The Pennsylvania State University
The Graduate School

MODEL MISSPECIFICATION AND FEATURE SCREENING FOR ULTRAHIGH DIMENSIONAL DATA

A Disertation in
Statistics
by
Junyi Lin

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2011

The thesis of Junyi Lin was reviewed and approved* by the following:

Runze Li
Professor of Statistics and Chair of Graduate Program
Dissertation Advisor, Chair of Committee

Altman, Naomi
Professor of Statistics

Ghosh, Debashis
Professor of Statistics

Hunter, David
Associate Professor of Statistics

Xu, Ping
Professor of Mathematics

*Signatures are on file in the Graduate School.

Abstract

The variance-bias trade-off has been partially discussed for linear and logistic regression models, but not for generalized linear models as a whole. In this dissertation, we derive the bias of the treatment effect in covariate-unadjusted models, when some important covariates are omitted. This result encourages the use of the covariate-adjusted approach in general. On the other hand, we show that for a broad class of generalized linear models, estimation of the treatment effect obtained from covariate-adjusted models have larger variances compared to those obtained from covariate-unadjusted models. This result reveals the potential loss of efficiency related to the covariate-adjusted approach, particularly when sample size is not large. These theoretical results are illustrated through examples, a simulation study and a real data example.

This dissertation is also concerned with feature screening for ultrahigh dimensional data. We propose two unified sure independence ranking and screening procedures based on conditional characteristic functions. The proposed procedures do not require specification of a regression function. In addition, they can be directly applied for univariate or multivariate continuous, discrete and categorical responses. We show that, with the number of predictors growing at an exponential rate of the available sample size, these unified procedures possess both the *ranking consistency* and *sure screening* properties. The ranking consistency property ensures that all important features will be ranked above the unimportant ones asymptotically, and the sure screening property guarantees that all important features will be retained with an overwhelming probability after screening. Both are desired properties in ultrahigh dimensional data analysis. We study the finite-sample performance of our proposed independence ranking and screening procedures through simulations and illustrate the proposed procedures via an empirical analysis of a real-world data set.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I must convey my greatest gratitude to my thesis advisor, Dr. Runze Li, who has been my academic advisor for my PhD program since my first year at Pennsylvania State University. As my advisor, Dr. Li always provides unlimited support for my study, critical guidance for my research, useful suggestions for my future career and tender solicitude for my life. He used his academic expertise to allow me great flexibilities in my research topics. After my comprehensive exam, he also offered me the opportunity to study at the Methodology Center, where I gained more chances to do real data analysis and collaborate with other senior researchers. Besides all the help in academic studies, he shows great considerations for his students, and is always supportive and understanding for even small events in my life. I truly feel that he is not only my academic advisor, and also a senior friend during my PhD life at Pennsylvania State University.

I also feel very grateful to Dr. Naomi Altman. She is always supportive and considerate to me. When I was her teaching assistant for the graduate course Experimental Design, she provided great help and useful suggestions so that I could be able to be familiar with the role quickly. She also offered me the opportunity to be in charge of a review session, which was very crucial for me to ponder the importance of the communication with non-statisticians. I also want to thank her for her help in my master thesis: reading my writing carefully, marking my grammar mistakes patiently, and sharing useful comments.

It is also my great pleasure to convey my gratitude to Dr. Debashis Ghosh. I took one of the most important courses with him, Survival Analysis, which was extremely helpful in my research on joint modeling and my summer intern at Bank of America. I want to thank him for his suggestions and comments on my work. I am also indebted to Dr. David Hunter, who is always patient and supportive to not only me but also all the other students in the Stat Department. It is a shame

that I have not had a chance to take any of his courses, which are generally recognized to be creative and helpful. Last but not the least, I want to deeply thank Dr. Ping Xu, who serves as my outsider committee member, for his precious time and valuable suggestions.

I am indebted to my dear wife Danqi Zhu. During my PhD life at Pennsylvania State University, her endless love, support and considerations for me are dispensable parts of my research. I also would like to thank my parents Yongsong Lin and Qiuhua Chen. They have devoted all their time to me, and their love supports me all the way.

This is also the good time to say thank you to Dr. Lei Nie, my intern advisor, and Agus Sudjianto, my intern manager. I want to thank them for offering my summer intern opportunities.

Finally, I want to say thank you to the department and to all my friends here. I will remember every moment I spent in this great department and every happiness I shared with my friends here.

This thesis research was supported by grants from the National Institute on Drug Abuse (NIDA) grant P50-DA10075, NIDA grant R21-DA024260, the National Science Foundation (NSF) grant DMS 0348869 and National Natural Science Foundation of China grant 11028103. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIDA or the NSF.

Chapter 1

Introduction

1.1 Model Misspecification in Generalized Linear Models

Generalized linear models are widely used in the analysis of randomized clinical trials and observational studies. Clinical trials are conducted to collect data for new drugs or devices. In designing a clinical trial, a sponsor usually has the goal of obtaining a statistically significant result showing a significant difference in outcomes between the groups who receive the study treatments. In a randomized clinical trial, each subject is randomly assigned to a treated group or a control group before the experiment. When the number of subjects is sufficient large, this random allocation of treatments to subjects provides a balance between treatment groups for confounding variables (variables correlate with both the response variable and explanatory variables). Thus a randomized clinical trial can provide the most supportive evidence that the study treatment causes the expected effect. By contrast, in an observational study, the assignment of subjects into a treated group or a control group is beyond the control of the investigator. Thus, investigators

only observe potential correlations between treatments and outcomes. More details about clinical trials are beyond the scope of this proposal, and readers who are interested in clinical trials can refer to Pocock (2004).

