathbf{X}_S^{\otimes 2}\right](\beta_S - \beta_S^*),$$

where $\tilde{\beta}_S$ is between β_S and β_S^*. By Condition (5),

$$\sigma_{\min}\kappa_{\min}\|\beta_S - \beta_S^*\|^2/2 \le E\left[\ell_S(\beta_S^*)\right] - E\left[\ell_S(\beta_S)\right] \le \sigma_{\max}\kappa_{\max}\|\beta_S - \beta_S^*\|^2/2.$$

□

Lemma 3.4.6. *Under Conditions (1) – (6), there exist some constants A_2 and A_3 that do not depend on n, such that $\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\| \leq A_2 K^{-1}\sqrt{q^2 \log p/n}$ and $|\ell_S(\hat{\boldsymbol{\beta}}_S) - \ell_S(\boldsymbol{\beta}_S^*)| \leq A_3 q^2 \log p/n$ hold uniformly over $S : |S| \leq q$, with probability at least $1 - 6\exp(-6q \log p)$.*

Proof. Define

$$\Omega(d) := \Big\{ \sup_{|S|\leq q, \boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)} \Big| \mathbb{G}_n \Big[L\left(\boldsymbol{\beta}_S^{\mathrm{T}} \mathbf{X}_S, Y\right) - L\left(\boldsymbol{\beta}_S^{*\mathrm{T}} \mathbf{X}_S, Y\right) \Big] \Big| < 20 A_1 d\sqrt{q \log p} \Big\}.$$

By Lemma 3.4.4, the event $\Omega(d)$ holds with probability at least $1 - 4\exp(-6q \log p)$. Thus, in the proof of Lemma 3.4.6, we shall assume $\Omega(d)$ hold with $d = A_2\sqrt{q^3 \log p/n}$ for some $A_2 > 20(\sigma_{\min}\kappa_{\min})^{-1} K^2 A_1$.

For any $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| = A_2 K^{-1}\sqrt{q^2 \log p/n}$, since $\sqrt{q^2 \log p/n} \leq \sqrt{q^3 \log p/n}/\sqrt{|S|}$, $\boldsymbol{\beta}_S \in \mathcal{B}_S^0(d)$. By Lemma 3.4.5,

$$\begin{aligned}
&\ell_S(\boldsymbol{\beta}_S^*) - \ell_S(\boldsymbol{\beta}_S) \\
&= \Big(\ell_S(\boldsymbol{\beta}_S^*) - E\left[\ell_S(\boldsymbol{\beta}_S^*)\right] - \left(\ell_S(\boldsymbol{\beta}_S) - E\left[\ell_S(\boldsymbol{\beta}_S)\right]\right) \Big) + \left(E\left[\ell_S(\boldsymbol{\beta}_S^*)\right] - E\left[\ell_S(\boldsymbol{\beta}_S)\right] \right) \\
&\geq \sigma_{\min}\kappa_{\min} \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2/2 - 20 A_1 d\sqrt{q \log p/n} \\
&= \sigma_{\min}\kappa_{\min} A_2^2 q^2 \log p/(K^2 n) - 20 A_1 A_2 q^2 \log p/n > 0.
\end{aligned}$$

Thus,

$$\inf_{|S|\leq q, \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| = A_2 K^{-1}\sqrt{q^2 \log p/n}} \ell_S(\boldsymbol{\beta}_S^*) - \ell_S(\boldsymbol{\beta}_S) > 0.$$

Then by the concavity of $\ell_S(\cdot)$, we obtain that $\max_{|S|\leq q} \left\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*\right\| \leq A_2 K^{-1}\sqrt{q^2 n^{-1} \log p}$.

On the other hand, for any $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\| \leq A_2 K^{-1}\sqrt{q^2 \log p/n}$,

$$\begin{aligned}
&\left|\ell_S(\boldsymbol{\beta}_S^*) - \ell_S(\boldsymbol{\beta}_S)\right| \\
&\leq \left| \ell_S(\boldsymbol{\beta}_S^*) - E\left[\ell_S(\boldsymbol{\beta}_S^*)\right] - \left(\ell_S(\boldsymbol{\beta}_S) - E\left[\ell_S(\boldsymbol{\beta}_S)\right]\right) \right| + \left| E\left[\ell_S(\boldsymbol{\beta}_S^*)\right] - E\left[\ell_S(\boldsymbol{\beta}_S)\right] \right| \\
&\leq \sigma_{\max}\kappa_{\max} \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_S^*\|^2/2 + 20 A_1 d\sqrt{q \log p/n} \leq A_3 q^2 n^{-1} \log p,
\end{aligned}$$

where $A_3 := 4\sigma_{\max}\lambda_{\max}A_2^2K^{-2} + 20A_1A_2$. □

3.5 Simulations

We compared the proposal with the other competing methods, including the penalized methods such as least absolute shrinkage and selection operator (LASSO), the differential geometric least angle regression (dgLARS) [90, 91], the forward regression (FR) approach [93], the sequentially conditioning (SC) approach [99], and the screening method such as sure independence screening (SIS) [87], which is popular in practice. As SIS does not directly generate a predictive model, we applied LASSO for the top $[n/\log(n)]$ variables chosen by SIS and denoted the procedure by SIS+LASSO.

As the FR, SC and STEPWISE approaches involve forward searching and to make them comparable, we applied the same stopping rule, for example, Equation (3.3) with the same γ, to their forward steps. In particular, the STEPWISE approach, with $\eta_1 = \gamma$ and $\eta_2 = 0$, is equivalent to FR and asymptotically equivalent to SC. By varying γ in FR and SC between γ_L and γ_H, we explored the impact of γ on inducing false positives and negatives. In our numerical studies, we fixed $\gamma_H = 10$ and set $\gamma_L = \eta_1$.

To choose η_1 and η_2 in (3.3) and (3.4) in STEPWISE, we performed 5-fold cross-validation to minimize the mean squared prediction error (MSPE), and reported the results in Table 3.1. Since the proposed STEPWISE algorithm uses the (E)BIC criterion, for a fair comparison we chose the tuning parameter in dgLARS by using the BIC criterion as well, and coined the corresponding approach as dgLARS(BIC).

The regularization parameter in LASSO was chosen via the following two approaches: 1) giving the smallest BIC for the models on the LASSO path, denoted by LASSO(BIC); 2) using the one-standard-error rule, denoted by LASSO(1SE), which chooses the most parsimonious model whose error is no more than one standard error above the error of the best model in cross-validation [126].

Table 3.1: The values of η_1 and η_2 used in the simulation studies

	Normal Model	Binomial Model	Poisson Model
Example 1	(0.5, 3)	(0.5, 3)	(1, 3)
Example 2	(0.5, 3)	(1, 3)	(1, 3)
Example 3	(1, 3)	(0.5, 3)	(0.5, 1)
Example 4	(1, 3.5)	(0, 1)	(1, 3)
Example 5	(0.5, 3)	(0.5, 2)	(0.5, 3)
Example 6	(0.5, 3)	(0.5, 3)	(1, 3)
Example 7	(0.5, 3)	(0.5, 3)	(0.5, 4.5)

Note: Values for η_1 and η_2 were searched on the grid $\{0, 0.25, 0.5, 1\}$ and $\{1, 2, 3, 3.5, 4, 4.5, 5\}$, respectively.

Denote by $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, the covariate vector for subject i $(1, \ldots, n)$ and the true coefficient vector. The following five examples generated $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}$, the inner product of the coefficient and covariate vectors for each individual, which were used to generate outcomes from the Normal, Binomial, and Poisson models.

Example 1: For each i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{p_0} \beta_j X_{ij} + \sum_{j=p_0+1}^{p} \beta_j X_{ij} \right), \quad i = 1, \ldots, n,$$

where $\beta_j = (-1)^{B_j}(4\log n/\sqrt{n} + |Z_j|)$, for $j = 1, \ldots, p_0$ and β_j =0, otherwise; B_j was a binary random variable with $P(B_j = 1) = 0.4$ and Z_j was generated by a standard normal distribution; p_0= 8; X_{ij}'s were independently generated from a standardized exponential distribution, that is, $\exp(1) - 1$. Here and also in the other examples, c (specified later) controls the signal strengths.

Example 2: This scenario is the same as **Example 1** except that X_{ij} was independently generated from a standard normal distribution.

Example 3: For each i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{p_0} \beta_j X_{ij} + \sum_{j=p_0+1}^{p} \beta_j X_{ij}^* \right), \quad i = 1, \ldots, n,$$

where $\beta_j = 2j$ for $1 \le j \le p_0$ and p_0= 5. We simulated every component of $\mathbf{Z}_i = (Z_{ij}) \in R^p$ and $\mathbf{W}_i = (W_{ij}) \in R^p$ independently from a standard normal distribution. Next, we generated $\mathbf{X}_i$ according to $X_{ij} = (Z_{ij} + W_{ij})/\sqrt{2}$ for $1 \le j \le p_0$ and $X_{ij}^* = (Z_{ij} + \sum_{j'=1}^{p_0} Z_{ij'})/2$ for $p_0 < j \le p$.

Example 4: For each i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{500} \beta_j X_{ij} + \sum_{j=501}^{p} \beta_j X_{ij} \right), \quad i = 1, \ldots, n,$$

where the first 500 X_{ij}'s were generated from the multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix with all of the diagonal entries being 1 and $\mathrm{cov}(X_{ij}, X_{ij'}) = 0.5^{|j-j'|}$ for $1 \leq j, j' \leq p$. The remaining $p - 500$ X_{ij}'s were generated through the autoregressive processes with $X_{i,501} \sim Unif(\text{-}2, 2)$, $X_{ij} = 0.5\ X_{i,j-1} + 0.5\ X_{ij}^*$, for $j = 502, \ldots, p$, where $X_{ij}^* \sim Unif(\text{-}2, 2)$ were generated independently. The coefficients β_j for $j = 1, \ldots, 7, 501, \ldots, 507$ were generated from $(-1)^{B_j}(4\log n/\sqrt{n} + |Z_j|)$, where B_j was a binary random variable with $P(B_j = 1) = 0.4$ and Z_j was from a standard normal distribution. The remaining β_j were zeros.

Example 5: For each i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(-0.5X_{i1} + X_{i2} + 0.5X_{i,100}\right), \quad i = 1, \ldots, n,$$

where $\mathbf{X}_i$ were generated from a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix with all of the diagonal entries being 1 and $\mathrm{cov}(X_{ij}, X_{ij'}) = 0.9^{|j-j'|}$ for $1 \leq j, j' \leq p$. All of the coefficients were zero except for X_{i1}, X_{i2} and $X_{i,100}$.

Example 6: For each i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{q} \beta_j X_{ij} + \sum_{j=q+1}^{p} \beta_j X_{ij} \right), \quad i = 1, \ldots, n,$$

where $\beta_{q+1} = \cdots = \beta_p = 0$, $(\beta_1 = \cdots = \beta_q) = (3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)$, and $q = 10$. X_{ij}'s were generated from the multivariate standard normal distribution for $1 \leq j \leq q$ and

$$X_{ij} = d_{ij} + b \sum_{l=1}^{q} X_{il}, \quad i = 1, \ldots, n,$$

for $q + 1 \leq j \leq p$, where $b = (3/4q)^{1/2}$ and $(d_{i(q+1)}, \ldots, d_{ip})$ are generated from the multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix with all of the diagonal entries being 1 and off-diagonal being 0.25, and are independent of X_{ij} for $1 \leq j \leq q$.

Example 7: For each i,

$$c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = c \times \left(\sum_{j=1}^{p} \beta_j X_{ij} \right), \quad i = 1, \ldots, n,$$

where $\beta_j = 0.2$ for $j = 1, \ldots, 7, 501, \ldots, 508$ and β_j =0, otherwise. X_{ij}'s were generated from the multivariate normal distribution for $1 \leq j \leq 500$ with mean 0 and a covariance matrix with all of the diagonal entries being 1, $\mathrm{cov}(X_{ij}, X_{ij'}) = 0.9^{|j-j'|}$ for $1 \leq j, j' \leq 15$, and $\mathrm{cov}(X_{ij}, X_{ij'}) = 0$ for $16 \leq j, j' \leq 500$. X_{ij}'s were generated from the multivariate double exponential distribution for $501 \leq j \leq 508$ with location parameter equal to 0 and a covariance matrix with all of the diagonal entries being 1, and independent of X_{ij} for $1 \leq j \leq 500$.

Examples 1 and 3 were adopted from Wang [93], while **Examples 2, 4, and 6** were borrowed from Fan *et al.* [87], Hwang *et al.* [95], and Ing *et al.* [113], respectively. We then generated the responses from the following three models.

Normal Model: $Y_i = c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$.

Binomial Model: $Y_i \sim$ Bernoulli($\exp(c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})/\{1 + \exp(c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})\}$).

Poisson Model: $Y_i \sim$ Poisson($\exp(c\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta})$).

Our simulated examples cover a wide range of models (Normal, Binomial, Poisson), having their covariates generated from various distributions (multivariate normal, exponential, double exponential, uniform, and mixture) with diverse set of complex covariance structures (independent, compound symmetry, autoregressive, unstructured) and comprised of strong and weak signals including hidden features.

We considered $n = 400$ and $p = 1{,}000$ throughout all of the examples. We specified the magnitude of the coefficients in the GLMs with a constant multiplier, c. For Examples 1-7, this constant was set, respectively for the Normal, Binomial and Poisson models, to be: (1, 1, 0.3), (1, 1.5, 0.3), (1, 1, 0.1), (1, 1.5, 0.3), (1, 3, 2), (1, 1, 1), and (1, 3, 2). For each parameter configuration, we simulated 500 independent data sets. We evaluated the performance of the methods by the criteria of true positives (TP), false positives (FP), the estimated probability of including the true models (PIT), the mean squared error (MSE) of $\hat{\boldsymbol{\beta}}$, and the mean squared prediction error (MSPE). To compute

the MSPE, we randomly partitioned the samples into the training (75%) and testing (25%) sets. The models obtained from the training datasets were used to predict the responses in the testing datasets. Tables 3.2–3.4 report the average TP, FP, PIT, MSE, and MSPE over 500 datasets along with the standard deviations. The findings are summarized below.

First, the proposed STEPWISE method was able to detect all the true signals with nearly zero FPs. Specifically, in all of the Examples, STEPWISE outperformed the other methods by detecting more TPs with fewer FPs, whereas LASSO, SIS+LASSO and dgLARS included much more FPs.

Second, though a smaller γ in FR and SC led to the inclusion of all TPs with a PIT close to 1, it incurred more FPs. On the other hand, a larger γ may eliminate some TPs, resulting in a smaller PIT and a larger MSPE.

Third, for the Normal model, the STEPWISE method yielded an MSE close to 0, the smallest among all the competing methods. The Binary and Poisson data challenged all of the methods, and the MSEs for all the methods were non-negligible. However, the STEPWISE method still produced the lowest MSE. The results seemed to verify the consistency of $\hat{\boldsymbol{\beta}}$, which distinguished the proposed STEPWISE method from the other regularized methods and highlighted its ability to provide a more accurate means to characterize the effects of high dimensional predictors.

Fourth, for all three models, STEPWISE procedure demonstrated a vivid advantage over other competing methods: for the Poisson model, it outperformed all methods by selecting the highest number of TP and keeping FP at the low rate. In fact, SIS+LASSO failed to detect any TP while including an incomparably high number of FP. High FP rates were observed in dgLARS method as well. Similarly, for the Binomial model, STEPWISE selected almost all TPs while including near zero FPs. LASSO, SIS+LASSO, and dgLARS selected a learge number of FPs while selecting less TPs than our proposed method.

Table 3.2: Normal model

Example	Method	TP	FP	PIT	MSE ($\times 10^{-4}$)	MSPE
1 (p_0 = 8)	LASSO(1SE)	8.00 (0.00)	5.48 (6.61)	1.00 (0.00)	2.45	1.148
	LASSO(BIC)	8.00 (0.00)	2.55 (2.48)	1.00 (0.00)	2.58	1.172
	SIS+LASSO(1SE)	8.00 (0.00)	6.59 (4.22)	1.00 (0.00)	1.49	1.042
	SIS+LASSO(BIC)	8.00 (0.00)	6.04 (3.33)	1.00 (0.00)	1.37	1.025
	dgLARS(BIC)	8.00 (0.00)	3.52(2.53)	1.00 (0.00)	2.25	1.130
	SC (γ_L)	8.00 (0.00)	3.01 (1.85)	1.00 (0.00)	1.09	0.895
	SC (γ_H)	7.60 (1.59)	0.00 (0.00)	0.94 (0.24)	14.56	5.081
	FR (γ_L)	8.00 (0.00)	2.96 (2.04)	1.00 (0.00)	1.08	0.896
	FR (γ_H)	7.88 (0.84)	0.00 (0.00)	0.98 (0.14)	3.74	2.040
	STEPWISE	8.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.21	0.972
2 (p_0 = 8)	LASSO(1SE)	8.00 (0.00)	4.74 (4.24)	1.00 (0.00)	2.46	1.154
	LASSO(BIC)	8.00 (0.00)	2.12 (2.02)	1.00 (0.00)	2.62	1.182
	SIS+LASSO	7.99 (0.10)	6.84 (4.57)	0.99 (0.10)	1.65	1.058
	SIS+LASSO(BIC)	7.99 (0.10)	6.11 (3.85)	0.99 (0.10)	1.56	1.041
	dgLARS(BIC)	8.00 (0.00)	3.26(2.62)	1.00 (0.00)	2.28	1.138
	SC (γ_L)	8.00 (0.00)	2.73 (1.53)	1.00 (0.00)	0.98	0.901
	SC (γ_H)	7.30 (2.11)	0.00 (0.00)	0.90 (0.30)	23.70	6.397
	FR (γ_L)	8.00 (0.00)	2.45 (1.65)	1.00 (0.00)	0.92	0.907
	FR (γ_H)	7.94 (0.60)	0.00 (0.00)	0.99 (0.00)	2.69	2.062
	STEPWISE	8.00 (0.00)	0.01 (0.10)	1.00 (0.00)	0.21	0.972
3 (p_0 = 5)	LASSO(1SE)	5.00 (0.00)	8.24 (2.63)	1.00 (0.00)	3.07	1.084
	LASSO(BIC)	5.00 (0.00)	12.33 (3.28)	1.00 (0.00)	27.97	2.398
	SIS+LASSO(1SE)	0.97 (0.26)	15.94 (2.93)	0.00 (0.00)	1406.22	76.024
	SIS+LASSO(BIC)	0.97 (0.26)	16.20 (2.81)	0.00 (0.00)	1354.54	71.017
	dgLARS(BIC)	5.00 (0.00)	53.91 (14.44)	1.00 (0.00)	6.63	0.979