When applied to randomized clinical trials or observational studies, GLMs are frequently misspecified due to the fact that some important covariates are omitted. In randomized clinical trials, important covariates may be unknown or unmeasured. In observational studies, omission of covariates arises when either there is the inability to collect all relevant factors or potential incomplete understanding of the study. It is also common that some covariates are excluded from the final model due to the concern that an overfitted model may yield inefficient estimates of parameters of interest. As a result it is important to study the impact of omitted covariates on the estimated effects of included covariates in these cases.

It is well known that under assumptions of classic linear regression, randomization leads to unbiased estimates of treatment effect even when important covariates are omitted (Cox, 1958). Furthermore, adjustment for desired covariates can improve the precision (small variance) of the treatment effect (Fisher, 1932). The improvement in the precision is due to the reduction of the residual variance.

However, it has been recognized that the property of unbiasedness cannot be extended to generalized linear models. For example, suppose (Cox, 1970) y is a binary outcome, T and $\mathbf{x}$ are binary covariates, and that the true model for the probability of response is given by

$$\text{logit}\{\Pr(y = 1|T, \mathbf{x})\} = \alpha + \beta T + \mathbf{x}\gamma. \tag{1.1}$$

Suppose the 2×2 classification (y, T) in sequence $(1, 1), (1, -1), (0, 1), (0, -1)$ has counts 900, 500, 100, 500 for $\mathbf{x} = 1$, and counts 500, 100, 500, 900 for $\mathbf{x} = -1$.

The data are summarized in Table 1.1 and Table 1.2. In Table 1.1, it is easy to

Table 1.1. Synthetic Data for logistic regression regarding bias

	x=1				**x**=-1		
	T=1	T=-1	Total		T=1	T=-1	Total
Y=1	900	500	1400	Y=1	500	100	600
Y=0	100	500	600	Y=0	500	900	1400
Total	1000	1000	2000	Total	1000	1000	2000

Table 1.2. Synthetic Data from Table 1.1 after combination over levels of **x**

	T=1	T=-1	Total
Y=1	1400	600	2000
Y=0	600	1400	2000
Total	2000	2000	4000

verify that T and $\mathbf{x}$ are independent, and the true model leads to an estimate of $\beta = \frac{1}{2}\log(9)$. In Table 1.2, the fitted logistic model is

$$\text{logit}\{\Pr(Y = 1|T)\} = \alpha^* + \beta^* T, \tag{1.2}$$

and omitting $\mathbf{x}$ yields an estimate of $\beta^* = \frac{1}{2}\log(49/9)$.

Gail et al. (1984) showed that the asymptotic bias from omitting covariates is zero if the link in the GLM is a linear function or a log function. Using Taylor series expansion, they presented an approximation of bias magnitude for regular cases. Neuhaus and Jewell (1993) proposed a geometric approach, which leads to analogous results as in Gail et al. (1984). A term called "Bias Factor" was developed in their paper by expanding around a different point from that in Gail et al. (1984). Drake and McQuarrie (1995) considered estimating the bias in observational studies of exposure effects, when some but not all covariates are omitted.

All these approaches adopt the functional relationship between the parameters in the true model and the misspecified model.

In some situations, the gains in precision regarding covariate adjustment do not apply to generalized linear models, either. For example, a logistic regression is fitted to a specific $2 \times 2 \times 2$ contingency table (Robinson and Jewell, 1991). Using

Table 1.3. Synthetic Data for logistic regression regarding variance

	x=0				**x**=1		
	T=1	T=0	Total		T=1	T=0	Total
Y=1	10	20	30	Y=1	40	80	120
Y=0	20	64	84	Y=0	5	16	21
Total	30	84	114	Total	45	96	141

Table 1.4. Synthetic Data from Table 1.3 after combination over levels of **x**

	T=1	T=0	Total
Y=1	50	100	150
Y=0	25	80	105
Total	75	180	255

both T and $\mathbf{x}$ as covariates leads to $\beta = \log(1.6)$ with associated standard error 0.354 while using T only yields $\beta^* = \log(1.6)$ as well with associated standard error 0.286. Note that in this example, T and $\mathbf{x}$ are correlated. Robinson and Jewell (1991) compared the variances of the two estimators, $\hat{\beta}$ and $\hat{\beta}^*$, in a logistic regression when both T and $\mathbf{x}$ are dichotomous variables. Neuhaus (1998) presented expressions for the effect of omitted covariates on the efficiency of the estimated effects of the included covariates in testing the hypothesis of no treatment effect ($\beta = \beta^* = 0$).

Meanwhile, some experience shows that for some clinical trials, statistical models which adjust for covariates are in close agreement with the simpler unadjusted

treatment comparisons. The properties of covariate-adjustment approach are not fully understood, and there remains confusion as to what is an appropriate statistical strategy. This thesis research is motivated by this confusion.