Continued on next page

Table 3.2 (cont'd)

Example	Method	TP	FP	PIT	MSE ($\times 10^{-4}$)	MSPE
	SC (γ_L)	4.48 (0.50)	0.25 (0.44)	0.48 (0.50)	21.74	3.086
	SC (γ_H)	4.48 (0.50)	0.14 (0.35)	0.48 (0.50)	21.70	2.065
	FR (γ_L)	5.00 (0.00)	0.23 (0.66)	1.00 (0.00)	0.27	0.973
	FR (γ_H)	5.00 (0.00)	0.14 (0.35)	1.00 (0.00)	0.15	0.074
	STEPWISE	5.00 (0.00)	0.03 (0.22)	1.00 (0.00)	0.18	0.976
4 ($p_0 = 14$)	LASSO(1SE)	14.00 (0.00)	29.84 (15.25)	1.00 (0.00)	13.97	1.148
	LASSO(BIC)	13.94 (0.24)	4.92 (5.54)	0.94 (0.24)	38.69	1.995
	SIS+LASSO(1SE)	11.44 (1.45)	15.19 (7.29)	0.05 (0.21)	133.38	4.714
	SIS+LASSO(BIC)	11.35 (1.51)	10.98 (7.19)	0.05 (0.21)	137.06	4.940
	dgLARS(BIC)	14.00 (0.00)	13.93 (6.68)	1.00 (0.00)	18.08	1.329
	SC (γ_L)	13.68 (0.60)	0.86 (0.62)	0.75 (0.44)	11.80	1.148
	SC (γ_H)	4.20 (2.80)	0.03 (0.17)	0.03 (0.17)	407.86	6.567
	FR (γ_L)	14.00 (0.00)	0.50 (0.76)	1.00 (0.00)	1.23	0.940
	FR (γ_H)	4.99 (3.07)	0.00 (0.00)	0.03 (0.17)	360.65	6.640
	STEPWISE	14.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.91	0.958
5 ($p_0 = 3$)	LASSO(1SE)	3.00 (0.00)	22.76 (9.05)	1.00 (0.00)	1.01	0.044
	LASSO(BIC)	3.00 (0.00)	8.29 (3.23)	1.00 (0.00)	1.75	0.054
	SIS+LASSO(1SE)	3.00 (0.00)	8.40 (3.10)	1.00 (0.00)	0.44	0.041
	SIS+LASSO(BIC)	3.00 (0.00)	9.58 (3.36)	1.00 (0.00)	0.29	0.040
	dgLARS(BIC)	3.00 (0.00)	13.39 (4.94)	1.00 (0.00)	1.28	0.048
	SC (γ_L)	3.00 (0.00)	1.47 (0.67)	1.00 (0.00)	0.03	0.038
	SC (γ_H)	2.01 (0.10)	0.01 (0.10)	0.01 (0.10)	4.51	0.008
	FR (γ_L)	3.00 (0.00)	1.21 (1.01)	1.00 (0.00)	0.03	0.038
	FR (γ_H)	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.01	0.003
	STEPWISE	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.01	0.039

Continued on next page

Table 3.2 (cont'd)

Example	Method	TP	FP	PIT	MSE ($\times 10^{-4}$)	MSPE
6 (p_0 = 10)	LASSO(1SE)	10.00 (0.00)	46.23 (6.61)	1.00 (0.00)	37.70	1.50
	LASSO(BIC)	9.86 (0.34)	59.05 (6.18)	0.86 (0.34)	698.70	13.82
	SIS+LASSO(1SE)	0.00 (0.00)	37.53 (3.29)	0.00 (0.00)	4620.54	127.51
	SIS+LASSO(BIC)	0.00 (0.00)	38.05 (3.21)	1.00 (0.00)	4644.21	118.64
	dgLARS(BIC)	10.00 (0.00)	156.15 (26.47)	1.00 (0.00)	20.96	0.88
	SC (γ_L)	2.99 (0.08)	1.41 (0.49)	0.00 (0.00)	2868.96	116.26
	SC (γ_H)	0.96 (1.26)	1.06 (0.23)	0.00 (0.00)	4775.34	51.16
	FR (γ_L)	7.45 (0.08)	3.34 (0.33)	0.00 (0.00)	657.77	30.03
	FR (γ_H)	1.48 (2.46)	2.00 (0.08)	0.00 (0.00)	4345.01	55.47
	STEPWISE	7.45 (0.08)	2.11 (0.33)	0.00 (0.00)	653.07	29.05
7 (p_0 = 15)	LASSO(1SE)	14.71 (0.49)	8.64 (4.88)	0.73 (0.44)	0.77	1.16
	LASSO(BIC)	14.67 (0.53)	6.57 (3.51)	0.71 (0.45)	0.76	1.16
	SIS+LASSO(1SE)	13.45 (1.77)	10.16 (4.20)	0.38 (0.48)	1.32	1.07
	SIS+LASSO(BIC)	13.44 (1.76)	9.55 (4.07)	0.36 (0.48)	1.34	1.05
	dgLARS(BIC)	14.69 (0.51)	7.02 (3.54)	0.72 (0.45)	0.76	1.14
	SC (γ_L)	11.12 (0.53)	3.01 (1.82)	0.00 (0.00)	3.34	0.94
	SC (γ_H)	3.19 (3.51)	0.06 (0.08)	0.00 (0.00)	5.97	14.17
	FR (γ_L)	12.20 (0.60)	3.15 (2.00)	0.00 (0.00)	2.27	0.91
	FR (γ_H)	6.64 (3.38)	0.01 (0.11)	0.00 (0.00)	6.49	12.43
	STEPWISE	12.20 (0.08)	0.09 (0.29)	0.00 (0.00)	2.99	1.03

Continued on next page

Table 3.2 (cont'd)

Example	Method	TP	FP	PIT	MSE ($\times 10^{-4}$)	MSPE

Note: TP, true positives; FP, false positives; PIT, probability of including all true predictors in the selected predictors; MSE, mean squared error of $\hat{\beta}$; MSPE, mean squared prediction error; numbers in the parentheses are standard deviations; LASSO(BIC), LASSO with the tuning parameter chosen to give the smallest BIC for the models on the LASSO path; LASSO(1SE), LASSO with the tuning parameter chosen by the one-standard -error rule; SIS+LASSO(BIC), sure independence screening by [87] followed by LASSO(BIC); SIS+LASSO(1SE), sure independence screening followed by LASSO(1SE); dgLARS(BIC), differential geometric least angle regression by [90, 91] with the tuning parameter chosen to give the smallest BIC on the dgLARS path; SC(γ), sequentially conditioning approach by [99]; FR(γ), forward regression by [93]; STEPWISE, the proposed method; In FR and SC, the smaller and large values of γ are presented as γ_L and γ_H, respectively; p_0 denotes the number of true signals; LASSO(1SE), LASSO(BIC), SIS, and dgLARS were conducted via R packages `glmnet` [127], `ncvreg` [128], `screening` [129], and `dglars` [130], respectively.

Table 3.3: Binomial model

Example	Method	TP	FP	PIT	MSE	MSPE
1 ($p_0 = 8$)	LASSO(1SE)	7.99 (0.10)	4.77 (5.56)	0.99 (0.10)	0.021	0.104
	LASSO(BIC)	7.99 (0.10)	3.19 (2.34)	0.99 (0.10)	0.021	0.104
	SIS+LASSO(1SE)	7.94 (0.24)	35.42 (6.77)	0.94 (0.24)	0.119	0.048
	SIS+LASSO(BIC)	7.94 (0.24)	16.83 (21.60)	0.94 (0.24)	0.119	0.073
	dgLARS(BIC)	8.00 (0.00)	3.27 (2.29)	1.00 (0.00)	0.019	0.102
	SC (γ_L)	8.00 (0.00)	2.81 (1.47)	1.00 (0.00)	0.009	0.073
	SC (γ_H)	1.02 (0.14)	0.00 (0.00)	0.00 (0.00)	0.030	0.028

Continued on next page

Table 3.3 (cont'd)

Example	Method	TP	FP	PIT	MSE	MSPE
	FR (γ_L)	8.00 (0.00)	3.90 (2.36)	1.00 (0.00)	0.032	0.066
	FR (γ_H)	2.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.025	0.027
	STEPWISE	7.98 (0.14)	0.08 (0.53)	0.98 (0.14)	0.002	0.094
2 (p_0 = 8)	LASSO(1SE)	7.98 (0.14)	3.29 (2.76)	0.98 (0.14)	0.054	0.073
	LASSO(BIC)	7.99 (0.10)	3.84 (2.72)	0.99 (0.10)	0.052	0.067
	SIS+LASSO(1SE)	7.92 (0.27)	28.20 (7.31)	0.92 (0.27)	0.038	0.030
	SIS+LASSO(BIC)	7.92 (0.27)	9.60 (12.92)	0.92 (0.27)	0.051	0.058
	dgLARS(BIC)	7.99 (0.10)	3.94 (2.65)	0.99 (0.10)	0.050	0.067
	SC (γ_L)	7.72 (0.45)	0.39 (0.49)	0.72 (0.45)	0.005	0.063
	SC (γ_H)	1.13 (0.37)	0.00 (0.00)	0.00 (0.00)	0.069	0.044
	FR (γ_L)	7.99 (0.10)	0.66 (0.76)	0.99 (0.10)	0.014	0.051
	FR (γ_H)	2.10 (0.30)	0.00 (0.00)	0.00 (0.00)	0.061	0.033
	STEPWISE	7.99 (0.10)	0.02 (0.14)	0.99 (0.10)	0.004	0.056
3 (p_0 = 5)	LASSO(1SE)	4.51 (0.52)	7.36 (2.57)	0.52 (0.50)	0.155	0.051
	LASSO(BIC)	4.98 (0.14)	5.97 (2.25)	0.98 (0.14)	0.118	0.037
	SIS+LASSO(1SE)	0.85 (0.46)	10.66 (3.01)	0.00 (0.00)	0.206	0.186
	SIS+LASSO(BIC)	0.85 (0.46)	12.10 (3.13)	0.00 (0.00)	0.197	0.185
	dgLARS(BIC)	4.92 (0.27)	16.21 (6.21)	0.92 (0.27)	0.112	0.035
	SC (γ_L)	4.32 (0.49)	0.47 (0.50)	0.33 (0.47)	0.016	0.048
	SC (γ_H)	2.62 (1.34)	0.42 (0.50)	0.00 (0.00)	0.104	0.066
	FR (γ_L)	4.98 (0.14)	0.67 (0.79)	0.98 (0.14)	0.020	0.033
	FR (γ_H)	2.98 (0.95)	0.40 (0.49)	0.00 (0.00)	0.087	0.043
	STEPWISE	4.97 (0.17)	0.04 (0.28)	0.97 (0.17)	0.014	0.034

Continued on next page

Table 3.3 (cont'd)

Example	Method	TP	FP	PIT	MSE	MSPE
4 (p_0 = 14)	LASSO(1SE)	9.96 (1.89)	6.78 (7.92)	0.01 (0.01)	0.112	0.107
	LASSO(BIC)	9.33 (1.86)	2.79 (2.87)	0.00 (0.00)	0.112	0.118
	SIS+LASSO(1SE)	10.03 (1.62)	28.01 (9.54)	0.03 (0.17)	0.098	0.070
	SIS+LASSO(BIC)	8.90 (1.99)	5.42 (10.64)	0.01 (0.10)	0.114	0.120
	dgLARS(BIC)	9.31 (1.85)	2.84 (2.86)	0.00 (0.00)	0.110	0.117
	SC (γ_L)	9.48 (1.40)	2.35 (2.14)	0.00 (0.00)	0.043	0.070
	SC (γ_H)	1.17 (0.40)	0.00 (0.00)	0.00 (0.00)	0.125	0.049
	FR (γ_L)	11.83 (1.39)	1.58 (1.60)	0.09 (0.29)	0.026	0.048
	FR (γ_H)	2.06 (0.24)	0.00 (0.00)	0.00 (0.00)	0.119	0.032
	STEPWISE	11.81 (1.42)	1.52 (1.58)	0.09 (0.29)	0.026	0.048
5 (p_0 = 3)	LASSO(1SE)	2.00 (0.00)	1.55 (1.76)	0.00 (0.00)	0.008	0.215
	LASSO(BIC)	2.00 (0.00)	1.86 (1.57)	0.00 (0.00)	0.008	0.213
	SIS+LASSO(1SE)	2.23 (0.42)	10.81 (6.45)	0.23 (0.42)	0.007	0.192
	SIS+LASSO(BIC)	2.10 (0.30)	3.60 (4.65)	0.10 (0.30)	0.007	0.206
	dgLARS(BIC)	2.00 (0.00)	1.64 (1.49)	0.00 (0.00)	0.008	0.213
	SC (γ_L)	2.27 (0.49)	7.16 (3.20)	0.29 (0.46)	0.060	0.166
	SC (γ_H)	1.87 (0.34)	0.03 (0.17)	0.00 (0.00)	0.005	0.030
	FR (γ_L)	2.96 (0.20)	8.88 (5.39)	0.96 (0.20)	0.013	0.147
	FR (γ_H)	1.97 (0.17)	0.03 (0.17)	0.00 (0.00)	0.005	0.019
	STEPWISE	2.89 (0.31)	0.76 (1.70)	0.89 (0.31)	0.001	0.194
6 (p_0 = 10)	LASSO(1SE)	6.10 (1.08)	31.66 (4.63)	0.00 (0.00)	0.41	0.07
	LASSO(BIC)	7.88 (0.97)	30.41 (4.63)	0.02 (0.16)	0.38	0.05

Continued on next page

Table 3.3 (cont'd)

Example	Method	TP	FP	PIT	MSE	MSPE
	SIS+LASSO(1SE)	0.00 (0.00)	28.40 (3.61)	0.00 (0.00)	0.45	0.15
	SIS+LASSO(BIC)	0.00 (0.00)	30.18 (3.24)	0.00 (0.00)	0.45	0.14
	dgLARS(BIC)	4.12 (2.29)	32.37 (7.61)	0.00 (0.00)	0.42	0.09
	SC (γ_L)	7.71 (0.58)	2.95 (0.95)	0.00 (0.00)	0.19	0.05
	SC (γ_H)	0.00 (0.00)	1.00 (0.00)	0.00 (0.00)	0.45	0.02
	FR (γ_L)	8.52 (0.89)	3.24 (1.15)	0.00 (0.00)	0.09	0.04
	FR (γ_H)	0.12 (0.32)	1.88 (0.32)	0.00 (0.00)	0.45	0.01
	STEPWISE	8.52 (0.92)	0.58 (0.77)	0.00 (0.00)	0.09	0.04
7 (p_0 = 15)	LASSO(1SE)	9.80 (1.14)	5.60 (4.29)	0.00 (0.00)	0.01	0.04
	LASSO(BIC)	9.74 (1.11)	4.68 (2.63)	0.00 (0.00)	0.01	0.04
	SIS+LASSO(1SE)	10.67 (1.39)	25.17 (5.85)	0.00 (0.00)	0.04	0.02
	SIS+LASSO(BIC)	10.62 (1.64)	24.02 (15.84)	0.00 (0.00)	0.01	0.01
	dgLARS(BIC)	9.84 (1.08)	4.58 (2.45)	0.00 (0.00)	0.01	0.04
	SC (γ_L)	8.92 (0.53)	1.75 (0.94)	0.00 (0.00)	0.01	0.02
	SC (γ_H)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01	0.02
	FR (γ_L)	9.56 (0.58)	1.52 (0.92)	0.00 (0.00)	0.07	0.02
	FR (γ_H)	2.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01	0.02
	STEPWISE	9.50 (0.54)	0.67 (0.83)	0.00 (0.00)	0.04	0.02

Note: Abbreviations are explained in the footnote of Table 3.2.

Table 3.4: Poisson model

Example	Method	TP	FP	PIT	MSE	MSPE
1 ($p_0 = 8$)	LASSO(1SE)	7.93 (0.43)	4.64 (4.82)	0.96 (0.19)	0.001	4.236
	LASSO(BIC)	7.99 (0.10)	14.37 (14.54)	0.99 (0.10)	0.001	3.133
	SIS+LASSO(1SE)	7.89 (0.37)	25.37 (8.39)	0.91 (0.29)	0.001	3.247
	SIS+LASSO(BIC)	7.89 (0.37)	17.77 (11.70)	0.91 (0.29)	0.001	3.078
	dgLARS(BIC)	8.00 (0.00)	13.28 (14.31)	1.00 (0.00)	0.001	3.183
	SC (γ_L)	7.96 (0.20)	4.94 (3.46)	0.96 (0.20)	0.001	2.874
	SC (γ_H)	5.05 (1.70)	0.04 (0.24)	0.07 (0.26)	0.001	3.902
	FR (γ_L)	7.93 (0.26)	4.86 (3.73)	0.93 (0.26)	0.001	2.837
	FR (γ_H)	5.13 (1.61)	0.06 (0.31)	0.07 (0.26)	0.001	3.833
	STEPWISE	7.91 (0.29)	2.77 (2.91)	0.91 (0.29)	0.001	3.410
2 ($p_0 = 8$)	LASSO(1SE)	8.00 (0.00)	2.23 (3.52)	1.00 (0.00)	0.001	3.981
	LASSO(BIC)	8.00 (0.00)	8.98 (8.92)	1.00 (0.00)	0.001	3.107
	SIS+LASSO(1SE)	7.98 (0.14)	22.85 (7.08)	0.98 (0.14)	0.001	2.824
	SIS+LASSO(BIC)	7.98 (0.14)	13.55 (8.24)	0.98 (0.14)	0.001	2.937
	dgLARS(BIC)	8.00 (0.00)	8.91 (9.10)	1.00 (0.00)	0.001	3.099
	SC (γ_L)	8.00 (0.00)	3.89 (2.89)	1.00 (0.00)	0.000	2.979
	SC (γ_H)	5.68 (1.45)	0.00 (0.00)	0.12 (0.33)	0.001	3.971
	FR (γ_L)	8.00 (0.00)	3.60 (2.80)	1.00 (0.00)	0.000	3.032
	FR (γ_H)	5.71 (1.42)	0.00 (0.00)	0.10 (0.30)	0.001	3.911
	STEPWISE	7.98 (0.14)	2.00 (2.23)	0.98 (0.14)	0.000	3.589
3 ($p_0 = 5$)	LASSO(1SE)	4.37 (0.51)	6.88 (2.61)	0.38(0.48)	0.001	1.959
	LASSO(BIC)	4.79 (0.41)	5.62 (2.17)	0.79 (0.41)	0.000	2.044

Continued on next page

Table 3.4 (cont'd)

Example	Method	TP	FP	PIT	MSE	MSPE
	SIS+LASSO(1SE)	0.86 (0.47)	10.11 (2.55)	0.00 (0.00)	0.002	3.266
	SIS+LASSO(BIC)	0.86 (0.47)	11.86 (2.99)	0.00 (0.00)	0.002	3.160
	dgLARS(BIC)	4.55 (0.51)	18.29 (6.13)	0.56 (0.49)	0.001	1.877
	SC (γ_L)	4.73 (0.45)	0.53 (0.66)	0.73 (0.45)	0.000	2.479
	SC (γ_H)	2.84 (0.63)	0.40 (0.49)	0.00 (0.00)	0.001	0.664
	FR (γ_L)	4.54 (0.52)	1.98 (2.19)	0.55 (0.50)	0.000	2.128
	FR (γ_H)	2.71 (0.70)	0.43 (0.50)	0.00 (0.00)	0.001	0.605
	STEPWISE	4.54 (0.52)	1.77 (2.01)	0.55 (0.50)	0.000	2.132
4 (p_0 = 14)	LASSO(1SE)	10.01 (1.73)	3.91 (6.03)	0.01 (0.10)	0.003	15.582
	LASSO(BIC)	12.11 (1.46)	36.56 (22.43)	0.19 (0.39)	0.002	5.688
	SIS+LASSO(1SE)	10.42 (1.66)	21.41 (8.87)	0.03 (0.17)	0.003	11.316
	SIS+LASSO(BIC)	10.73 (1.66)	32.67 (8.92)	0.03 (0.17)	0.003	8.545
	dgLARS(BIC)	12.05 (1.52)	38.70 (28.97)	0.18 (0.38)	0.002	5.111
	SC (γ_L)	10.33 (1.63)	10.48 (6.66)	0.02 (0.14)	0.002	4.499
	SC (γ_H)	5.32 (1.92)	0.52 (1.37)	0.00 (0.00)	0.003	14.005
	FR (γ_L)	12.00 (1.71)	8.93 (6.36)	0.23 (0.42)	0.001	4.503
	FR (γ_H)	5.65 (2.13)	0.38 (1.15)	0.00 (0.00)	0.003	13.802
	STEPWISE	11.80 (1.72)	5.97 (5.37)	0.19 (0.39)	0.001	5.809
5 (p_0 = 3)	LASSO(1SE)	2.00 (0.00)	1.13 (2.85)	0.00 (0.00)	0.003	2.674
	LASSO(BIC)	2.01 (0.10)	2.82 (2.52)	0.01 (0.10)	0.003	2.583
	SIS+LASSO(1SE)	2.87 (0.34)	9.28 (3.85)	0.87 (0.34)	0.002	2.455
	SIS+LASSO(BIC)	2.87 (0.34)	9.88 (4.29)	0.87 (0.34)	0.002	2.355
	dgLARS(BIC)	2.00 (0.00)	2.88 (2.38)	0.00 (0.00)	0.003	2.562

Continued on next page

Table 3.4 (cont'd)

Example	Method	TP	FP	PIT	MSE	MSPE
	SC (γ_L)	2.75 (0.44)	3.27 (1.75)	0.75 (0.44)	0.001	2.339
	SC (γ_H)	2.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.003	1.086
	FR (γ_L)	3.00 (0.00)	2.80 (1.73)	1.00 (0.00)	0.001	2.326
	FR (γ_H)	2.40 (0.49)	0.00 (0.00)	0.40 (0.49)	0.002	0.981
	STEPWISE	3.00 (0.00)	0.35 (0.59)	1.00 (0.00)	0.001	2.977
6 (p_0 = 10)	LASSO(1SE)	6.08 (1.16)	32.54 (4.83)	0.00 (0.00)	0.01	7.64
	LASSO(BIC)	8.15 (0.94)	37.56 (7.96)	0.06 (0.23)	0.01	5.93
	SIS+LASSO(1SE)	0.00 (0.00)	26.34 (3.87)	0.00 (0.00)	0.01	12.27
	SIS+LASSO(BIC)	0.00 (0.00)	28.03 (4.29)	0.00 (0.00)	0.01	12.06
	dgLARS(BIC)	8.42 (1.22)	75.21 (15.56)	0.21 (0.41)	0.01	4.55
	SC (γ_L)	9.45 (0.72)	9.80 (3.12)	0.76 (0.42)	0.00	4.77
	SC (γ_H)	3.73 (1.73)	2.15 (0.45)	0.00 (0.00)	0.01	3.43
	FR (γ_L)	9.54 (1.05)	11.26 (2.97)	0.55 (0.50)	0.01	4.42
	FR (γ_H)	2.85 (1.85)	2.70 (0.57)	0.00 (0.00)	0.01	3.28
	STEPWISE	9.54 (1.05)	4.30 (2.03)	0.55 (0.50)	0.01	6.01
7 (p_0 = 15)	LASSO(1SE)	11.98 (2.20)	4.00 (3.56)	0.12 (0.32)	0.01	20.71
	LASSO(BIC)	14.93 (0.29)	51.44 (11.68)	0.94 (0.23)	0.00	7.17
	SIS+LASSO(1SE)	12.51 (1.20)	17.32 (5.91)	0.03 (0.18)	0.00	15.24
	SIS+LASSO(BIC)	12.50 (1.29)	33.31 (7.69)	0.05 (0.23)	0.01	11.93
	dgLARS(BIC)	14.94 (0.27)	56.10 (17.38)	0.95 (0.20)	0.00	6.58
	SC (γ_L)	12.65 (1.43)	20.94 (4.47)	0.08 (0.28)	0.00	3.92
	SC (γ_H)	6.55 (1.98)	0.54 (0.84)	0.00 (0.00)	0.00	5.67
	FR (γ_L)	13.60 (1.12)	19.11 (4.03)	0.21 (0.41)	0.00	3.96

Continued on next page

Table 3.4 (cont'd)

Example	Method	TP	FP	PIT	MSE	MSPE
	FR (γ_H)	7.13 (1.87)	0.31 (0.55)	0.00 (0.00)	0.00	6.54
	STEPWISE	13.54 (1.31)	5.30 (3.12)	0.21 (0.39)	0.00	9.23

Note: Abbreviations are explained in the footnote of Table 3.2.

3.6 Applications: Real Data Analysis

3.6.1 A Study of Gene Regulation in the Mammalian Eye

To demonstrate the utility of our proposed method, we analyzed a microarray dataset from Scheetz *et al.* [41] with 120 twelve-week male rats selected for eye tissue harvesting. The dataset contained more than 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array); see Scheetz *et al.* [41] for a more detailed description of the data.

Although our method was applicable to the original 31,042 probe sets, many probes turned out to have very small variances and were unlikely to be informative for correlative analyses. Therefore, using variance as the screening criterion, we selected 5,000 genes with the largest variances in expressions and correlated them with gene *TRIM32* that has been found to cause Bardet-Biedl syndrome, a genetically heterogeneous disease of multiple organ systems including the retina [42]. We applied the proposed STEPWISE method to the dataset with $n = 120$ and $p = 5{,}000$, and treated the *TRIM32* gene expression as the response variable and the expressions of 5,000 genes as the predictors. With no prior biological information available, we started with the empty set. To choose η_1 and η_2, we carried out 5-fold cross-validation to minimize the mean squared prediction error (MSPE) by using the following grid search: $\eta_1 = \{0, 0.25, 0.5, 1\}$ and $\eta_2 = \{1, 2, 3, 4, 5\}$, and set $\eta_1 = 1$ and $\eta_2 = 4$. We also performed the same procedure to choose the γ for FR and SC. The regularization parameters in LASSO and dgLARS were selected to minimize BIC values.

In the forward step, STEPWISE selected the probes of *1376747_at*, *1381902_at*, *1382673_at* and *1375577_at*, and the backward step eliminated probe *1375577_at*. The STEPWISE procedure produced the following final predictive model:

TRIM32 = 4.6208 + 0.2310 × (*1376747_at*) + 0.1914 × (*1381902_at*) + 0.1263 × (*1382673_at*).

Table A.1 in Appendix A presents the numbers of overlapping genes among competing methods. It shows that the two out of three probes, *1381902_at* and *1376747_at*, selected from our method are also discovered by the other methods, except for dgLARS.

Next, we performed Leave-One-Out Cross-Validation (LOOCV) to obtain the distribution of the model size (MS) and MSPE for the competing methods. As reported in Table 3.5 and Figure 3.1, LASSO, SIS+LASSO and dgLARS tended to select more variables than the forward approaches and STEPWISE. Among all of the methods, STEPWISE selected the fewest variables but with almost the same MSPE as the other methods.

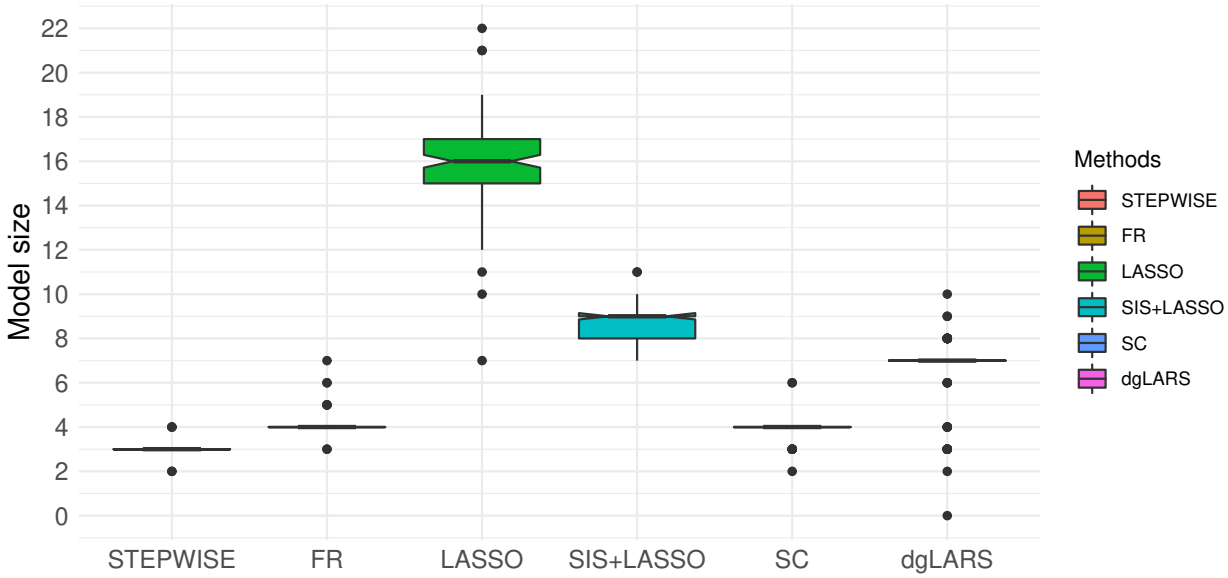


Figure 3.1: Box plot of model sizes for each method over 120 different training samples from the mammalian eye data set. STEPWISE was performed with $\eta_1 = 1$ and $\eta_2 = 4$, and FR and SC were conducted with $\gamma = 1$.

Table 3.5: Comparisons of MSPE between competing methods using the mammalian eye data set.

	STEPWISE	FR	LASSO	SIS+LASSO	SC	dgLARS
Training set	0.005	0.005	0.005	0.006	0.005	0.014
Testing set	0.011	0.012	0.010	0.009	0.014	0.020

Note: The mean squared prediction error (MSPE) was averaged over 120 splits. LASSO, least absolute shrinkage and selection operator with regularization parameter that gives the smallest BIC; SIS+LASSO, sure independence screening by [87] followed by LASSO; dgLARS, differential geometric least angle regression by [90, 91] that gives the smallest BIC; SC(γ), sequentially conditioning approach by [99]; FR(γ), forward regression by [93]; STEPWISE, the proposed method. STEPWISE was performed with $\eta_1 = 1$ and $\eta_2 = 4$, FR and SC were performed with $\gamma = 1$.

3.6.2 An Esophageal Squamous Cell Carcinoma Study

Esophageal squamous cell carcinoma (ESCC), the most common histological type of esophageal cancer, is known to be associated with poor overall survival, making early diagnosis crucial for treatment and disease management [47]. Several studies have investigated the roles of circulating microRNAs (miRNAs) in diagnosis of ESCC [45].

Using a clinical study that investigated the roles of miRNAs on the ESCC [57], we aimed to use miRNAs to predict ESCC risks and estimate their impacts on the development of ESCC. Specifically, with a dataset of serum profiling of 2,565 miRNAs from 566 ESCC patients and 4,965 controls without cancer, we demonstrated the utility of the proposed STEPWISE method in predicting ESCC with miRNAs. To proceed, we used a balance sampling scheme (283 cases and 283 controls) in the training dataset. The design of yielding an equal number of cases and controls in the training set has proved to be useful [57] for handling imbalanced outcomes as we encountered here. To validate our findings, samples were randomly divided into a training (n_1 = 566, p = 2,565) and testing set (n_2 = 4,965, p = 2,565).

The training set consisted of 283 patients with ESCC (median age of 65 years, 79% male) and 283 control patients (median age of 68 years, 46.3% male), and the testing set consisted of 283 patients with ESCC (median age of 67 years, 85.7% male) and 4,682 control patients (median age of 67.5 years, 44.5% male). Control patients without ESCC came from three sources: 323 individuals

from National Cancer Center Biobank (NCCB); 2,670 individuals from the Biobank of the National Center for Geriatrics and Gerontology (NCGG); and 1,972 individuals from Minoru Clinic (MC). More detailed characteristics of cases and controls in the training and testing sets are given in Table A.4.

We defined the binary outcome variable to be 1 if the subject was a case and 0 otherwise. As age and gender (0 = female, 1 = male) are important risk factors for ESCC [131, 132] and it is common to adjust for them in clinical models, we set the initial set in STEPWISE to be $F_0 = \{\text{age, gender}\}$. With $\eta_1 = 0$ and $\eta_2 = 3.5$ that were also chosen from 5-fold CV, our procedure recruited three miRNAs. More specifically, *miR-4783-3p*, *miR-320b*, *miR-1225-3p* and *miR-6789-5p* were selected among 2,565 miRNAs by the forward stage from the training set, and then the backward stage eliminated *miR-6789-5p*. In comparison, with $\gamma = 0$, both FR and SC selected four miRNAs, *miR-4783-3p*, *miR-320b*, *miR-1225-3p*, and *miR-6789-5p*.

The list of selected miRNAs by different methods is given in Table A.2 in Appendix A. Our findings were biologically meaningful, as the selected miRNAs had been identified by other cancer studies as well. Specifically, *miR-320b* was found to promote colorectal cancer proliferation and invasion by competing with its homologous *miR-320a* [133]. In addition, serum levels of *miR-320* family members were associated with clinical parameters and diagnosis in prostate cancer patients [134]. Mullany *et al.* [135] showed that *miR-4783-3p* was one of the miRNAs that could increase the risk of colorectal cancer death among rectal cancer cases. Finally, *miR-1225-5p* inhibited proliferation and metastasis of gastric carcinoma through repressing insulin receptor substrate-1 and activation of β-catenin signaling [136].

Aiming to identify a final model without resorting to a pre-screening procedure that may miss out on important biomarkers, we applied STEPWISE to reach the following predictive model for ESCC based on patients' demographics and miRNAs:

$\text{logit}^{-1}(-35.70 + 1.41 \times \textit{miR-4783-3p} + 0.98 \times \textit{miR-320b} + 1.91 \times \textit{miR-1225-3p} + 0.10 \times \textit{Age} - 2.02 \times \textit{Gender})$, where $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$.

In the testing dataset, the model had an area under the receiver operating curve (AUC) of 0.99 and

achieved a high accuracy of 0.96, with a sensitivity and specificity of 0.97 and 0.95, respectively. Also using the testing cohort, we evaluated the performance of the models sequentially selected by STEPWISE. Starting with a model containing age and gender, STEPWISE selected *miR-4783-3p*, *miR-320b* and *miR-1225-3p* in turn. Figure 3.3, showing the corresponding receiver operating curves (ROC) for these sequential models, revealed the improvement by sequentially adding predictors to the model and justified the importance of these variables in the final model. In addition, Figure 3.3 (e) illustrated that adding an extra miRNA selected by FR and SC made little improvement of the model's predictive power.

Furthermore, we conducted subgroup analysis within the testing cohort to study how the sensitivity of the final model differed by cancer stage, one of the most important risk factors. The sensitivity for stages 0, i.e., non-invasive cancer, 9 (n = 27), 1 (n = 128), 2 (n = 57), 3 (n = 61), and 4 (n = 10) was 1.00, 0.98, 0.97, 0.97, and 1.00, respectively. We next evaluated how the specificity varied across controls coming from different data sources. The specificity for the various control groups, namely, NCCB (n = 306), NCGG (n = 2,512), and MC (n = 1,864), was 0.99, 0.99, and 0.98, respectively. The results indicated the robust performance of the miRNA-based model toward cancer stages as well as data sources.

Finally, to compare STEPWISE with the other competing methods, we repeatedly applied the aforementioned balance sampling procedure and split the ESCC data into the training and testing sets 100 times. We obtained MSPE and the average of accuracy, sensitivity, specificity, and AUC. Figure 3.2 reported the model size of each method. Though STEPWISE selected fewer variables compared to the other variable selection methods (for example, LASSO selected 11-31 variables and dgLARS selected 12-51 variables), it achieved comparable prediction accuracy, specificity, sensitivity and AUC (see Table 3.6), evidencing the utility of STEPWISE for generating parsimonious models while maintaining competitive predictability.

We used `R` software [137] to obtain the numerical results in Sections 4 and 5 with following packages: `ggplot2` [138], `ncvreg` [128], `glmnet` [127], `dglars` [130], and `screening` [129].