Previous theoretical results are obtained under the assumption that there are no interactions between treatment and covariates, while this assumption may not necessarily be realistic. In clinical trails, non-crossover (or quantitative) interactions between covariates and treatment are to be expected and crossover (or qualitative) interactions between covariates and treatment are possible (Peto, 1982; Gail and Simon, 1985). In epidemiologic studies, analysis of interaction between gene and environment or interaction between two major exposure variables can be the major interest (Stümer and Brenner, 2002; Zou, 2008; Richardson and Kaufman, 2009). Interactions between predictive variables are also important in psychology and the social science (McClelland and Judd, 1993).

Contribution of this Dissertation:

It is natural to ask whether those previous theoretical results still hold when there are treatment/covariate interactions. In a framework of generalized linear models which contain possible interactions between the main predictive variable (e.g. treatment) and covariates, we first investigate the bias caused by omitted covariates. Some bias approximations are appropriate when there are no interactions between the main predictive variable and covariates. However, they may not be meaningful in the presence of interactions. Consequently, we derive the relationship between the effects of the main predictive variable in the potentially misspecified model and the true model. Our results indicate that, in the presence of interactions, the difference between these two effects can be large even

when omitted covariates have negligible effects. This result is different from the conclusion that omitting a covariate (that does not interact with the main predictive variable) with a small effect leads to negligible bias as in Gail et al (1984). Furthermore, we present a new approach to assess the variance comparison. We prove that covariate-adjusted approaches lead to variation inflation for a broad class of generalized linear models. Conclusions on logistic regression in Robinson and Jewell (1991) can be obtained from this class of regression models as a special case.

1.2 Ultrahigh Dimensional Data

Scientific data of unprecedented size and complexity arise in a large variety of fields such as computational biology, climatology, neurology, health science, economics and finance. The total number of features can be much larger than the available sample size. Fan, Samworth and Wu (2009) pointed out that statisticians are confronting simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability in ultrahigh dimensional statistical learning problems. Traditional variable selection procedures such as the AIC (Akaike, 1973), BIC (Schwartz, 1978) and many other existing procedures for high-dimensional data, including the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001; Kim, Choi and Oh, 2008; Zou and Li, 2008), adaptive LASSO (Zou, 2006) and Dantzig selector (Candes and Tao, 2007; Candes, Wakin and Boyd, 2007) may not perform well due to the ultrahigh dimensionality. Fan and Lv (2008) emphasized the importance of feature screening in the ultrahigh dimensional data analysis, and proposed sure independence screening (SIS) to extract important features in the context of homoscedastic linear models. The SIS ranks the importance of candidate predictors

using the Pearson correlation coefficients for univariate and continuous response variables. Fan and Song (2009) proposed to rank the importance of each feature using the magnitude of the marginal likelihood under the framework of generalized linear models. Fan, Feng and Song (2011) generalized the SIS to nonparametric additive models. The aforementioned procedures require data analysts to specify a regression model between the response and the predictors. Their performances rely heavily on how well the postulated working model approximates the underlying true model. We refer to these procedures as model-based screening procedures. The theoretic properties of the model-based procedures were established based on the assumption that the imposed working model is the underlying true model. It is well known that the model assumptions are typically quite restrictive, and it is difficult to test the validity of these model assumptions in an ultrahigh dimensional setting.

Note that the general methodology of independence screening is to develop a marginal utility which is derived from the marginal regression of $\mathbf{y}$ onto X_j, for $j = 1 \ldots, p$. As discussed in Chapter 3, the marginal regression of $\mathbf{y} \mid X_j$ may misspecify the joint regression of $\mathbf{y} \mid \mathbf{x}$. However, the main purpose of independence screening is ranking the importance of predictors rather than estimating regression functions.

Contribution of this Dissertation:

In this dissertation, we first propose a unified sure independence ranking and screening (USIRS, for short) procedure to simultaneously tackle the issues of unprecedented size and complexity. The USIRS selects the important predictors in a model-free fashion in the sense that it does not require the researchers to specify a

functional regression relationship between the responses and the predictors. The proposed procedure enables us to identify important features on which the correlations among the response variables depend. The USIRS is readily applicable to the classical linear models, generalized linear models, index models, heteroscedastic models, transformation models, additive models and many others. It can be readily applied to regressions with either univariate or multivariate responses, regardless of whether the response variables are continuous, discrete or categorical. Moreover, the USIRS is computationally efficient and simple to implement, in that it does not require an iterative numerical optimization algorithm for calculating the marginal utility. This is desired in ultrahigh dimensional data analysis. We conduct Monte Carlo simulations to examine the finite sample performance of the USIRS, and compare the performance of the USIRS with the SIS (Fan and Lv, 2008), the nonparametric SIS (Fan, Feng and Song, 2011) and the SIS for generalized linear models (Fan and Song, 2009). The numerical comparisons show that the USIRS is a dramatic improvement upon existing procedures under some statistical settings.