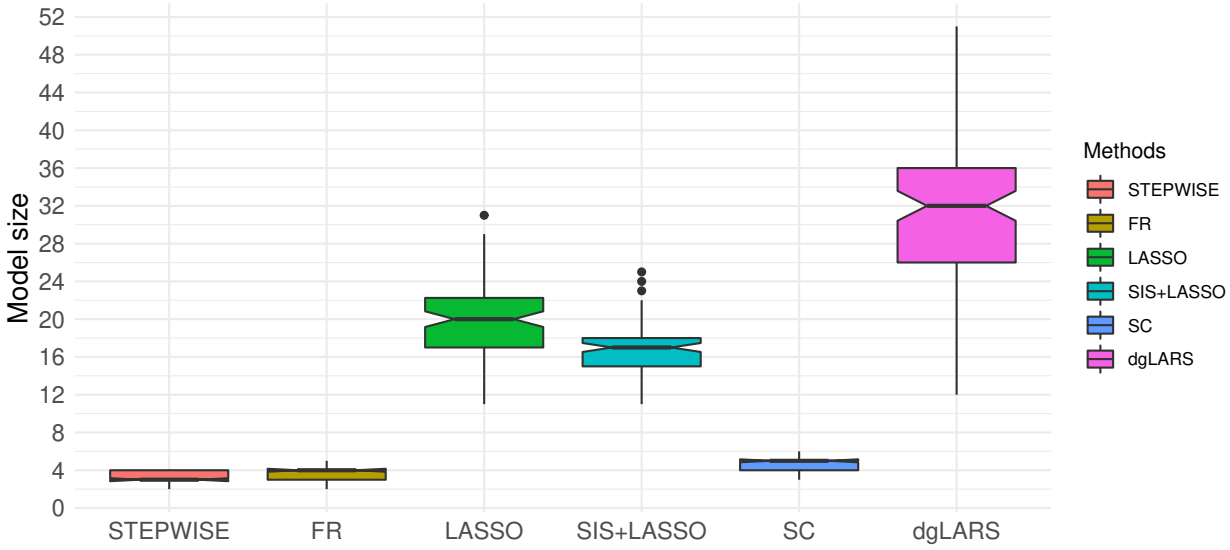
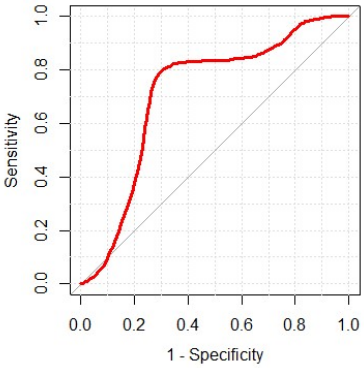


Figure 3.2: Box plot of model sizes for each method based on 100 ESCC training datasets. Performance of STEPWISE is reported with $\eta_1 = 0$ and $\eta_2 = 3.5$. Performance of SC and FR are reported with $\gamma = 0$.

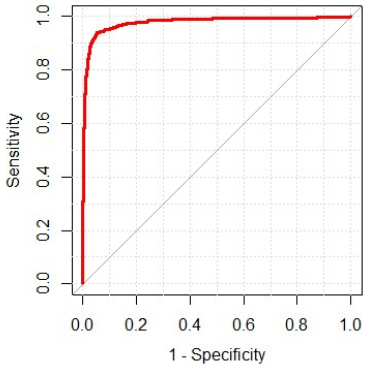
Table 3.6: Comparisons of competing methods over 100 independent splits of the ESCC data into training and testing sets

Training set	MSPE	Accuracy	Sensitivity	Specificity	AUC
STEPWISE	0.02	0.97	0.98	0.97	1.00
SC	0.01	0.99	0.98	0.98	1.00
FR	0.02	0.99	0.97	0.97	1.00
LASSO	0.01	0.98	1.00	0.97	1.00
SIS+LASSO	0.01	0.99	1.00	0.99	1.00
dgLARS	0.04	0.96	0.99	0.94	1.00
Test set	MSPE	Accuracy	Sensitivity	Specificity	AUC
STEPWISE	0.04	0.96	0.97	0.95	0.99
SC	0.03	0.96	0.97	0.96	0.99
FR	0.04	0.96	0.97	0.95	0.99
LASSO	0.03	0.96	0.99	0.95	1.00
SIS+LASSO	0.02	0.97	0.99	0.96	1.00
dgLARS	0.05	0.94	0.98	0.94	1.00

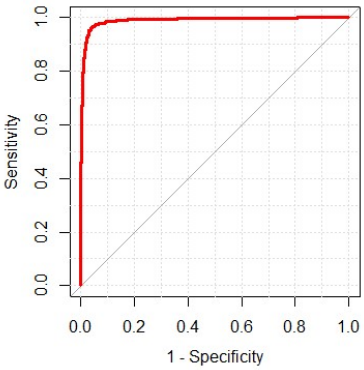
Note: Values are averaged over 100 splits. STEPWISE was performed with $\eta_1 = 0$ and $\eta_2 = 1$. SC and FR were performed with $\gamma = 1$. The regularization parameters in LASSO and dgLARS were selected to minimize the BIC.



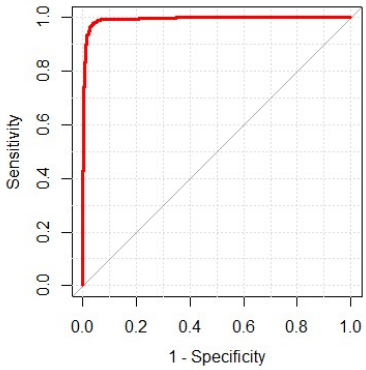
(a) Model 1, AUC= 0.71



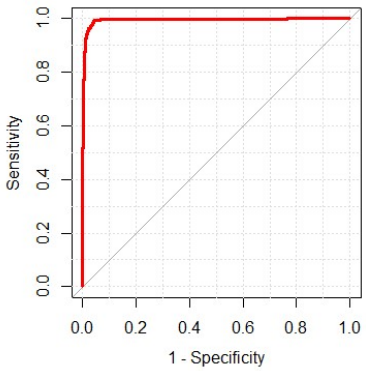
(b) Model 2, AUC=0.97



(c) Model 3, AUC=0.98



(d) Model 4, AUC=0.99



(e) Model 5, AUC=0.99

Figure 3.3: Comparisons of ROC curves for the selected models in the ESCC data set by the sequentially selected order. Model 1 includes Age and Gender feature, and the following features are sequnatially added to the model: *miR-4783-3p, miR-320b, miR-1225-3p, miR-6789-5p.*

3.6.3 Neurobehavioral Impairment from Total Sleep Deprivation

In this study we aim to explore gene expression biomarker candidates for neurobehavioral impairment from total sleep deprivation. Specifically, using a clinical study, we investigate the role of genes on the Total Sleep Deprivation (TSD) and we use these biomarkers, that is, gene expressions, to predict TSD and estimate their effect on the development of TSD.

To perform analysis, data was obtained from NCBI GEO online repository, accession GSE98582. Blood samples were obtained from 17 healthy adults (ages 22–37, 7 females) who were not using drugs. Subjects remained in the sleep laboratory at the Sleep and Performance Research Center of Washington State University (Spokane, WA) for six consecutive nights. Meals were semi-standardized with selection from among a limited number of menu options; blood draws were performed immediately prior to meals. Blood samples were collected with an intravenous catheter approximately every 4 h during time awake on days two, four, and six. At each of the 12 timepoints, 2.5 mL blood was collected in a PAXgene™ Blood RNA tube, and the number of lapses per test bout was recorded from a 10 min PVT assay. Overall, the dataset contains 555 samples and 8284 gene features.

We define the binary outcome variable to be 1 if a sample corresponds to the case when TSD is observed and 0 otherwise. Total, 342 samples with TSD and 213 controls (without TSD) were taken. Further, we split the dataset into training and testing sets in order to perform the data analysis. To preserve the underlying distribution of the response variable, a stratified sampling technique was implemented. We kept 70% of data in the training set (389 samples with 8284 features) and the remaining 30% (166 samples with 8284 features) was used for model validation.

With $\eta_1 = 0.5$ and $\eta_2 = 3$ that were chosen from 5-fold cross-validation, the STEPWISE procedure recruited five genes. Particularly, *PF4V1*, *USP32P1*, *EMR1*, *NBR2*, and *DUSP23* were selected among 8284 genes. In addition, our procedure was applied to identify a final model for predicting TSD based on gene biomarkers. As a result, the following model was produced:

$\text{logit}^{-1}(-322.02 + 13.01 \times PF4V1 - 9.96 \times USP32P1 + 15.17 \times EMR1 + 17.66 \times NBR2 + 15.34 \times DUSP23)$, where $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$.

In the testing dataset, the final model had an area under the receiver operating curve (AUC) of 0.997 and achieved an accuracy of 0.983, with a sensitivity and specificity of 0.991 and 0.972, respectively. To compare STEPWISE with other competing methods, we repeatedly applied the aforementioned sampling procedure and split the dataset into training and testing sets 100 times. We obtained MSPE, the average of accuracy, sensitivity, specificity, and AUC. Figure 3.4 reports the model size of each method. Again, we observe that although STEPWISE procedure selects fewer variables that other methods, it achieves comparable prediction accuracy, specificity, sensitivity, and AUC. The results are presented in the Table 3.7.

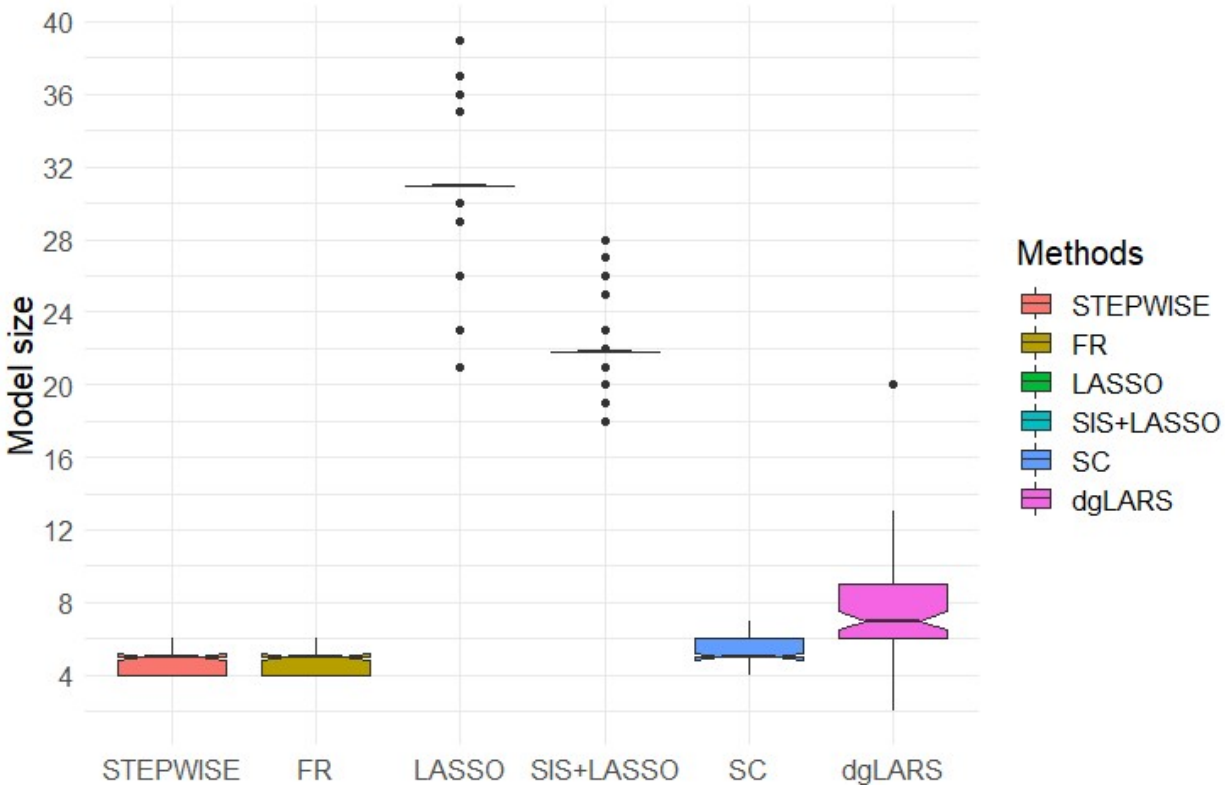


Figure 3.4: Box plot of model sizes for each method based on 100 total sleep deprivation training datasets. Performance of STEPWISE is reported with $\eta_1 = 0.5$ and $\eta_2 = 3$. Performance of SC and FR are reported with $\gamma = 0.5$

Table 3.7: Comparisons of competing methods over 100 independent splits of the Total Sleep Deprivation data into training and testing sets

Training set	MSPE	Accuracy	Sensitivity	Specificity	AUC
STEPWISE	0.02	0.98	0.98	0.97	0.99
SC	0.01	0.98	0.98	0.98	1.00
FR	0.02	0.98	0.98	0.97	0.99
LASSO	0.00	1.00	1.00	1.00	1.00
SIS+LASSO	0.00	1.00	1.00	1.00	1.00
dgLARS	0.07	0.91	0.92	0.89	0.95
Test set	MSPE	Accuracy	Sensitivity	Specificity	AUC
STEPWISE	0.04	0.97	0.96	0.94	0.98
SC	0.03	0.96	0.97	0.95	0.99
FR	0.04	0.97	0.96	0.94	0.98
LASSO	0.01	0.98	0.98	0.99	1.00
SIS+LASSO	0.01	0.99	0.99	0.98	1.00
dgLARS	0.08	0.88	0.90	0.86	0.95

Note: Values are averaged over 100 splits. STEPWISE was performed with $\eta_1 = 0.5$ and $\eta_2 = 3$. SC and FR were performed with $\gamma = 0.5$. The regularization parameters in LASSO and dgLARS were selected to minimize the BIC.

CHAPTER 4

MULTI-STAGE HYBRID MACHINE LEARNING METHOD

4.1 Machine Learning Ensemble Methods: Categories and Types

In Machine Learning, ensemble methods are used to achieve better predictive performance by combining predictions from multiple models instead of using a single model [139]. These methods tend to provide better results when the models used in ensemble methods are significantly diverse [140, 141]. Ensemble methods are mainly divided into two categories: sequential ensemble techniques and parallel ensemble techniques. Former generates base-learners (each method used in the model) in a sequence making them dependent on one another. The model performance tends to improve by assigning higher weights to previously misrepresented learners. In contrast, parallel ensemble techniques generate base learners in a parallel. This is done in order to introduce independence among base learners, which significantly reduces the error due to averaging the results obtained from base learner models.

Besides being divided into categories, ensemble methods can be distinguished by their types. The most popular and well-known types are Bagging, Boosting, and Stacking methods. Bagging methods train each base learner on a different sample of a training dataset (normally, these are bootstrap samples taken from the original training dataset). Predictions made by each of ensemble members are then combined by averaging the results, which is done to incorporate all possible outcomes of the prediction and randomize the outcome [142]. In order to improve the predictive power of the model, boosting ensemble technique learns from mistakes made by previous predictors. It adds predictors to the model sequentially, where successor predictors correct mistakes of the preceding predictors [143]. The gradient decent algorithm is used to identify points that need improvement the most.

Finally, Stacking technique (also known as stacked generalization) trains a learning algorithm to combine predictions of multiple other learning algorithms. It makes a final prediction by using

the predictions made by other algorithms, that is, output values of these methods become the input values of the stacked model. In this chapter, we will utilize parallel ensemble techniques and improve the performance of the STEPWISE algorithm. All methods used in the final model are discussed in subsequent sections.

4.2 A Review on Existing Methods

In this section we propose and describe methods included in the multi-stage hybrid machine learning model. They can be divided into two groups: model-based and model-free methods. Model-based methods specifically define the relationship between the response and explanatory variables (also known as predictors) via a certain link function. These models are considered to have a mathematical structure and involve various parameters that need to be estimated based on observed data. In addition, these models are accompanied by a set of statistical assumptions such as an underlying distribution of the response variable, relationship among predictors (mainly concerning their independence), variability of data, and etc.

It becomes crucial to satisfy those assumptions as it guarantees the reliability of results. Thus, implementation of model-based methods should be done by carefully examining and confirming validity of the model assumptions and choosing the appropriate link function. Due to straightforward interpretation and relatively small model complexity, model-based methods gained popularity among practitioners. We selected least absolute shrinkage and selection operator (LASSO) and our proposed STEPWISE method to represent model-based methods in the model.

In contrast, model-free methods do not make any assumptions on the parametric form of the underlying model explicitly. In other words, they adapt to the data characteristics without pre-specified model structure. Model-free algorithms are designed to automatically learn, adjust their actions and improve results with minimal or no human intervention. These algorithms help to gain some insights from data and enable to build right predictions and minimize chances of making any kind of errors. Given complex data, model-free methods are able to construct non-parametric representations (also known as non-parametric models). These methods develop their models based on

constant learning and retraining.

Normally, model-free methods are used in problems where mathematical models are unavailable or hard to construct. Therefore, almost all model-free methods have some optimization techniques at their core. For instance, gradient decent optimization algorithm is commonly used in many non-parametric methods. As a result, model-free methods tend to achieve high accuracy in their predictions and are successfully used with complex data. We selected several such methods to include in our model. Specifically, random forest (RF), support vector machine (SVM), extreme gradient boosting machine (XGBoost), and artificial neural network(ANN).

These methods are just selective examples of many other methods that are currently available. We decided to include this particular set of methods in our model because it is diverse in its nature and these methods are applicable in various real-life scenarios. For instance, random forest reduces drawback of large variance and is not prone to overfit the model; extreme gradient boosting machine provide lots of flexibility, can optimize on different loss functions and applicable to case with low variance and high bias; support vector machine is more effective in high dimensional spaces and relatively memory efficient; least absolute shrinkage and selection operator performs both automated variable selection and regularization, and helps minimize the impact of multicollinearity among predictors; artificial neural network can handle comprehensive data structures due to its complexity. All these methods are described in subsequent sections.

4.2.1 Random Forest (RF)

Random forests combine tree predictors that depend on values of a random vector sampled independently and identically for all trees in the forest. This methodology was proposed by Breiman [144] and quickly gained popularity among researchers and practitioners due to its simplicity and high accuracy.

Random Forest is a generic method, but mostly has been used with classification trees. Random Forests grow multiple classification trees, and classify a new object from an input vector by putting it down each of the trees in the forest. Then each tree gives a classification, and the majority of

"votes" determines a class of a prediction. More specifically, each tree in random forest grows as follows. If a number of observations in the training set is n, it takes a sample of n observations at random with replacement, from the original data. This creates a training set for growing the tree. Then, if there are M input predictors, a number $m < M$ is specified such that at each node, m variables are selected at random out of the M and the best split on these m divides the node. The value of m is being held constant during the forest growing. Then each tree grows to the largest extent possible with no pruning applied.

Mainly, an error rate produced by random forests depends on two factors: First, increasing a correlation between any two trees in the forest increases the error rate. Second, increasing the strength of the individual trees decreases the forest error rate. Additionally, the error rate can be controlled by manipulating an m parameter. Reducing m reduces both the correlation and the strength and increasing it increases both. The advantages of using Random Forest is the ability to cope with thousands of features without variable deletion, provide variable importance assessment in the classification, generate an internal unbiased estimate of the generalization error, and have methods for balancing error in class population unbalanced data sets.

4.2.2 Support Vector Machines (SVM)

In essence, the Support Vector Machine is a method proposed by Vapnik [145] that conceptually implements the following idea: it takes input vectors and maps them non-linearly to a very high dimension feature space. Then in this feature space it constructs a linear decision surface. Special properties of the decision surface guarantees high generalization ability of the learning method. SVMs can handle any number of classes, as well as observations of any dimension and can take almost any shape including linear, radial, and polynomial, among others. Particularly, SVMs construct a hyperplane or a set of hyperplanes, that is decision boundaries, in a high- or infinite-dimensional space, which can be used for classification, regression, and other type of problems. Good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class, also known as support vectors.

If classes are linearly separable, Hard Margin Classifier (HMC) can be implemented. HMC finds an optimal hyperplane, that separates classes while maximazing the distance to the closest points from the classes. The maximized distance is referred to as the Margin. HMC estimates the coefficients of hyperplanes by solving a quadratic programming problem with linear inequality constrains. If a perfect linear separation is not achievable or desirable, a Soft Margin Classifier (SMC) can be considered. While the data can be still linearly separable, the decision boundaries obtained using the HMC might not generalize well to new data and accuracy will suffer. To solve this issue, SMC loosens the constrains and allows some points to be wrongly classified. The set of points is called allowable budget.

Finally, if classes are not linearly separable, Support Vector Machine projects data to higher dimensions, where they are linearly separable and constructs a hyperplane. Then it transforms this hyperplane back to the initial space and obtains a non-linear decision boundary. It is achieved by using a kernel trick that computes a dot product in some feature space without even knowing what the space is and what is a mapping function. Most commonly employed kernel functions are linear, polynomial, and radial basis functions. The advantages of Support Vector Machine are that it always guarantees to find a global optimum as it just solves convex optimization problem, relatively robust to outliers (soft margin), and is flexible (implements various kernel functions). The main drawback of SVM is that it slows down the training process as data become taller (when the number of observations is significantly greater than the number of predictors) as it has to estimate parameters for each row.

4.2.3 Gradient Boosting Machine (GBM)

Gradient Boosting Machines, proposed by Friedman [146], quickly gained popularity due to their high accuracy and effectiveness in solving complex problems. Typically, it is hard for other methods to outperform the performance of GBMs and it is the algorithm of choice for many teams of machine learning competitions. GBMs build an ensemble of shallow trees in sequence where each tree learns and improves on the previous one. Even though shallow trees by themselves are

more of weak predictive models, they are able to boost and produce a powerful committee.

The main idea of boosting algorithms is to add new models to the ensemble sequentially. It starts with a weak tree and sequentially continues to build new trees, where each new tree in the sequence fixes up where the previous one made the mistakes (for instance, each new tree in the sequence focuses on the training rows where the previous tree had the largest prediction errors). Specifically, at any instant the model outcomes are weighed according to the outcomes of the previous instant. The outcomes that are predicted correctly are given a lower weight and the ones that are miss-classified are given higher weights.

Gradient Boosting Machines are considered a gradient decent algorithm. Gradient descent is a generic optimization algorithm that is capable of finding optimal solutions to a wide range of problems and can be used on any loss function that is differentiable. The fundamental idea of gradient descent is to search parameter values iteratively that will minimize a loss function. Here it is used to estimate the weights assigned to correctly and incorrectly predicted outcomes. Unlike bagging algorithms, GBM deals with bias variance trade-off by controlling both bias and variance and is proven to be more effective when applied to models with high bias and low variance.

There are various versions of Gradient Boosting Machine. Particularly useful are Stochastic GBM and Extreme GBM. Stochastic GBM takes a random subsample of the training dataset that offers additional reduction in tree correlation which improves a prediction accuracy. Extreme GBM is an optimized distributed gradient boosting machine that improves the accuracy and speed of the method by employing parallelism in its algorithm and adding regularization parameters to the model. The main disadvantage of GBM method is that it is a complex and less intuitive algorithms. In addition, it is time and computationally expensive method.

4.2.4 Artificial Neural Network (ANN)

An important subfield of Machine Learning is Deep Learning, which focuses on building predictive models based on artificial neural networks with two or more hidden layers. Artificial Neural Networks (ANN), first proposed by McCullough and Pitts [147], a model structure used in most

of the Deep Learning models, is inspired by the biological neural networks and mimics the way humans gain certain types of information through a combination of data inputs, weights, and bias. Like other machine learning algorithms, neural networks perform learning by mapping features to targets through a process of simple data transformations and feedback signals. Fundamental to most of the deep learning methods is the feedforward ANN. Feedforward ANNs are densely connected layers where inputs influence each successive layer which then influences the final output layer. Basic neural networks have three layers: an input layer, a hidden layer, and an output layer.

The input layer consists of all of the original input features. Most of the learning happens in the hidden layer, and the output layer produces the final predictions. The layers and nodes are the building blocks of our ANN and they decide how complex the network will be. Layers are called dense if all the nodes in successive layer are connected. Consequently, the more layers and nodes you add the more opportunities you create for new features to be learned.

There is no unique approach for determining the number of hidden layers and nodes; basically, these are the first hyperparameters among many others to tune. Mainly, features in your data largely determine the number of nodes you define in these hidden layers. The modeling task drives the choice of output layer. For regression problems, the output layer contains just one node that outputs the final prediction. If you are predicting a binary output, your output layer will still contain only one node and that node will predict the probability of success. Finally, if you predict an output with several classes, the output layer will contain the same number of nodes as the number of classes.

A crucial component of artificial neural networks is activation. Each node in ANN is connected to all the nodes in the previous layer. Each connection gets a weight and then that node adds all the incoming inputs multiplied by its corresponding connection weight plus an extra bias parameter. This summation becomes an input to an activation function. The activation function is a mathematical function that determines whether to fire a signal to the next layer.

On the forward pass, the ANN will select a batch of observations, randomly assign weights across all the node connections, and predict the output. Then, it assesses its own accuracy and adjusts the weights across all the node connections in order to improve the accuracy. This process is called

backpropagation. To carry out backpropagation, first, you need to establish an objective (loss) function to measure performance. On each forward pass the ANN will measure its performance based on the loss function chosen. The ANN will then work backwards through the layers, compute the gradient of the loss with regards to the network weights, adjust the weights a little in the opposite direction of the gradient, grab another batch of observations to run through the model, and repeat until the loss function is minimized. The performance of ANN can be optimized by tuning its hyperparameters. It can be done through adjusting model capacity (layers and modes), adding batch normalization, adjusting learning rate, trying out different activation functions and so on. Another possible way of improving ANN's performance is placing constraints on a model's complexity with regularization, also referred to as dropout implementation.

4.2.5 Least Absolute Shrinkage and Selection Operator (LASSO)

Nowadays, data sets typically contain a large number of features. As the number of features grows, certain assumptions required by traditional methods (e.g., linear models) break down and these models tend to overfit the data, causing the out of sample error to increase and making the results unreliable. One possible solution is to use Regularization methods, which constrain or regularize the estimated coefficients and can reduce the variance and uncertainty in the estimation.

As it was mentioned, having a large number of features invites various issues in using classic regression models. For instance, the model becomes much less interpretable, there could be infinite number of estimates for the model coefficients, and the predictors are likely to be highly correlated, which can invite multicollinearity issues. The regularized techniques constrain the total size of all the coefficient estimates that helps to reduce the magnitude and fluctuations of the coefficients and will reduce the variance of the model.

Arguably, one of the most well-known and frequently used regularized method is the Least Absolute Shrinkage and Selection Operator (LASSO). The method was poroposed by Tibshirani [84] and it minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint, it pushes coefficients all the way

to zero and produces some coefficients that are exactly 0, which eventually provides interpretable models. The LASSO method can be summarized as follows

Suppose we have data $(\mathbf{X}_i, y_i)$, $i = 1, 2, \ldots, N$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ are the predictors and y_i is the response variable. We assume that either the observations are independent or the responses are conditionally independent of the predictors. Finally, we assume that all predictors are standardized. Letting $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$, the LASSO estimate $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is defined as

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin}\left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j X_{ij} \right)^2 \right\}, \qquad \sum_j |\beta_j| \leq \lambda,$$

where $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage applied to the estimates. The LASSO method provides properties of both automated feature selection and ridge regression, and it exhibits the stability of the latter one. The main disadvantage of the technique is that it achieves these results at the cost of producing biased estimates.

4.2.6 STEPWISE Method

STEPWISE method is a procedure proposed in this thesis and described in Chapter 3. The proposed method fits GLMs with ultrahigh-dimensional predictors. It starts with an empty set or pre-specified predictors, scans all features and sequentially selects features, and conducts backward elimination once the forward selection is completed. The forward selection steps recruit variables in an inclusive way by allowing some false positives for the sake of avoiding false negatives, while backward selection steps eliminate the potential false positives from the recruited variables.

STEPWISE algorithm embraces model selection and estimation, controls both false negatives and positives by using different stopping criteria in the forward and backward selection steps, yields consistent estimates, and accommodates a wide range of data types, such as binary, categorical, and count data. In addition, under a sparsity assumption of the true model, it can discover all of the relevant predictors within a finite number of steps, and can produce a final model in ultrahigh dimensional settings without applying a pre-screening step which may introduce unintended false negatives.

4.3 Multi-Stage Hybrid Machine Learning Method

4.3.1 Introduction

The proposed multi-stage hybrid machine learning method carries out a stacking technique. Stacking method is designed to boost predictive accuracy by blending the predictions of multiple machine learning models. Stacked generalization or stacked was proposed by Wolpert [148] and is widely used by other researchers and practitioners, [149, 150, 151].

Stacking is a technique in which the predictions produced by a collection of models are given as inputs to a second-level learning algorithm. This second-level algorithm is trained optimally to combine the model predictions and form a final set of predictions. Specifically, stacking method trains a new learning algorithm to combine predictions of several base-learners, also known as, individual models. First, base-learners are trained using the training data, then a combiner, called a super learner, is trained to make a final prediction based on the predictions of the base learners.

It is important that the dataset collected for the stacked model consists of out-of-sample model predictions. In other words, to obtain the prediction for a certain data point in the data set, the model parameters should be estimated on a training set which does not include that particular data point. This is normally achieved via K-fold cross-validation. The training data is split into almost equal K subsets and K versions of the model are trained, each on the data with a different subset removed. Thus, model predictions for the kth subset are produced from the model trained on a set that did not include that subset.

To set up a multi-stage hybrid machine learning method, the following steps are being completed. First, we specify a list of base learners and a super learner, also know as, a meta algorithm. We select Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting Machine (XGBoost), Least Absolute Shrinkage and Selection Operator (LASSO), Artificial Neural Network (ANN), and STEPWISE method as the base learners. Linear weighted summation combiner is being used as a super learner. Next, we train each of these base learners on the training data. Specifically, we employ K-fold cross-validation for each of the base learners and collect the cross-

validated class probabilities from them. The cross-validated class probabilities are combined to create a new feature matrix.

Finally, we use this new feature matrix to train the meta learning algorithm, which can be used to generate predictions on new data. In other words, output values of each base learner become input values for the super learner. To generate ensemble predictions, first we have to generate class probabilities for each of the base learners. Then feed these class probabilities into the super learner, which will generate the ensemble prediction. Algorithm 1 summarizes the proposed method and provides a high-level explanation.

Algorithm 1 MULTI-STAGE HYBRID MACHINE LEARNING METHOD

1. Set up the stacked model
    - Specify a list of M base learners with a determined set of model parameters
    - Specify a super learner algorithm
2. Train the model
    - Train each of the M base learners on the training set
    - Perform cross-validation technique on each of the base-learners and collect cross-validated class probabilities from each (denoted as $p_1, ..., p_M$)
    - Combine the N (the number of observations in the training set) cross-validated class probability values from each of the M base-learners into a new $N \times M$ feature matrix. This matrix, along with the original response vector (y), is called level - one data
    - Train the super learner on level - one data
3. Predict on new data
    - To generate stacked predictions, first generate class probabilities from the base-learners
    - Feed those class probabilities to a super learner and produce new final predictions

Our method is called hybrid, because it employs both model-free and model-based methods. And we call it multi-stage, because it consists of two stages: setting up and training base-learners, and training the super learner and generating class probabilities based on it. It is worth to mention that the stacked model works the best when a diverse set of methods is selected as base learners.

Therefore, our model includes both types of methods.

4.3.2 Algorithm

Define $D = \{(\mathbf{X}_i, Y_i),\ i = 1, \ldots, n\}$ a dataset (D is also referred to as level-0 data) consisting of a vector $\mathbf{X}_i$ representing the attribute values of the i-th instance, and Y_i representing the class value. Let A_j, $j = 1, \ldots, J$, be a base-learner algorithm, also known as, level-0 estimator. Given D dataset, we randomly split it into K almost equal parts, $D_1, \ldots, D_K$. Let D_k, $k = 1, \ldots, K$, and $D^{(-k)} = D \setminus D_k$ be the test and training sets for the kth fold of K-fold cross-validation. Given J base-learner algorithms, we separately invoke the jth algorithm on the data in the training set $D^{(-k)}$ to induce a model $M_j^{(-k)}$ for $j = 1, \ldots, J$. For each instance $(\mathbf{X}_i, Y_i) \in D_k$, let $g_j(\mathbf{X}_i)$ denote the class probability estimated by the model $M_j^{(-k)}$ for $\mathbf{X}_i$. At the end of the cross-validation process, after applying the testing dataset D_k for each $k = 1, \ldots, K$ to each $M_j^{(-k)}$ for $j = 1, \ldots, J$, the base-learner model class probabilities form a meta-instance $\Big(g_1(\mathbf{X}_i), \ldots, g_J(\mathbf{X}_i), Y_i\Big)$ with the output variable for the original instance. A new dataset

$$D_{CV} = \left\{ \Big(g_1(\mathbf{X}_i), \ldots, g_J(\mathbf{X}_i), Y_i\Big), i = 1, \ldots, N \right\}$$

is assembled, also known as, level-1 data. Note that the original $\mathbf{X}_i$ is replaced with the corresponding level-0 output vectors $\Big\{g_1(\mathbf{X}_i), \ldots, g_J(\mathbf{X}_i)\Big\}$.

Now, at the second stage, referred to as level-1 learning stage, we derive our final level-1 model, $\tilde{M}$, from D_{CV}. The level-1 model will be constructed of the following form:

$$\tilde{M}(\mathbf{X}) = \sum_{j=1}^{J} \alpha_j \times g_j(\mathbf{X}),$$

where $g_j(\mathbf{X})$ is the jth level-0 class probability and α_j is its corresponding weight. Values for α's are computed based on corresponding level-0 model performances and are derived as

$$\alpha_j = K_j^2 \Big/ \sum_{j=1}^{J} K_j^2,$$

where K is a Kappa coefficient defined as

$$K = \frac{\text{accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}},$$

where

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

and

$$\text{expected accuracy} = \left(\frac{TP + FP}{N} \times \frac{TP + FN}{N}\right) + \left(\frac{TN + FP}{N} \times \frac{TN + FN}{N}\right),$$

where TP, FP, TN, and FN are True Positives, False Positives, True Negatives, and False Negatives, respectively. Breiman, [152], suggests that non-negative constraint $\alpha_j \geq 0$ provides consistently good results. Now, to make the final prediction we use the models M_j for $j = 1, \ldots, J$ in conjunction with $\tilde{M}$. Given a new instance, $\mathbf{X}_l$ models M_j produce a vector $(g_1(\mathbf{X}_l), \ldots, g_J(\mathbf{X}_l))$. Then this vector is used as an input value to the level-1 model, $\tilde{M}$, whose output is the final prediction for that instance.

4.4 Application: Bladder Cancer Prediction

4.4.1 Data Description

To utilize the aforementioned method we obtained data from Usuba *et al.* [69]. The goal of building this model is to identify important miRNA biomarkers that have an impact on bladder cancer development and can help in early detection of the disease. Moreover, ensemble method will enhance the predictive power of the proposed STEPWISE method taken separately.

Data consists of 972 samples profiling 2565 miRNAs. Specifically, 392 serum samples were obtained from bladder cancer patients who were admitted or referred to the National Cancer Center Hospital (NCCH) between 2008 and 2016. A total of 580 serum samples from non-cancer individuals were collected from 2 independent cohorts: the first cohort included individuals whose

serum samples were collected and stored by the National Center for Geriatrics and Gerontology (NCGG) Biobank between 2010 and 2012 and the second cohort included volunteers aged over 35 years who were recruited from the Yokohama Minoru Clinic in 2015.

We defined the binary outcome variable to be 1 if the subject was a case and 0 otherwise. To proceed, we randomly divided samples into training and testing sets. Training set consists of 80 % of original data (310 samples with bladder cancer and 468 non-cancer controls) and the testing set consists of the remaining 20 % (82 samples with bladder cancer and 112 non-cancer controls). Table A.3 summarises characteristics of the samples used in the study.

4.4.2 Results

We perform data analysis in two parts. First, we employ STEPWISE procedure separately and evaluate its performance based on obtained results. Then, we implement multi-stage hybrid machine learning method and demonstrate its advantages over the former model. To begin with, we further split the training set into training and validation sets by using 5-fold cross-validation technique in order to identify the best configuration of η_1 and η_2 parameters. Specifically, we use a greed search approach and specify a set of values for each of these two parameters: η_1 is being searched on the grid $\{0, 0.25, 0.5, 0.75, 1\}$ and η_2 on $\{1, 2, 3, 3.5, 4, 4.5, 5\}$. The results from the cross-validation procedure are presented in Table 4.1. It shows that STEPWISE method with $\eta_1 = 0.5$ and $\eta_2 = 3$ performed the best, so this pair of values will be used further in analysis.

Next, we applied the proposed STEPWISE method to the training set with $n = 778$ and $p = 2565$. With no prior biological information available, we started with an empty set. In the forward step, STEPWISE selected *mir*-6087, *mir*-5100, *mir*-1914-3p, *mir*-6831-5p, *mir*-2110, *mir*-6717-5p, *mir*-1343-3p, *mir*-6069, *mir*-6780b-5p, *mir*-1343-5p miRNAs, and the backward step eliminated *mir*-6780b-5p, *mir*-1343-5p miRNAs. The STEPWISE procedure produced the following final predictive model:

$\text{logit}^{-1}(88.33 - 8.08 \times \textit{miR-6087} + 2.53 \times \textit{miR-5100} - 3.54 \times \textit{miR-1914-3p} + 1.22 \times \textit{miR-6831-5p} - 1.57 \times \textit{miR-2110} + 2.26 \times \textit{miR-6717-5p} - 2.51 \times \textit{miR-1343-3p} + 0.75 \times \textit{miR-6069}$, where $\text{logit}^{-1}(x) =$

$\exp(x)/(1+\exp(x))$.