We also observe that these existing procedures, as well as their iterative counterparts, fail to identify a class of important predictors which exhibit "symmetric patterns" to the response variable $\mathbf{y}$. In our context the symmetry pattern means that $E(X_j \mid \mathbf{y}) = 0$ for an important predictor X_j. To address this issue, in this dissertation, we develop a complementary methodology through a second-moment approach. This proposed second-moment unified sure independence ranking and screening (USIRS-II, for short) approach can be regarded as an important member in the marginal regression family. It retains the model-free feature, which is a very appealing property in ultrahigh dimensional regression analysis, particularly when there is little information about the model structure. It also inherits the

merits of the USIRS in that it allows different types of response variable (e.g., quantitative and categorical), no matter the response is univariate or multivariate. It thus can readily be used to capture important predictors which describe correlations between different response variables. In parallel to the USIRS, the new approach has both the sure screening and ranking consistency properties even when the dimension p grows at an exponential rate in n. Besides its nice properties in an asymptotic sense, the numerical studies demonstrate that it has a competent finite-sample performance in a wide range of regression models.

1.3 Dissertation Structure

This dissertation is organized as follows. In Chapter 2, we provide a brief literature review about model misspecification and statistical methodologies in high and ultrahigh dimensional data. In Chapter 3, we compare treatment effects in the true model and the misspecified model with treatment/covariates interactions. Chapter 4 and Chapter 5 focus on details about the unified sure independence ranking and screening procedure (USIRS) and the complimentary second-moment unified sure independence ranking and screening approach (USIRS-II) for ultrahigh dimensional data. Future studies are summarized in Chapter 6.

Chapter 2

Literature Review

2.1 Covariate-Adjusted Approaches in Generalized Linear Models

Suppose y is the response variable, $\mathbf{x}$ is a vector of covariates, and T is the main explanatory variable (e.g., a treatment indicator variable that takes values 1 or -1). Given $\mathbf{x}$ and T, the conditional expectation of y satisfies

$$\begin{aligned} E(y|T,\mathbf{x}) &= h(\eta) \\ \eta &= \alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma}, \end{aligned} \tag{2.1a}$$

where $h(\cdot)$ is a known function.

Suppose also that it is mistakenly assumed that

$$\begin{aligned} E(y|T,\mathbf{x}) = E(y|T) &= h(\eta^*) \\ \eta^* &= \alpha^* + T\beta^*. \end{aligned} \tag{2.1b}$$

2.1.1 Maximum Likelihood Estimation over Misspecified Models

Maximum likelihood estimation (MLE) has become one of the most important approaches for statistical inference, after Fisher (1922, 1925) advocated the method of maximum likelihood. A key assumption leading to the properties of the maximum likelihood estimator (Wald, 1949; LeCam, 1953) is that the true parameter lies within a specified parametric family of models; that is, the model must be correctly specified. However, in many cases, this assumption may not be satisfied.

Based on results of Wald (1949), Huber provided some general conditions, under which the MLE converges to a well-defined limit even when the model is not correctly specified. Akaike (1973) pointed out that when the true model is unknown, the MLE is a natural estimator for the parameters which minimizes the Kullback-Leibler Information Criterion (Kullback and Leibler, 1951).

For model (2.1a) and model (2.1b), let $\boldsymbol{\theta}' = (\alpha, \beta, \boldsymbol{\gamma}')$, $(\boldsymbol{\theta}^*)' = (\alpha^*, \beta^*)$. Assume $f(y|\boldsymbol{\theta}, T, \mathbf{x})$ is the true (conditional) density function of y, and $f(y|\boldsymbol{\theta}^*, T)$ is the misspecified (conditional) density function. The MLE, $\hat{\boldsymbol{\theta}}^*$, of $\boldsymbol{\theta}^*$ under the misspecified model (2.1b) converges to the value that minimizes the Kullback-Leibler divergence between model (2.1a) and model (2.1b):

$$E\Big[\log\big\{f(y|\boldsymbol{\theta}, T, \mathbf{x})/f(y|\boldsymbol{\theta}^*, T)\big\}\Big]. \tag{2.2}$$

Here the expectation is taken with respect to the true model.

White (1982) gave the asymptotic normal distribution of $\hat{\boldsymbol{\theta}}^*$ under certain regularity conditions. More precisely, he showed that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*) \xrightarrow{D} N\big(0, C(\boldsymbol{\theta}^*)\big), \tag{2.3}$$

where $C(\boldsymbol{\theta}^*) = A(\boldsymbol{\theta}^*)^{-1}B(\boldsymbol{\theta}^*)A(\boldsymbol{\theta}^*)^{-1}$, and

$$A(\boldsymbol{\theta}^*) = E\Big\{\partial^2 \log f(y|\boldsymbol{\theta}^*, T)/\partial\boldsymbol{\theta}^*\partial(\boldsymbol{\theta}^*)'\Big\},$$
$$B(\boldsymbol{\theta}^*) = E\Big\{\partial \log f(y|\boldsymbol{\theta}^*, T)/\partial\boldsymbol{\theta}^* \cdot \partial \log f(y|\boldsymbol{\theta}^*, T)/\partial(\boldsymbol{\theta}^*)'\Big\}.$$

Again, the expectation is taken with respect to the true model.