Table 4.1: Results of the 5-fold cross-validation procedure for the STEPWISE method

		η_2						
		1	2	3	3.5	4	4.5	5
η_1	0	0.885	0.893	0.893	0.895	0.895	0.905	0.905
	0.25	0.885	0.893	0.893	0.895	0.895	0.905	0.905
	0.5	0.917	0.917	0.931	0.925	0.925	0.911	0.911
	0.75	0.900	0.900	0.895	0.895	0.883	0.879	0.879
	1	0.875	0.870	0.870	0.863	0.861	0.861	0.859

Values for η_1 and η_2 were searched on the grid $\{0, 0.25, 0.5, 0.75, 1\}$ and $\{1, 2, 3, 3.5, 4, 4.5, 5\}$, respectively. The optimal configuration of the parameters was discovered by comparing AUC-ROCs (area under the receiver operating curve). The pair of parameter values that maximized the AUC value was selected for further analysis.

In the testing dataset, the model had AUC of 0.92 and achieved an accuracy of 0.91, with sensitivity, specificity, and precision of 0.93, 0.86, and 0.90, respectively. Finally, we repeatedly applied the sampling procedure and split the data into the training and testing sets 100 times. We obtained the average accuracy, sensitivity, specificity, precision, and AUC. The results are presented in the Table 4.2.

Table 4.2: Assessment of the proposed STEPWISE procedure using the bladder cancer data set

	Accuracy	Sensitivity	Specificity	Precision	AUC
Training set	0.92	0.94	0.92	0.92	0.94
Testing set	0.91	0.93	0.90	0.89	0.92

Note: values of accuracy, sensitivity, specificity, precision, and AUC were averaged over 100 splits.

In order to develop the final multi-stage hybrid machine learning model, we first built the other base-learner models included in the stacked method. Specifically, RF, SVM, XGBoost, LASSO, and ANN. After carrying out 5-fold cross-validation procedure, the following sets of hyperparameters have been identified for each of these methods: 600 trees were selected for RF model along with

5 instances in the terminal nodes and 250 randomly selected predictors in each tree; SVM achieved its best results with parameter $C = 0.01$ and polynomial structure of the second order; XGBoost performed the best with 300 trees, tree depth and learning rate equal to 8 and 0.1, respectively, and minimum 5 samples in the terminal nodes; ANN constructed its model with two hidden layers having 70% and 35% of predictors on its layers, respectively, learning rate equal to 0.1 and dropout rate to be 0.6; LASSO picked its penalty parameter to be 0.0361. Results of their performances over training and testing sets are summarized in Table 4.3.

Weights, α's, for the super learner combiner are computed based on base-learners' performance achieved during the first stage of modeling. Particularly, 0.17, 0.17, 0.14, 0.18, 0.18, and 0.16 are weights assigned to STEPWISE, RF, SVM, XGBosst, ANN, and LASSO, respectively. Table 4.4 presents results obtained from evaluating the multi-stage hybrid machine learning model. It can be observed that hybrid model significantly improved the performance of STEPWISE method. In addition, it also outperformed other methods included in the model.

Finally, we performed sensitivity analysis to quantify the relationship between the model performance and the weights assigned to the base-learners. Mainly, we tried out 7 model settings with different weight configurations and compared them with our existing model. Specifically, we developed a model with equal weights assigned to each method and the remaining 6 models have a high weight of 0.8 assigned to one of the base-learners while keeping other weights equal to 0.04. The results are summarized in the Table 4.5 and illustrate the advantage of our model over other competing model settings. The proposed multi-stage hybrid model outperformed Models 2-7 in all evaluation metrics; Model 1 achieved comparable results as it was expected since assigned weighted were similar to ours.

Table 4.3: Comparison of base-learner methods included in the multi-stage hybrid machine learning model over 100 independent splits of the bladder cancer data into training and testing sets

Training set	Accuracy	Sensitivity	Specificity	Precision	AUC
RF	0.95	0.95	0.96	0.94	0.95
SVM	0.91	0.93	0.92	0.93	0.92
XGBoost	0.95	0.94	0.96	0.97	0.96
ANN	0.95	0.96	0.94	0.97	0.96
LASSO	0.94	0.93	0.95	0.94	0.95
Test set	Accuracy	Sensitivity	Specificity	Precision	AUC
RF	0.93	0.94	0.93	0.92	0.93
SVM	0.89	0.92	0.90	0.87	0.89
XGBoost	0.94	0.93	0.95	0.93	0.95
ANN	0.94	0.95	0.93	0.91	0.95
LASSO	0.92	0.91	0.92	0.89	0.94

Note: values of accuracy, sensitivity, specificity, precision, and AUC were averaged over 100 splits; RF - Random Forest; SVM - Support Vector Machine; XGBoost - Extreme Gradient Boosting Machine; ANN - Artificial Neural Network; LASSO - Least Absolute Shrinkage and Selector Operator

Table 4.4: Evaluation of the proposed multi-stage hybrid machine learning model with the bladder cancer data set

	Accuracy	Sensitivity	Specificity	Precision	AUC
Training set	0.99	1.00	0.99	0.99	1.00
Testing set	0.98	0.98	0.98	0.97	0.99

Note: values of accuracy, sensitivity, specificity, precision, and AUC were averaged over 100 splits.

Table 4.5: Comparison of various model configurations included in the sensitivity analysis

Training set	Accuracy	Sensitivity	Specificity	Precision	AUC
Model 1	0.98	0.98	0.97	0.96	0.97
Model 2	0.96	0.96	0.96	0.95	0.95
Model 3	0.93	0.94	0.94	0.92	0.93
Model 4	0.97	0.98	0.97	0.97	0.95
Model 5	0.97	0.97	0.96	0.97	0.95
Model 6	0.94	0.95	0.93	0.94	0.93
Model 7	0.94	0.94	0.93	0.94	0.93

Note: values of accuracy, sensitivity, specificity, precision, and AUC were averaged over 100 splits; Model 1 corresponds to the equal-weights scenario; the Model 2-7 correspond to the scenarios with a high weight of 0.8 assigned to one of the base-learners while keeping other weights equal to 0.04; a high weight was assigned to methods in the following order: Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting machine (XGBoost), Artificial Neural Network (ANN), least absolute shrinkage and selector Operator (LASSO), STEPWISE procedure

4.5 Web Application

An R-Shiny web application was developed to enable users employ the proposed multi-stage hybrid machine learning method in practice. The main goal of this app is to help users analyze their own data and build predictive models according to our algorithm. The web application can be accessed online at Multi-Stage_Hybrid_ML_Method. Specifically, it is aimed to solve classification problems and has the following features.

First, users will have an option to either upload their own data sets or use pre-built sets. Pre-built option includes two well-known data sets: Iris and Abalone. Once the data is uploaded/selected from the given options, users can split the data into training and testing sets. For instance, specifying validation split to be 0.8 will split data into train and test sets with 4:1 ratio. In addition, one can apply pre-processing steps to the data, which is an important part of data analysis. As of now, there are three options available: standardizing numerical features (making features have mean equal to zero and variance equal to 1), imputing missing values via K-Nearest Neighbor technique (a method that takes into account the relationship among predictors), and removing Zero and Near-zero variance variables (removes feature that have no effect/ minimal effect on the response

feature). Figure A.1 illustrates this step.

Next, users are offered to tune hyperparameters of the base learner methods. Almost all Machine learning methods have hyper-parameters that can be tuned. Tuning them will potentially improve the model performance and reduce chances of overfiting the model. In order to accomplish this task, the web application employs greed search technique: users will specify a set of values for each of the hyper-parameters provided in the menu bar. Two resampling methods are available: k-fold cross-validation and repeated k-fold cross-validation. Once all necessary items are selected, the web application will tune these parameters and will display numeric and visual results in "Numerical Results" and "Visualize Results" tabs respectively. Figure A.2 illustrates a tuning hyperparameters procedure for a Random Forest method.

Lastly, after the tuning hyperparameters step is complete and a set of values for these parameters is selected, users can proceed further and start building their final predictive model. At this step, they can indicate parameter values for each base-learner obtained from the previous step and set weights for them. The app will train the model and display the results, which include model performance metrics (computed for both training and testing sets), feature importance, and final predictions. Figure A.3 illustrates an output results for the final predictive model including model evaluations for both training and testing sets.

CHAPTER 5

CONCLUSIONS, DISCUSSION, AND DIRECTIONS FOR FUTURE RESEARCH

In this thesis we have proposed to apply STEPWISE to produce final models in ultrahigh-dimensional settings, without resorting to a pre-screening step. We have shown that the method identifies or includes the true model with probability going to 1, and produces consistent coefficient estimates, which are useful for properly interpreting the actual impacts of risk factors. The theoretical properties of STEPWISE are established under mild conditions, which are worth discussing. Because in practice covariates are often standardized for various reasons, Condition (2) is assumed without loss of generality.

Conditions (3) and (4) are generally satisfied under common GLM models, including Gaussian, Binomial, Poisson, and Gamma distributions. Condition (5) is also often satisfied in practice. Proposition 2 in Zhang *et al.* [121] may be used as a tool to verify Condition (5) as well. Conditions (1) and (6) are in good faith with the unknown true model size $|\mathcal{M}|$ and minimum signal strength $n^{-\alpha}$ in practice. The "irrepresentable" condition (6) is strong and may not hold in some real datasets (see, e.g. [153, 154]). However, the condition holds under some commonly used covariance structures, including AR(1) and compound symmetry structure [153].

As shown in simulation studies and real data analyses, STEPWISE tends to generate models as predictive as the other well-known methods, with fewer variables (Figure 3.2). Parsimonious models are useful for biomedical studies as they explain data with a small number of important predictors, and offer practitioners a realistic list of biomarkers to investigate. With categorical outcome data frequently observed in biomedical studies (e.g. histology types of cancer), STEPWISE can be extended to accommodate multinomial classification, with more involved notation and computation. We will pursue this elsewhere.

As it was shown and discussed in the previous chapters of the thesis, STEPWISE procedure controls both false positives and false negatives in high-dimensional settings. It is achieved by employing different stopping criteria in the forward and backward selection steps that adds flexibility to our

algorithm. Mainly, versatility of the stopping criterion in the forward selection step allows to avoid false negatives by including some false positives in the early stages of the model building. While, using stopping criterion in the backward elimination step allows removing the potential false positives from the selected variables.

In addition, two extra parameters η_1 and η_2 involved in computing the stopping criteria determine how restrictive the variable screening process should be. Specifically, large values of η_1 in the forward selection step will recruit less variables and vice versa. Similarly, large η_2 values of the stopping criterion in the backward elimination step will remove more features. Thus, this framework can address different needs. For instance, if controlling false positives is the priority, then we recommend applying large values for parameters, and if it is more meaningful to control false negatives, then small values must be used. It is worth noting that our method includes forward selection as a special case when the parameter value is equal to 0, making it even more flexible.

Moreover, in this thesis we prove that, under a sparsity assumption of the true model, the proposed STEPWISE approach can discover all of the relevant predictors within a finite number of steps. Sparse models are common in high-dimensional settings. Among hundreds or thousands predictors involved in the model development, only a handful number of predictors have a significant relationship with the response variable. Including too many predictors in the model may result in overfitting, while keeping a few variables may lead to high bias and low predictive accuracy. Thus, identifying true signals and significant predictors correctly and including them in the final predictive model is a crucial step in a model building process.

Finally, we developed a multi-stage hybrid machine learning method to boost a predictive accuracy and improve a performance of the proposed method. It carries out stacking technique and combines model-free and model-based methods including the proposed STEPWISE method. Ting and Witten [155] suggested that the users of stacking method have a free choice of base-learner models. Therefore, we have selected heterogeneous machine learning methods (e.g., boosting, bagging, neural nets, and model-based methods) that have different strengths and disadvantages. Having a diverse set of base-learners makes our method applicable in various scenarios. In addition, they

claimed and demonstrated that successful stacked generalization implies using class probabilities rather than class predictions, and supported their claim with empirical examples. We also adopted this technique in our model.

Ueda [149] defined several combination types that can be used to combine base-learner outputs via weighted sum (WS), class-dependent weighted sum (CSW), and linear stacked generalization (LSG). Erdogan and Sen [156] showed that none of these methods is superior than others and the performance is data-driven and data-dependent. We have selected WS technique to be implemented in the super-learner method. A Kappa statistic (K) was used to estimate and assign weights to the individual base-learner outputs, which is considered to be more accurate metric for model evaluation [157, 158]. These weights reflect their performance on level-0 data: greater weights are assigned to base-learners with stronger performance and vice versa. This weights assignment method is believed to be more effective as it incorporates significance of each method included in the model [159]. Finally, Breiman [152] reported that non-negative constraint over the assigned weights will provide consistently good results. This constraint was added to our model as well.

The numerical examples we provided have vividly demonstrated an improved predictive power of the proposed method. Moreover, we performed sensitivity analysis to illustrate the superiority of the weight assignment technique used in the model over the other competing techniques. Lastly, we proposed and developed a web application that enables users employ the proposed multi-stage hybrid machine learning method in practice.

There are several open questions. First, in our numerical experiments, we used cross-validation to choose values for η_1 and η_2, which seemed to work well. However, more in-depth research is needed to find their optimal values to strike a balance between false positives and false negatives. Second, despite our consistent estimates, drawing inference based on them remains challenging. Statistical inference, which accounts for uncertainty in estimation, is key for properly interpreting analysis results and drawing appropriate conclusions. Our asymptotic results, nevertheless, are a stepping stone toward this important problem. Third, although the proposed STEPWISE procedure is designed to deal with the binary classification, it can be extended to accommodate multinomial

classification, a commonly observed problem in biological or biomedical research. Most multinomial classification methods rely on sequential binary classification by way of one-versus-all or direct pairwise comparison [160], which requires selecting a reduction method from multiclass to binary. Further investigation will be needed to identify such methods as it is not a trivial task and is on a case-by-case basis.

APPENDICES

APPENDIX A

SUPPLEMENT MATERIALS

A.1 Supplementary Materials

An R package, STEPWISE, was developed and is available at `https://github.com/AlexPijyan/STEPWISE`, along with the examples shown in the dissertation.

A.2 Additional Results in the Real Data Analysis

Table A.1: Comparison of genes selected by each competing method from the mammalian eye data set

	STEPWISE	FR	LASSO	SIS+LASSO	SC	dgLARS
STEPWISE	3	3	2	2	2	0
FR		4	2	2	2	0
LASSO			16	5	2	0
SIS+LASSO				9	2	0
SC					4	0
dgLARS						7

Note: Diagonal and off-diagonal elements of the table represent the model sizes for each method and the number of overlapping genes selected by the two methods corresponding to the row and column, respectively.

Table A.2: Selected miRNAs for ESCC training dataset

Methods	selected miRNAs
STEPWISE	*miR-4783-3p*; *miR-32*0b; *miR-1225-3p*
FR	*miR-4783-3p*; *miR-320b*; *miR-1225-3p*; *6789-5p*
SC	*miR-4783-3p*; *miR-320b*; *miR-1225-3p*; *6789-5p*
LASSO	*miR-6789-5p*; *miR-6781-5p*; *miR-1225-3p*; *miR-1238-5p*; *miR-320b*; *miR-6794-5p*; *miR-6877-5p*; *miR-6785-5p*; *miR-718*; *miR-195-5p*
SIS+LASSO	*miR-6785-5p*; *miR-1238-5p*; *miR-1225-3p*; *miR-6789-5p*; *miR-320b*; *miR-6875-5p*; *miR-6127*; *miR-1268b*; *miR-6781-5p*; *miR-125a-3p*
dgLARS	*miR-891b*; *miR-6127*; *miR-151a-5p*; *miR-195-5p*; ; *miR-3688-5p* *miR-125b-1-3p*; *miR-1273c*; *miR-6501-5p*; *miR-4666a-5p*; *miR-514a-3p*

Note: LASSO, SIS+LASSO, dgLARS selected 20, 17, and 33 miRNAs, respectively, and we only reported top 10 miRNAs.

Table A.3: Clinicopathologic characteristics of participants in bladder cancer study

Covariates	Training Set n_1 (%)	Testing set n_2 (%)
Bladder Cancer patients		
Total number of patients	310	82
Age, median (range)	68 (32-90)	70 (34-93)
Gender:		
Male	233 (74.9%)	54 (65.8%)
Female	77 (25.1%)	28 (34.2%)
Tumor Stage (%):		
<pT2	239 (77.09%)	62 (75.21%)
≥ pT2	71 (22.55%)	20 (23.93%)
Healthy Control patients		
Total number of patients	468	112
Age, median (range)	66.5 (35-90)	68.5 (41-92)
Gender:		
Male	241 (51.43%)	45 (40.17%)
Female	227 (48.57%)	67 (59.83%)

Table A.4: Clinicopathologic characteristics of study participants of the ESCC data set

Covariates	Training Set n_1 (%)	Testing set n_2 (%)
Esophageal squamous cell carcinoma (ESCC) patients		
Total number of patients	283	283
Age, median (range)	65 [40, 86]	67 [37, 90]
Gender:		
Male	224 (79.0%)	247 (87.3%)
Female	59 (21.0%)	36 (12.7%)
Stage:		
0	24 (8.5%)	27 (9.5%)
1	127 (44.9%)	128 (45.2%)
2	58 (20.5%)	57 (20.1%)
3	67 (23.7%)	61 (21.6%)
4	7 (2.4%)	10 (3.6%)
Non-ESCC Controls		
Total number of patients	283	4,682
Age, median (range)	68 [27, 92]	67.5 [20, 100]
Gender:		
Male	131 (46.3%)	2,086 (44.5%)
Female	152 (53.7%)	2,596 (55.5%)
Data sources of the controls:		
National Cancer Center Biobank (NCCB)	17 (6.0%)	306 (6.5%)
National Center for Geriatrics and Gerontology (NCGG)	158 (55.8%)	2,512 (53.7%)
Minoru clinic (MC)	108 (38.2%)	1,864 (39.8%)

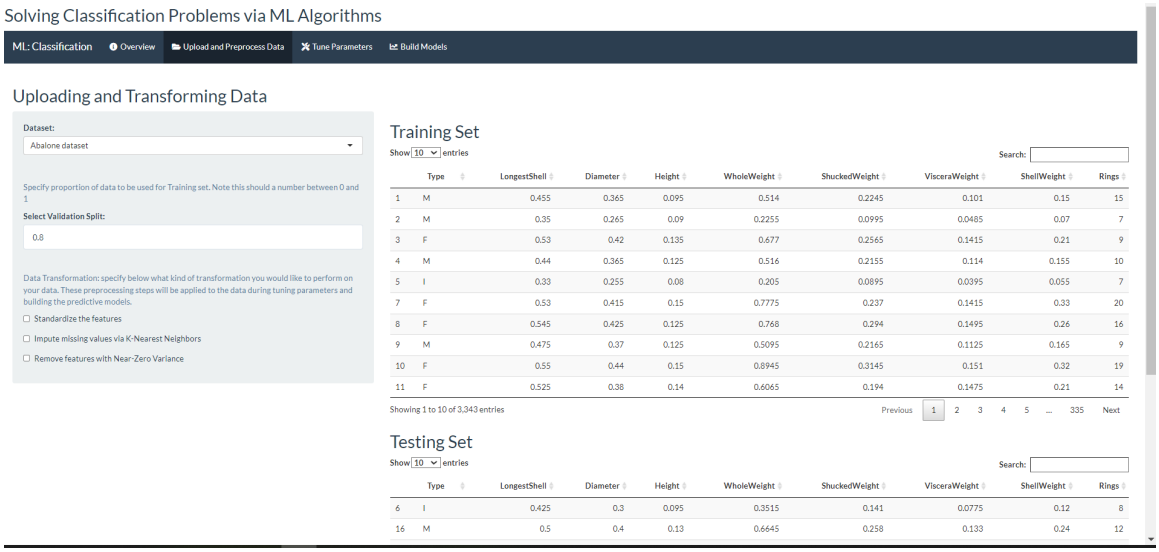


Figure A.1: R-Shiny Web Application for solving classification problems. The plot illustrates uploading and splitting a dataset into training and testing sets.