For the true model (2.1a), $A(\boldsymbol{\theta}) = B(\boldsymbol{\theta})$, and (2.3) reduces to the classic asymptotic normal distribution of MLEs. In general, $A(\boldsymbol{\theta}^*) \neq B(\boldsymbol{\theta}^*)$, and thus special care must be taken for statistical inference in the presence of misspecification. Gail (1988) and Neuhaus (1998) observed that $A(\boldsymbol{\theta}^*) = B(\boldsymbol{\theta}^*)$ for all binary regression models. In other words, for binary regression models, the information matrix of the misspecified model provides the correct value of the covariance matrix when using the results of White (1982).

2.1.2 Covariate-Adjusted Approaches or Not?

The analysis of the primary objectives of randomized clinical trials often are not adjusted for covariates, except possibly for stratification variables. People often debate whether a covariate-adjusted approach should be adopted or not as the primary analysis. Grouin et al. (2004,2005) provided guidance: when subgroup analysis can be done; when they should be done; and their interpretations.

Hauck et al. (1998) reviewed the literature regarding logistic and Cox (proportional hazards) regression models and advocated the adjustment for important prognostic covariates in order to come as close as possible to the clinically most relevant subject-specific measure of treatment effect since omitting covariates from the analysis of randomized trials leads to a loss of efficiency as well as a change

in the treatment effect being estimated. They pointed out that additional benefits from adjustment for important covariates would be an increase in efficiency of tests for no treatment effect and improved external validity, which is particularly relevant to meta-analyses.

Ford and Norrie (2002) reviewed previously published literature and showed that the impact of including covariates in models used to estimate the magnitude of treatment effects in long-term clinical trials is different from what would be predicted from results for the classic linear model. They used a data from clinical trials to evaluate the role of covariates in estimating treatment effects and risk in long-term clinical trials.

Based on a recent survey of 50 trial reports in four major journals, Pocock et al (2002) examined how literature in the medical journals used baseline data on each patient at randomization to do (i) subgroup analysis; (ii) covariate-adjusted analysis; (iii) baseline comparions. Some major problems and key issues were highlighted, including inconsistencies in the use of covariate-adjustment and the lack of clear guidelines on covariate selection. They recommended to adjust for the appropriate covariates (that is, the strong predictors of outcome) and to make one's statistical policy for covariate adjustment completely objective. Pocock et al (2002) also advocated manipulation of variable selection procedure, and encouraged more methodological research on this topic. They believed that variable selection procedures have a useful role in formulating covariate-adjustment in larger trials.

As pointed out in Neuhaus et al. (1991), there exists a close connection between bias analysis and generalized linear mixed models. If we view the omitted covariate as a random intercept, then the bias due to omission of covariates can be used to explain the differences between cluster-specific models (or conditional models) and population-averaged models (or marginal models) (Zeger, et al. 1988). Cluster-

specific models are similar to models without omissions ($\eta = \alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma}$) while population-averaged models are analogous to ones with omitted covariates ($\eta^* = \alpha^* + T\beta^*$).

Senn (2004) discussed the existing controversy about the use of marginal and conditional models, particularly in the analysis of data from longitudinal studies. The seeming differences between marginal and conditional models were shown to be caused by preimposed unidentifiable constraints on the random effects. Senn (2004) also discussed the advantages of conditional models over marginal models, and regarded the conditional model as fundamental, from which marginal predictions can be made.

This debate over covariate-adjustment approach extends to other areas as well (e.g., epidemiology). Weng et al (2009) generated twenty-five scenarios with case-control samples from 10 simulated populations to compare the performance of 4 covariate selection approaches in the presence of confounders of various strengths. They concluded that the full model is more appropriate unless the standard error in the reduced model is substantially smaller.

2.1.3 Bias Approximations

Theoretical results show that omitting predictive covariates in GLMs often causes inaccurate treatment effect estimates. Assume that T , independent of $\mathbf{x}$, is a treatment indicator variable that takes values 1 or -1 with probability p and $1-p$ respectively. Gail et al. (1984) confirmed that randomization does not always lead to asymptotically unbiased estimates of T when $\mathbf{x}$ is omitted. Let

$$\zeta_1 = E(y|T=1) = E_{\mathbf{x}}\{h(\alpha + \beta + \mathbf{x}'\boldsymbol{\gamma})\}, \tag{2.4a}$$

$$\zeta_2 = E(y|T=-1) = E_{\mathbf{x}}\{h(\alpha - \beta + \mathbf{x}'\boldsymbol{\gamma})\}. \tag{2.4b}$$

If h^{-1} exists and is well defined at ζ_1 and ζ_2, then

$$\beta^* = \frac{1}{2}\{h^{-1}(\zeta_1) - h^{-1}(\zeta_2)\}. \tag{2.5}$$

Using Taylor's expansion, they showed that

$$\beta^* - \beta \doteq \frac{1}{4}\boldsymbol{\gamma}' Var(\mathbf{x})\boldsymbol{\gamma}\big\{h''(\alpha+\beta)/h'(\alpha+\beta) - h''(\alpha-\beta)/h'(\alpha-\beta)\big\}. \tag{2.6}$$

Note that the bias will be small if $\boldsymbol{\gamma}$ is small or the covariate $\mathbf{x}$ has little variability. This approximation provides reasonable magnitude of the bias. Particularly, they pointed out that $\beta^* = \beta$ when $h(\eta) = a\eta + b$ or $h(\eta) = c\exp(a\eta) + b$ for some real constants a, b and c. This can be verified through (2.5).

Neuhaus and Jewell (1993) presented a geometric approach to assess the direction of the bias resulting from omitted covariates in GLMs. Under their assumptions, T is not required to be an indicator variable. Let $\delta = \mathbf{x}'\boldsymbol{\gamma}$ and $\mu_k = E(y|T+k,\delta) = h\{\alpha + (T+k)\beta + \delta\}$, then $\beta = h^{-1}(\mu_1) - h^{-1}(\mu_0)$. When T and $\mathbf{x}$ are independent and the covariate vector $\mathbf{x}$ is omitted, one obtains the marginal effect of a unit increase in the covariate T by considering

$$\beta^* = h^{-1}(\mu_1^*) - h^{-1}(\mu_0^*) = H(\beta), \tag{2.7}$$

where $\mu^* = E_\delta(\mu_k) = \int h\{\alpha + (T+k)\beta + \delta\} f(\delta) d\delta$.

Note that $H(0) = 0$. Thus, when $\beta = 0$ in the true model, unbiased estimate of the treatment effect can be obtained through a model with omitted covariates.

Furthermore, $H(\beta)$ can be approximated by expanding about $\beta = 0$:

$$\beta^* \doteq \beta H'(0) = \beta g'[E\{g^{-1}(\delta)\}]E[1/g'\{g^{-1}(\delta)\}], \tag{2.8}$$

where $g(\cdot) = h^{-1}(\cdot)$. The direction of the bias can be determined by checking whether $1/g'(\cdot)$ is concave or convex. Table 2.1 gives the bias factor $H'(0)$ for several popular link functions.

Table 2.1. Bias factors for generalized linear model parameter estimates obtained with omitted covariates

Link function	$g(\mu)$	Bias factor $H'(0)$ in (2.8)
Linear	$a + b\mu$	1
Log	$(1/b)\log(a + b\mu)$	1
Logistic	$\log\{\mu/(1-\mu)\}$	$1 - \frac{Var(\mu_0)}{E(\mu_0)\{1-E(\mu_0)\}}$
Complementary log-log	$\log\{-\log(1-\mu)\}$	$\frac{E\{(1-\mu_0)\log(1-\mu_0)\}}{\{1-E(\mu_0)\}\log\{1-E(\mu_0)\}}$

Note that (2.5) and (2.7) are essentially the same. (2.6) is obtained by a Taylor expansion about $\gamma = 0$ while (2.8) is obtained by expanding about $\beta = 0$.

Drake and McQuarrie (1995) considered estimating the bias in observational studies of exposure effects for generalized linear models with canonical link. In their case, the bias is due to omission of some but not all confounders. They pointed out that the bias approximation can be derived from a system of linear equations after using first order Taylor expansion.

2.1.4 Asymptotic Relative Efficiency for Model Misspecification

Let's first consider a simple classic linear regression. Suppose the structure of a population is described by the following two linear models:

$$E(y|T,X) = \eta,\ \eta = \alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma}, \tag{2.9a}$$

$$E(y|T) = \eta^*,\ \eta^* = \alpha^* + T\beta^*. \tag{2.9b}$$

Here T is not required to be an indicator and X is a scalar. Denote $\hat{\beta}$ and $\hat{\beta}^*$ to be MLEs of β and β^*, respectively. The asymptotic relative efficiency of $\hat{\beta}$ to $\hat{\beta}^*$ is defined to be

$$\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^*) = \frac{\text{Var}(\hat{\beta}^*)}{\text{Var}(\hat{\beta})}. \tag{2.10}$$

For $\hat{\beta}$ and $\hat{\beta}^*$ associated with (2.9a) and (2.9b), the formula for ARE($\hat{\beta}$ to $\hat{\beta}^*$) is given by (Robinson and Jewell, 1991)

$$\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^*) = \frac{1-\rho^2_{T\mathbf{x}}}{1-\rho^2_{Y\mathbf{x}\cdot T}}, \tag{2.11}$$

where $\rho_{T\mathbf{x}}$ is the simple correlation between T and $\mathbf{x}$, and $\rho_{y\mathbf{x}\cdot T}$ is the partial correlation between y and $\mathbf{x}$ conditional on T.

Under assumptions of classic linear regression, if $\rho_{T\mathbf{x}} = 0$, then $\beta = \beta^*$. In this case, we can see that ARE($\hat{\beta}$ to $\hat{\beta}^*$)≥ 1. This explains why adjusting for desired covariates improves the precision of the estimates of treatment effects. Note also that $\rho_{Y\mathbf{x}\cdot T} = 0$ is equivalent to $\gamma = 0$ (hence $\beta = \beta^*$). In this case, ARE($\hat{\beta}$ to $\hat{\beta}^*$)≤ 1. This explains why it is not wise to adjust for non-predictive covariates.