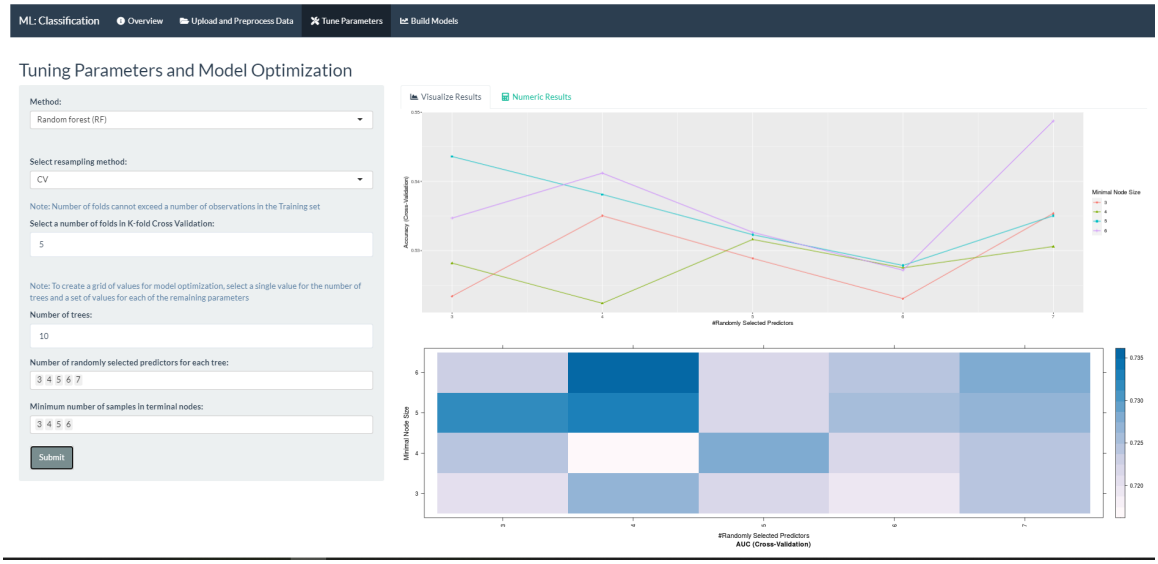


Figure A.2: R-Shiny Web Application for solving classification problems. The plot depicts a tuning parameters step for a Random Forest method.

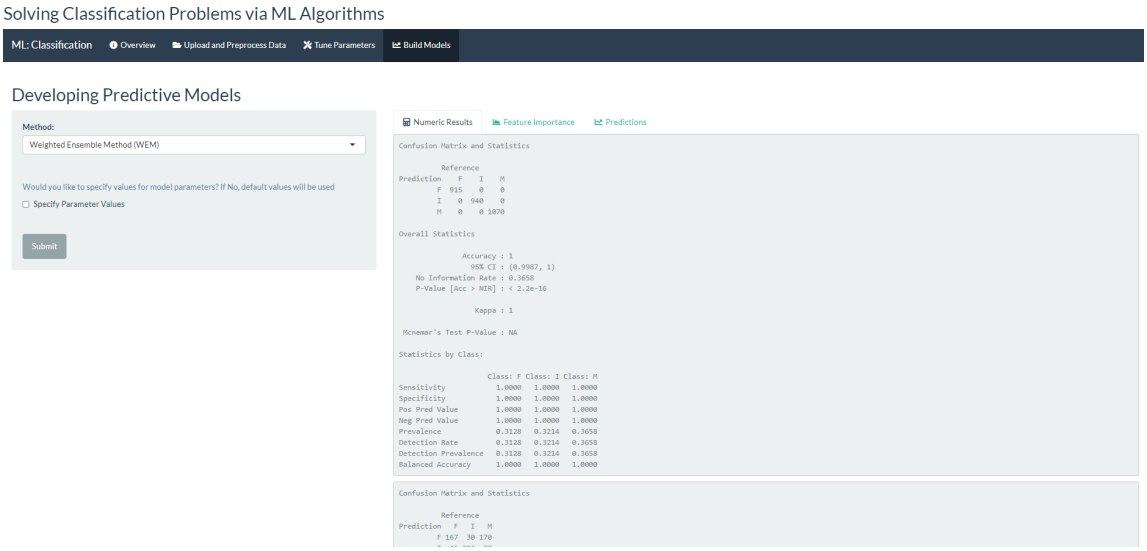


Figure A.3: R-Shiny Web Application for solving classification problems. The plot depicts an output of the final predictive model developed by the web application.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *science*, vol. 332, no. 6025, pp. 60–65, 2011.

[2] D. Reinsel, J. Gantz, and J. Rydning, "Data age 2025: The evolution of data to life-critical," *Don't Focus on Big Data*, vol. 2, 2017.

[3] M. Prosperi, J. S. Min, J. Bian, and F. Modave, "Big data hurdles in precision medicine and precision public health," *BMC Medical Informatics and Decision Making*, vol. 18(139), no. 1, 2018.

[4] J. Kang, J. Cho, and H. Zhao, "Practical issues in building risk-predicting models for complex diseases," *Journal of biopharmaceutical statistics*, vol. 20, no. 2, pp. 415–440, 2010.

[5] K. M. Corey, S. Kashyap, E. Lorenzi, S. A. Lagoo-Deenadayalan, K. Heller, K. Whalen, S. Balu, M. T. Heflin, S. R. McDonald, and M. Swaminathan, "Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (pythia): A retrospective, single-site study," *PLoS medicine*, vol. 15, no. 11, p. e1002701, 2018.

[6] P. Johansson, T. Jaarsma, G. Andersson, and J. Lundgren, "The impact of internet-based cognitive behavioral therapy and depressive symptoms on self-care behavior in patients with heart failure. a secondary analysis of a randomised controlled trial," *International Journal of Nursing Studies*, p. 103454, 2019.

[7] S. Mehta, A. Shelling, A. Muthukaruppan, A. Lasham, C. Blenkiron, G. Laking, and C. Print, "Predictive and prognostic molecular markers for cancer medicine," *Therapeutic advances in medical oncology*, vol. 2, no. 2, pp. 125–148, 2010.

[8] S. G. Baker and B. S. Kramer, "Simple methods for evaluating 4 types of biomarkers: Surrogate endpoint, prognostic, predictive, and cancer screening," *Biomarker Insights*, vol. 15, p. 1177271920946715, 2020.

[9] S. G. Baker, "Improving the biomarker pipeline to develop and evaluate cancer screening tests," *JNCI: Journal of the National Cancer Institute*, vol. 101, no. 16, pp. 1116–1119, 2009.

[10] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[11] I. Ruczinski, C. Kooperberg, and M. L. LeBlanc, "Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 178–195, 2004.

[12] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, and D. L. Willey, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature*, vol. 409, no. 6822, pp. 928–934, 2001.

[13] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, and R. A. Holt, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[14] A. J. Brookes, "The essence of snps," *Gene*, vol. 234, no. 2, pp. 177–186, 1999.

[15] S. Ye, S. Dhillon, X. Ke, A. R. Collins, and I. N. Day, "An efficient procedure for genotyping single nucleotide polymorphisms," *Nucleic acids research*, vol. 29, no. 17, pp. e88–e88, 2001.

[16] A.-C. Syvänen, "Accessing genetic variation: genotyping single nucleotide polymorphisms," *Nature Reviews Genetics*, vol. 2, no. 12, pp. 930–942, 2001.

[17] D. F. Easton, K. A. Pooley, A. M. Dunning, P. D. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field, and R. Luben, "Genome-wide association study identifies novel breast cancer susceptibility loci," *Nature*, vol. 447, no. 7148, pp. 1087–1093, 2007.

[18] A. Tenesa and M. G. Dunlop, "New insights into the aetiology of colorectal cancer from genome-wide association studies," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 353–358, 2009.

[19] G. C. Barnett, C. M. West, A. M. Dunning, R. M. Elliott, C. E. Coles, P. D. Pharoah, and N. G. Burnet, "Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype," *Nature Reviews Cancer*, vol. 9, no. 2, pp. 134–142, 2009.

[20] C. L. Sawyers, "The cancer biomarker problem," *Nature*, vol. 452, no. 7187, pp. 548–552, 2008.

[21] A. V. Kapp, S. S. Jeffrey, A. Langerød, A.-L. Børresen-Dale, W. Han, D.-Y. Noh, I. R. Bukholm, M. Nicolau, P. O. Brown, and R. Tibshirani, "Discovery and validation of breast cancer subtypes," *BMC genomics*, vol. 7, no. 1, pp. 1–15, 2006.

[22] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, and L. A. Akslen, "Molecular portraits of human breast tumours," *nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[23] G. Sotiropoulou, G. Pampalakis, E. Lianidou, and Z. Mourelatos, "Emerging roles of micrornas as molecular switches in the integrated circuit of the cancer cell," *Rna*, vol. 15, no. 8, pp. 1443–1461, 2009.

[24] C. Blenkiron and E. A. Miska, "Mirnas in cancer: approaches, aetiology, diagnostics and therapy," *Human molecular genetics*, vol. 16, no. R1, pp. R106–R113, 2007.

[25] E. S. Lander and N. J. Schork, "Genetic dissection of complex traits," *Science*, vol. 265, no. 5181, pp. 2037–2048, 1994.

[26] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, and I. Simon, "Transcriptional regulatory networks in saccharomyces cerevisiae," *science*, vol. 298, no. 5594, pp. 799–804, 2002.

[27] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, and G. Cavet, "Genetics of gene expression surveyed in maize, mouse and man," *Nature*, vol. 422, no. 6929, pp. 297–302, 2003.

[28] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung, "Genetic analysis of genome-wide variation in human gene expression," *Nature*, vol. 430, no. 7001, pp. 743–747, 2004.

[29] G. Yvert, R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak, "Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors," *Nature genetics*, vol. 35, no. 1, pp. 57–64, 2003.

[30] F. S. Collins, M. S. Guyer, and A. Chakravarti, "Variations on a theme: cataloging human dna sequence variation," *Science*, vol. 278, no. 5343, pp. 1580–1581, 1997.

[31] M. E. Zwick, D. J. Cutler, and A. Chakravarti, "Patterns of genetic variation in mendelian and complex traits," *Annual review of genomics and human genetics*, vol. 1, no. 1, pp. 387–407, 2000.

[32] A. M. Slavotinek, E. M. Stone, K. Mykytyn, J. R. Heckenlively, J. S. Green, E. Heon, M. A. Musarella, P. S. Parfrey, V. C. Sheffield, and L. G. Biesecker, "Mutations in mkks cause bardet-biedl syndrome," *Nature genetics*, vol. 26, no. 1, pp. 15–16, 2000.

[33] D. Y. Nishimura, C. C. Searby, R. Carmi, K. Elbedour, L. Van Maldergem, A. B. Fulton, B. L. Lam, B. R. Powell, R. E. Swiderski, K. E. Bugge, *et al.*, "Positional cloning of a novel gene on chromosome 16q causing bardet–biedl syndrome (bbs2)," *Human Molecular Genetics*, vol. 10, no. 8, pp. 865–874, 2001.

[34] K. Mykytyn, T. Braun, R. Carmi, N. B. Haider, C. C. Searby, M. Shastri, G. Beck, A. F. Wright, A. Iannaccone, and K. Elbedour, "Identification of the gene that, when mutated, causes the human obesity syndrome bbs4," *Nature genetics*, vol. 28, no. 2, pp. 188–191, 2001.

[35] E. M. Stone, A. J. Lotery, F. L. Munier, E. Héon, B. Piguet, R. H. Guymer, K. Vandenburgh, P. Cousin, D. Nishimura, and R. E. Swiderski, "A single efemp1 mutation associated with both malattia leventinese and doyne honeycomb retinal dystrophy," *Nature genetics*, vol. 22, no. 2, pp. 199–202, 1999.

[36] E. Héon, B. Piguet, F. Munier, S. R. Sneed, C. M. Morgan, S. Forni, G. Pescia, D. Schorderet, C. M. Taylor, and L. M. Streb, "Linkage of autosomal dominant radial drusen (malattia leventinese) to chromosome 2p16-21," *Archives of Ophthalmology*, vol. 114, no. 2, pp. 193–198, 1996.

[37] E. M. Stone, B. E. Nichols, A. E. Kimura, T. A. Weingeist, A. Drack, and V. C. Sheffield, "Clinical features of a stargardt-like dominant progressive macular dystrophy with genetic linkage to chromosome 6q," *Archives of Ophthalmology*, vol. 112, no. 6, pp. 765–772, 1994.

[38] R. P. Lifton, A. G. Gharavi, and D. S. Geller, "Molecular mechanisms of human hypertension," *Cell*, vol. 104, no. 4, pp. 545–556, 2001.

[39] D. Y. Nishimura, R. E. Swiderski, W. L. Alward, C. C. Searby, S. R. Patil, S. R. Bennet, A. B. Kanis, J. M. Gastier, E. M. Stone, and V. C. Sheffield, "The forkhead transcription factor gene fkhl7 is responsible for glaucoma phenotypes which map to 6p25," *Nature genetics*, vol. 19, no. 2, pp. 140–147, 1998.

[40] J. H. Fingert, E. Héon, J. M. Liebmann, T. Yamamoto, J. E. Craig, J. Rait, K. Kawase, S.-T. Hoh, Y. M. Buys, and J. Dickinson, "Analysis of myocilin mutations in 1703 glaucoma patients from five different populations," *Human molecular genetics*, vol. 8, no. 5, pp. 899–905, 1999.

[41] T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, V. C. Sheffield, and E. M. Stone, "Regulation of gene expression in the mammalian eye and its relevance to eye disease," *Proceedings of the National Academy of Sciences*, vol. 103, no. 39, pp. 14429–14434, 2006.

[42] A. P. Chiang, J. S. Beck, H.-J. Yen, M. K. Tayeh, T. E. Scheetz, R. E. Swiderski, D. Y. Nishimura, T. A. Braun, K.-Y. A. Kim, J. Huang, K. Elbedour, R. Carmi, D. C. Slusarski, T. L. Casavant, E. M. Stone, and V. C. Sheffield, "Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11)," *Proceedings of the National Academy of Sciences*, vol. 103, no. 16, pp. 6287–6292, 2006.

[43] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[44] M. Arnold, M. Laversanne, L. M. Brown, S. S. Devesa, and F. Bray, "Predicting the future burden of esophageal cancer by histological subtype: international trends in incidence up to 2030," *Official journal of the American College of Gastroenterology| ACG*, vol. 112, no. 8, pp. 1247–1255, 2017.

[45] B.-X. Li, Q. Yu, Z.-L. Shi, P. Li, and S. Fu, "Circulating microRNAs in esophageal squamous cell carcinoma: association with locoregional staging and survival," *International Journal of Clinical and Experimental Medicine*, vol. 8, no. 5, pp. 7241–7250, 2015.

[46] S. H. Lu, "Alterations of oncogenes and tumor suppressor genes in esophageal cancer in china: Molecular pathology and epidemiology," *Mutation research. Reviews in mutation research*, vol. 462, no. 2-3, pp. 343–353, 2000.

[47] S. He, J. Peng, L. Li, Y. Xu, X. Wu, J. Yu, J. Liu, J. Zhang, R. Zhang, and W. Wang, "High expression of cytokeratin CAM5.2 in esophageal squamous cell carcinoma is associated with poor prognosis," *Medicine*, vol. 98(37), no. 37, p. e17104, 2019.

[48] X. Qian, C. Tan, F. Wang, B. Yang, Y. Ge, Z. Guan, and J. Cai, "Esophageal cancer stem cells and implications for future therapeutics," *OncoTargets and therapy*, vol. 9, p. 2247, 2016.

[49] B. Zhang, X. Pan, G. P. Cobb, and T. A. Anderson, "Micrornas as oncogenes and tumor suppressors," *Developmental biology*, vol. 302, no. 1, pp. 1–12, 2007.

[50] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, and A. A. Ferrando, "Microrna expression profiles classify human cancers," *nature*, vol. 435, no. 7043, pp. 834–838, 2005.

[51] R. Hamano, H. Miyata, M. Yamasaki, K. Sugimura, K. Tanaka, Y. Kurokawa, K. Nakajima, S. Takiguchi, Y. Fujiwara, and M. Mori, "High expression of lin28 is associated with tumour aggressiveness and poor prognosis of patients in oesophagus cancer," *British journal of cancer*, vol. 106, no. 8, pp. 1415–1423, 2012.

[52] N. Liu, N.-Y. Chen, R.-X. Cui, W.-F. Li, Y. Li, R.-R. Wei, M.-Y. Zhang, Y. Sun, B.-J. Huang, and M. Chen, "Prognostic value of a microrna signature in nasopharyngeal carcinoma: a microrna expression analysis," *The lancet oncology*, vol. 13, no. 6, pp. 633–641, 2012.

[53] G. Gao, H. A. Gay, R. D. Chernock, T. R. Zhang, J. Luo, W. L. Thorstad, J. S. Lewis Jr, and X. Wang, "A microrna expression signature for the prognosis of oropharyngeal squamous cell carcinoma," *Cancer*, vol. 119, no. 1, pp. 72–80, 2013.

[54] H. Konishi, D. Ichikawa, S. Komatsu, A. Shiozaki, M. Tsujiura, H. Takeshita, R. Morimura, H. Nagata, T. Arita, and T. Kawaguchi, "Detection of gastric cancer-associated micrornas on microrna microarray comparing pre-and post-operative plasma," *British journal of cancer*, vol. 106, no. 4, pp. 740–747, 2012.

[55] Y. Tomimaru, H. Eguchi, H. Nagano, H. Wada, S. Kobayashi, S. Marubashi, M. Tanemura, A. Tomokuni, I. Takemasa, and K. Umeshita, "Circulating microrna-21 as a novel biomarker for hepatocellular carcinoma," *Journal of hepatology*, vol. 56, no. 1, pp. 167–175, 2012.