For generalized linear models (2.1a) and (2.1b), the above conclusions do not always hold. Neuhaus (1998) derived the estimation efficiency with omitted covariates in generalized linear models. Using the results of White (1982) on estimation in misspecified models, Neuhaus (1998) showed that when T and $\mathbf{x}$ are independent, the Pitman efficiency of $\hat{\beta}$ to $\hat{\beta}^*$ at $\beta = 0$ is

$$\begin{aligned} &\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^* \text{ at } \beta = 0) \\ &= \frac{E\{V + \phi^{-1}(\mu_0 - E\mu_0)^2\}E[1/\{V\{g'(\mu_0)\}^2\}]}{[E\{1/g'(\mu_0)\}]^2} \geq 1. \end{aligned} \tag{2.12}$$

Here $\mu_0 = h(\alpha + \mathbf{x}'\boldsymbol{\gamma})$, $g(\cdot) = h^{-1}(\cdot)$, V is the variance function and ϕ is the dispersion parameter. Particularly, when $\mathbf{x}$ is a nonconfounding covariate ($\beta = \beta^*$), then

$$\begin{aligned} &\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^* \text{ at } \beta = 0) \\ &= \frac{\text{Var}(\hat{\beta}^*|\beta = 0)}{\text{Var}(\hat{\beta}|\beta = 0)} = E\{V + \phi^{-1}(\mu_0 - E\mu_0)^2\} \cdot \{g'(E\mu_0)\}^2 \cdot E[1/\{V\{g'(\mu_0)\}^2\}]. \end{aligned} \tag{2.13}$$

Robinson and Jewell (1991) presented ARE($\hat{\beta}$ to $\hat{\beta}^*$ at $\beta = 0$) for the logistic regression model with T and $\mathbf{x}$ both being binary. They showed that

$$\text{ARE}(\hat{\beta} \text{ to } \hat{\beta}^* \text{ at } \beta = 0) = \frac{E(p_{0j})E(q_{0j})}{E(p_{0j}q_{0j})} \geq 1, \tag{2.14}$$

with equality occurring if and only if $\mathbf{x}$ is independent of (y, T). Here, $p_{ij} = \Pr(y = 1|T = i, \mathbf{x} = j)$ for $i, j = 0, 1$, $q_{ij} = 1 - p_{ij}$.

2.1.5 The Presence of Interactions between Treatment and Covariates

Previous theoretical results are obtained under the assumption that there are no interactions between treatment and covariates, while this assumption may not always hold. The importance of treatment/covariate interactions has been recognized recently, and the evaluation of evidence of that treatment effects vary among different subsets of patients has gained increasing attention in the analysis of large clinical trials. Qualitative (or crossover) interactions occur when one treatment is superior for some subsets of patients while the alternative treatment is superior for other subsets. Quantitative (or non-crossover) interactions arise when there is variation in the magnitude, but not in the direction, of treatment effects among subsets.

Gail and Simon (1985) developed a likelihood ratio test for qualitative interactions. Let ω_i denote the true difference in treatment effects within disjoint patient subset i for $i = 1, \cdots, I$. Assume estimates D_i of ω_i have independent normal distribution with mean ω_i and known variance σ_i^2. In case of large samples, consistent estimates of σ_i^2 can be used instead. The hypothesis of no crossover interactions is equivalent to the vector $\Omega = (\omega_1, \cdots, \omega_I)$ satisfying either $\omega_i \geq 0$ for all i or $\omega_i \leq 0$ for all i. Let $\mathbf{O}^+ = \{\Omega : \ \omega_i \geq 0 \text{ for all } i\}$, and $\mathbf{O}^- = \{\Omega : \ \omega_i \leq 0 \text{ for all } i\}$. The likelihood ratio test of this hypothesis is based on the test statistic

$$\frac{\max_{\Omega \in \mathbf{O}^+ \cup \mathbf{O}^-} \exp[\sum_{i=1}^{I} \{-(D_i - \omega_i)^2/(2\sigma_i^2)\}]}{\max_{\Omega} \exp[\sum_{i=1}^{I} \{-(D_i - \omega_i)^2/(2\sigma_i^2)\}]}. \tag{2.15}$$

Since the maximum value of the denominator is 1, the likelihood ratio test is thus

$$\max_{\Omega \in \mathbf{O}^+ \cup \mathbf{O}^-} \exp[\sum_{i=1}^{I} \{-(D_i - \omega_i)^2/(2\sigma_i^2)\}] < k, \tag{2.16}$$

or equivalently

$$\min_{\Omega \in \mathbf{O}^+} \sum_{i=1}^{I} \{(D_i - \omega_i)^2/\sigma_i^2\} > c, \tag{2.17}$$

and

$$\min_{\Omega \in \mathbf{O}^-} \sum_{i=1}^{I} \{(D_i - \omega_i)^2/\sigma_i^2\} > c, \tag{2.18}$$

with $c = -2 \log k$. Let

$$Q^- \equiv \sum (D_i^2/\sigma_i^2) I(D_i > 0), \tag{2.19a}$$

$$Q^+ \equiv \sum (D_i^2/\sigma_i^2) I(D_i < 0), \tag{2.19b}$$

where $I(\cdot)$ is an indicator function. The rejection region is given by $\min(Q^+, Q^-) > c$. Gail and Simon (1985) also provided a table of values of c for different significance levels.