[56] N. H. Heegaard, A. J. Schetter, J. A. Welsh, M. Yoneda, E. D. Bowman, and C. C. Harris, "Circulating micro-rna expression profiles in early stage nonsmall cell lung cancer," *International journal of cancer*, vol. 130, no. 6, pp. 1378–1386, 2012.

[57] K. Sudo, K. Kato, J. Matsuzaki, N. Boku, S. Abe, Y. Saito, H. Daiko, S. Takizawa, Y. Aoki, H. Sakamoto, Y. Aoki, H. Sakamoto, S. Niida, F. Takeshita, T. Fukuda, and T. Ochiya, "Development and validation of an esophageal squamous cell carcinoma detection model by large-scale microrna profiling," *JAMA Network Open*, vol. 2, no. 5, pp. e194573–e194573, 2019.

[58] M. Burger, J. W. Catto, G. Dalbagni, H. B. Grossman, H. Herr, P. Karakiewicz, W. Kassouf, L. A. Kiemeney, C. La Vecchia, and S. Shariat, "Epidemiology and risk factors of urothelial bladder cancer," *European urology*, vol. 63, no. 2, pp. 234–241, 2013.

[59] M. Burger, W. Oosterlinck, B. Konety, S. Chang, S. Gudjonsson, R. Pruthi, M. Soloway, E. Solsona, P. Sved, and M. Babjuk, "Icud-eau international consultation on bladder cancer 2012: non–muscle-invasive urothelial carcinoma of the bladder," *European urology*, vol. 63, no. 1, pp. 36–44, 2013.

[60] M. Babjuk, M. Burger, R. Zigeuner, S. F. Shariat, B. W. Van Rhijn, E. Compérat, R. J. Sylvester, E. Kaasinen, A. Böhle, and J. P. Redorta, "Eau guidelines on non–muscle-invasive urothelial carcinoma of the bladder: update 2013," *European urology*, vol. 64, no. 4, pp. 639–653, 2013.

[61] A. B. Apolo, M. Milowsky, and D. F. Bajorin, "Clinical states model for biomarkers in bladder cancer," *Future Oncology*, vol. 5, no. 7, pp. 977–992, 2009.

[62] J. P. Stein, G. Lieskovsky, R. Cote, S. Groshen, A.-C. Feng, S. Boyd, E. Skinner, B. Bochner, D. Thangathurai, and M. Mikhail, "Radical cystectomy in the treatment of invasive bladder cancer: long-term results in 1,054 patients," *Journal of clinical oncology*, vol. 19, no. 3, pp. 666–675, 2001.

[63] H. von der Maase, L. Sengelov, J. T. Roberts, S. Ricci, L. Dogliotti, T. Oliver, M. J. Moore, A. Zimmermann, and M. Arning, "Long-term survival results of a randomized trial comparing gemcitabine plus cisplatin, with methotrexate, vinblastine, doxorubicin, plus cisplatin in patients with bladder cancer," *Journal of clinical oncology*, vol. 23, no. 21, pp. 4602–4608, 2005.

[64] P. Whelan, "Bladder cancer–contemporary dilemmas in its management.," *European urology*, vol. 53, no. 1, pp. 24–26, 2007.

[65] M. F. Botteman, C. L. Pashos, A. Redaelli, B. Laskin, and R. Hauser, "The health economics of bladder cancer," *Pharmacoeconomics*, vol. 21, no. 18, pp. 1315–1330, 2003.

[66] I. Osman, D. F. Bajorin, T.-T. Sun, H. Zhong, D. Douglas, J. Scattergood, R. Zheng, M. Han, K. W. Marshall, and C.-C. Liew, "Novel blood biomarkers of human urinary bladder cancer," *Clinical Cancer Research*, vol. 12, no. 11, pp. 3374–3380, 2006.

[67] A. Apolo, I. Osman, R. Shen, A. Olshen, and D. Bajorin, "Peripheral-blood gene expression profiling in bladder cancer (bc) patients (pts)," *Journal of Clinical Oncology*, vol. 26, no. 15_suppl, pp. 5076–5076, 2008.

[68] X. Zhang and Y. Zhang, "Bladder cancer and genetic mutations," *Cell biochemistry and biophysics*, vol. 73, no. 1, pp. 65–69, 2015.

[69] W. Usuba, F. Urabe, Y. Yamamoto, J. Matsuzaki, H. Sasaki, M. Ichikawa, S. Takizawa, Y. Aoki, S. Niida, and K. Kato, "Circulating mirna panels for specific and early detection in bladder cancer," *Cancer science*, vol. 110, no. 1, pp. 408–419, 2019.

[70] H. A. Uyhelji, D. M. Kupfer, V. L. White, M. L. Jackson, H. Van Dongen, and D. M. Burian, "Exploring gene expression biomarker candidates for neurobehavioral impairment from total sleep deprivation," *BMC genomics*, vol. 19, no. 1, pp. 1–17, 2018.

[71] E. L. Bliss, L. D. Clark, and C. D. West, "Studies of sleep deprivation—relationship to schizophrenia," *AMA Archives of Neurology & Psychiatry*, vol. 81, no. 3, pp. 348–359, 1959.

[72] P. L. Franzen, D. J. Buysse, R. E. Dahl, W. Thompson, and G. J. Siegle, "Sleep deprivation alters pupillary reactivity to emotional stimuli in healthy young adults," *Biological psychology*, vol. 80, no. 3, pp. 300–305, 2009.

[73] D. C. Baldwin Jr and S. R. Daugherty, "Sleep deprivation and fatigue in residency training: results of a national survey of first-and second-year residents," *Sleep*, vol. 27, no. 2, pp. 217–223, 2004.

[74] X. Liu, "Sleep and adolescent suicidal behavior," *Sleep*, vol. 27, no. 7, pp. 1351–1358, 2004.

[75] J. Lim and D. F. Dinges, "Sleep deprivation and vigilant attention," *Annals of the New York Academy of Sciences*, vol. 1129, no. 1, pp. 305–322, 2008.

[76] N. Goel, H. Rao, J. S. Durmer, and D. F. Dinges, "Neurocognitive consequences of sleep deprivation," in *Seminars in neurology*, vol. 29, pp. 320–339, © Thieme Medical Publishers, 2009.

[77] W. D. Killgore, "Effects of sleep deprivation on cognition," *Progress in brain research*, vol. 185, pp. 105–129, 2010.

[78] S. M. Doran, H. P. Van Dongen, and D. F. Dinges, "Sustained attention performance during sleep deprivation: evidence of state instability.," *Archives italiennes de biologie*, vol. 139, no. 3, pp. 253–267, 2001.

[79] S. K. Davies, J. E. Ang, V. L. Revell, B. Holmes, A. Mann, F. P. Robertson, N. Cui, B. Middleton, K. Ackermann, and M. Kayser, "Effect of sleep deprivation on the human metabolome," *Proceedings of the National Academy of Sciences*, vol. 111, no. 29, pp. 10761–10766, 2014.

[80] N. Goel, ""omics" approaches for sleep and circadian rhythm research: biomarkers for identifying differential vulnerability to sleep loss," *Current Sleep Medicine Reports*, vol. 1, no. 1, pp. 38–46, 2015.

[81] B. D. W. Group, A. J. Atkinson Jr, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, and R. T. Schooley, "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clinical pharmacology & therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[82] P. Van Dongen, M. D. Baynard, G. Maislin, and D. F. Dinges, "Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability," *Sleep*, vol. 27, no. 3, pp. 423–433, 2004.

[83] E. S. Arnardottir, E. V. Nikonova, K. R. Shockley, A. A. Podtelezhnikov, R. C. Anafi, K. Q. Tanis, G. Maislin, D. J. Stone, J. J. Renger, C. J. Winrow, and A. I. Pack, "Blood-gene expression reveals reduced circadian rhythmicity in individuals resistant to sleep deprivation," *Sleep*, vol. 37, no. 10, pp. 1589–1600, 2014.

[84] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[85] C. J. Flynn, C. M. Hurvich, and J. S. Simonoff, "On the sensitivity of the lasso to the number of predictor variables," *Statistical Science*, vol. 32, no. 1, pp. 88–105, 2017.

[86] S. A. van de Geer, "On the asymptotic variance of the debiased Lasso," *Electronic Journal of Statistics*, vol. 13, no. 2, pp. 2970–3008, 2019.

[87] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[88] E. Barut, J. Fan, and A. Verhasselt, "Conditional sure independence screening," *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 1266–1277, 2016.

[89] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[90] L. Augugliaro, A. M. Mineo, and E. C. Wit, "Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models," *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology*, vol. 75, pp. 471–498, 2013.

[91] H. Pazira, L. Augugliaro, and E. Wit, "Extended differential geometric lars for high-dimensional glms with general dispersion parameter," *Statistics and Computing*, vol. 28, no. 4, pp. 753–774, 2018.

[92] H. An, D. Huang, Q. Yao, and C.-H. Zhang, "Stepwise searching for feature variables in high-dimensional linear regression," *[Online]. Available: http://eprints.lse.ac.uk/51349/*, 2008.

[93] H. Wang, "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1512–1524, 2009.

[94] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[95] J.-S. Hwang and T.-H. Hu, "A stepwise regression algorithm for high-dimensional variable selection," *Journal of Statistical Computation and Simulation*, vol. 85, no. 9, pp. 1793–1806, 2015.

[96] J.-S. Hwang and T.-H. Hu, "Stepwise paring down variation for identifying influential multi-factor interactions related to a continuous response variable," *Statistics in Biosciences*, vol. 4, no. 2, pp. 197–212, 2012.

[97] T. M. Loughin, "A systematic comparison of methods for combining p-values from independent tests," *Computational statistics & data analysis*, vol. 47, no. 3, pp. 467–485, 2004.

[98] H. A. David, "Order statistics new york," 1981.

[99] Q. Zheng, H. G. Hong, and Y. Li, "Building generalized linear models with ultrahigh dimensional features: A sequentially conditional approach," *Biometrics*, vol. 76, no. 1, pp. 47–60, 2020.

[100] P. McCullagh, *Generalized Linear Models*. New York: Routledge, 1989.

[101] S. D. Zhao and Y. Li, "Principled sure independence screening for Cox models with ultra-high-dimensional covariates," *Journal of Multivariate Analysis*, vol. 105, no. 1, pp. 397–411, 2012.

[102] A. Gorst-Rasmussen and T. Scheike, "Independent screening for single-index hazard rate models with ultrahigh dimensional features," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 2, pp. 217–245, 2013.

[103] H. G. Hong, J. Kang, and Y. Li, "Conditional screening for ultra-high dimensional covariates with survival outcomes," *Lifetime data analysis*, vol. 24, no. 1, pp. 45–71, 2018.

[104] X. He, L. Wang, and H. G. Hong, "Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data," *The Annals of Statistics*, vol. 41, no. 1, pp. 342–369, 2013.

[105] R. Song, W. Lu, S. Ma, and X. Jessie Jeng, "Censored rank independence screening for high-dimensional survival data," *Biometrika*, vol. 101, no. 4, pp. 799–814, 2014.

[106] H. G. Hong, X. Chen, D. C. Christiani, and Y. Li, "Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes," *Biometrics*, vol. 74, no. 2, pp. 421–429, 2018.

[107] J. Li, Q. Zheng, L. Peng, and Z. Huang, "Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes," *Biometrics*, vol. 72, no. 4, pp. 1145–1154, 2016.

[108] H. G. Hong, Q. Zheng, and Y. Li, "Forward regression for Cox models with high-dimensional covariates," *Journal of Multivariate Analysis*, vol. 173, pp. 268–290, 2019.

[109] M.-Y. Cheng, T. Honda, and J.-T. Zhang, "Forward variable selection for sparse ultra-high dimensional varying coefficient models," *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 1209–1221, 2016.

[110] S. Luo and Z. Chen, "Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space," *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 1229–1240, 2014.

[111] C. T. Volinsky and A. E. Raftery, "Bayesian information criterion for censored survival models," *Biometrics*, vol. 56, no. 1, pp. 256–262, 2000.

[112] R. Xu, F. Vaida, and D. P. Harrington, "Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models," *Statistica Sinica*, vol. 19, no. 2, p. 819, 2009.

[113] C.-K. Ing and T. L. Lai, "A stepwise regression method and consistent model selection for high-dimensional sparse linear models," *Statistica Sinica*, vol. 21, pp. 1473–1513, 2011.

[114] P. Bühlmann and B. Yu, "Boosting with the l 2 loss: regression and classification," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[115] J. Chen and Z. Chen, "Extended BIC for small-*n*-large-*P* sparse GLM," *Statistica Sinica*, vol. 22, no. 2, pp. 555–574, 2012.

[116] P. Bühlmann and B. Yu, "Sparse boosting," *Journal of Machine Learning Research*, vol. 7, pp. 1001–1024, 2006.

[117] S. A. van de Geer, "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, vol. 36, no. 2, pp. 614–645, 2008.

[118] Y. Fan and C. Y. Tang, "Tuning parameter selection in high dimensional penalized likelihood," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 3, pp. 531–552, 2013.

[119] M. Kwemou, "Non-asymptotic oracle inequalities for the Lasso and group Lasso in high dimensional logistic model," *ESAIM: Probability and Statistics*, vol. 20, pp. 309–331, 2016.

[120] Y. Jiang, Y. He, and H. Zhang, "Variable selection with prior information for generalized linear models via the prior LASSO method," *Journal of the American Statistical Association*, vol. 111, no. 513, pp. 355–376, 2016.

[121] C.-H. Zhang and J. Huang, "The sparsity and bias of the Lasso selection in high-dimensional linear regression," *The Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[122] J. Fan and R. Song, "Sure independence screening in generalized linear models with np-dimensionality," *The Annals of Statistics*, vol. 38, no. 6, pp. 3567–3604, 2010.

[123] S. Luo, J. Xu, and Z. Chen, "Extended Bayesian information criterion in the Cox model with a high-dimensional feature space," *Annals of the Institute of Statistical Mathematics*, vol. 67, no. 2, pp. 287–311, 2015.

[124] A. Van der Vaart and J. Wellner, "Weak convergence, in 'weak convergence and empirical processes'," 1996.

[125] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *The Annals of Probability*, vol. 22, no. 1, pp. 28–76, 1994.

[126] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[127] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[128] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics*, vol. 5, no. 1, pp. 232–253, 2011.

[129] X. Wang and C. Leng, "R package: screening." `https://github.com/wwrechard/screening`, 2016.

[130] L. Augugliaro, A. M. Mineo, and E. C. Wit, "dglars: An R package to estimate sparse generalized linear models," *Journal of Statistical Software*, vol. 59, no. 8, pp. 1–40, 2014.

[131] Y. Zhang, "Epidemiology of esophageal cancer," *World Journal of Gastroenterology: WJG*, vol. 19, no. 34, pp. 5598–5606, 2013.

[132] L. N. Mathieu, N. F. Kanarek, H.-L. Tsai, C. M. Rudin, and M. V. Brock, "Age and sex differences in the incidence of esophageal adenocarcinoma: results from the Surveillance, Epidemiology, and End Results (SEER) registry (1973–2008)," *Diseases of the Esophagus*, vol. 27, no. 8, pp. 757–763, 2014.

[133] J. Zhou, M. Zhang, Y. Huang, L. Feng, H. Chen, Y. Hu, H. Chen, K. Zhang, L. Zheng, and S. Zheng, "Microrna-320b promotes colorectal cancer proliferation and invasion by competing with its homologous microrna-320a," *Cancer Letters*, vol. 356, no. 2, pp. 669–675, 2015.

[134] V. Lieb, K. Weigelt, L. Scheinost, K. Fischer, T. Greither, M. Marcou, G. Theil, H. Klocker, H.-J. Holzhausen, X. Lai, J. Vera, A. Ekici, W. Horninger, P. Fornara, B. Wullich, H. Taubert, and S. Wach, "Serum levels of mir-320 family members are associated with clinical parameters and diagnosis in prostate cancer patients," *Oncotarget*, vol. 9, no. 12, pp. 10402–10416, 2018.

[135] L. E. Mullany, J. S. Herrick, R. K. Wolff, J. R. Stevens, and M. L. Slattery, "Association of cigarette smoking and microrna expression in rectal cancer: insight into tumor phenotype," *Cancer Epidemiology*, vol. 45, pp. 98–107, 2016.

[136] H. Zheng, F. Zhang, X. Lin, C. Huang, Y. Zhang, Y. Li, J. Lin, W. Chen, and X. Lin, "Microrna-1225-5p inhibits proliferation and metastasis of gastric carcinoma through repressing insulin receptor substrate-1 and activation of β-catenin signaling," *Oncotarget*, vol. 7, no. 4, pp. 4647–4663, 2016.

[137] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[138] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[139] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[140] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning*, vol. 51, no. 2, pp. 181–207, 2003.

[141] P. S. A. Krogh *et al.*, "Learning with ensembles: How over-fitting can be useful," in *Proceedings of the 1995 Conference*, vol. 8, p. 190, 1996.

[142] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[143] R. E. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, no. 2, pp. 197–227, 1990.

[144] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[145] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[146] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[147] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[148] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[149] N. Ueda, "Optimal linear combination of neural networks for improving classification performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 2, pp. 207–215, 2000.

[150] N. Rooney, D. Patterson, and C. Nugent, "Non-strict heterogeneous stacking," *Pattern recognition letters*, vol. 28, no. 9, pp. 1050–1061, 2007.

[151] B. Zenko, L. Todorovski, and S. Dzeroski, "A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods," in *Proceedings 2001 IEEE International Conference on Data Mining*, pp. 669–670, IEEE, 2001.

[152] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, no. 1, pp. 49–64, 1996.

[153] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, no. Nov, pp. 2541–2563, 2006.

[154] P. Bühlmann and S. Van De Geer, *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

[155] K. M. Ting and I. H. Witten, "Issues in stacked generalization," *Journal of artificial intelligence research*, vol. 10, pp. 271–289, 1999.

[156] H. Erdogan and M. U. Sen, "A unifying framework for learning the linear combiners for classifier ensembles," in *2010 20th International Conference on Pattern Recognition*, pp. 2985–2988, IEEE, 2010.

[157] D. V. Cicchetti and A. R. Feinstein, "High agreement but low kappa: Ii. resolving the paradoxes," *Journal of clinical epidemiology*, vol. 43, no. 6, pp. 551–558, 1990.

[158] R. J. Cook, "Kappa and its dependence on marginal rates," *Wiley StatsRef: Statistics Reference Online*, 2014.

[159] A. Taha, "Intelligent ensemble learning approach for phishing website detection based on weighted soft voting," *Mathematics*, vol. 9, no. 21, p. 2799, 2021.

[160] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.