Stürner and Brenner (2002) extended the concept of flexible matching strategies to the field of gene-environment interactions. They assessed the power and efficiency of such studies to detect and estimate gene-environment interactions under a variety of assumptions regarding the prevalence and effects of the environmental exposure and the genetic susceptibility as well as their association in the population. Zou (2008) presented a new way that uses the conventional asymmetric intervals for risk ratios to set confidence limits for measures of additive interaction

in a four-by-two table. A four-by-two table, with its four rows representing the presence and absence of gene and environmental factors, has been suggested as the fundamental unit in the assessment of gene-environment interaction.

2.2 Statistical Modeling for High Dimensional Data

High dimensional data appear in various scientific fields. How to model the relationship between response variables and associated features in high dimensional data is key to many scientific discoveries. Variable selection and feature screening are frequently used to get parsimonious and unbiased models; that is, feature screening and variable selection can improve estimation accuracy and enhance model interpretability by identifying important predictors from a large amount of candidates. In this section, we will briefly review those techniques. Let $\mathbf{x} = (X_1, X_2, \ldots, X_p)'$ and y be the predictor vector and the response vector, respectively. We also assume that $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ is a random sample from $(\mathbf{x}, y)$.

2.2.1 Variable Selection via Penalized Likelihood

Various penalty functions have been proposed for model selection in the past decade, including the LASSO or L_1 penalty (Tibshirani, 1996), the ridge regression or L_2 penalty (Hoerl and Kennard, 1970; Frank and Friedman, 1993) and the SCAD penalty (Fan and Li, 2001). Assume that y_i has a density function $f_i(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i)$ given $\mathbf{x}_i$, where g is a known link function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector. Let $\ell_i(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i) = \log f_i(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i)$ denote the con-

ditional log-likelihood function of y_i. A general form of the penalized log-likelihood function is

$$\sum_{i=1}^{n} \ell_i\Big(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i\Big) - n\sum_{j=1}^{p} P_\lambda(|\beta_j|), \tag{2.20}$$

where P_λ is the penalty function and λ is the tuning parameter.

When $P_\lambda(|\beta_j|) = \lambda|\beta_j|$, (2.20) reduces to the Least Absolute Shrinkage and Selection Operator (LASSO), which corresponds to L_1 penalty and minimizes

$$-\sum_{i=1}^{n} \ell_i\Big(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i\Big) + n\lambda\sum_{j=1}^{p} |\beta_j|. \tag{2.21}$$

Note that this is equivalent to minimizing $-\sum_{i=1}^{n} \ell_i\Big(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i\Big)$ with constrain that $\sum_{j=1}^{p} |\beta_j| \leq t$ for some real number t. The LASSO gives sparse solutions in the sense that it shrinks coefficients and produces some coefficients that are exactly 0. The LASSO is very popular because of its practical implementation, although it is biased due to the shrinkage.

When $P_\lambda(|\beta_j|) = \lambda|\beta_j|^2$, (2.20) is equivalent to the ridge regression. L_2 penalty is applied in this case and it minimizes

$$-\sum_{i=1}^{n} \ell_i\Big(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i\Big) + n\lambda\sum_{j=1}^{p} |\beta_j|^2. \tag{2.22}$$

It is worth mentioning that some classic variable selection (e.g., the best subset variable selection) approaches for linear regression models are special cases of regularized approaches when $P_\lambda(|\beta|)$ is L_0 penalty. This can be written as minimizing

the following objective function

$$\sum_{i=1}^{n} ||y_i - \mathbf{x}_i'\boldsymbol{\beta}||^2 + \frac{n}{2}\lambda^2 \sum_{j=1}^{p} I\{|\beta_j| \neq 0\}. \tag{2.23}$$

For instance, if $\lambda = \sigma\sqrt{2/n}$, then (2.23) reduces to the AIC criteria (Akaike, 1974). When $\lambda = \sigma\sqrt{\log(n)/n}$, then (2.23) is equivalent to the BIC criteria (Schwarz, 1978). Note that L_0 penalty function is discontinuous and thus it requires exhaustive search over 2^p subsets. The computational challenges limit the implementations of the best subset variable selection in high dimensional problems.

The L_r penalty has the form $P_{\lambda,r}(|\beta_j|) = \lambda|\beta_j|^r$, which is known as "bridge regression" (Frank and Friedman, 1993; Fu, 1998). It represents a family of penalty functions by bridging best subset regression and ridge regression. The solution is only continuous when $r \geq 1$ (Fan and Li, 2001). If $r > 1$, bridge regression can not produce sparse estimates (e.g., ridge regression).

Fan and Li (2001) studied the selection of penalty functions thoroughly, and advocated the smoothly clipped absolute deviation penalty (SCAD), which is symmetric and non-convex on $(0, \infty)$. The penalty function takes the following form

$$P_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & \text{if } 0 \leq |\beta_j| \leq \lambda; \\ \frac{1}{2(a-1)}\{(a^2-1)\lambda^2 - (|\beta_j| - a\lambda)^2\} & \text{if } \lambda \leq |\beta_j| \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta_j| \geq a\lambda, \end{cases}$$

where a is suggested to take value 3.7.

SCAD enjoys good properties like asymptotical *unbiasedness, sparsity and continuity*. Unbiasedness means the resulting estimator is unbiased when the true unknown parameter is large enough. Sparsity indicates setting small estimated coefficients as 0 to reduce model complexity. Continuity guarantees the resulting

estimator is continuous to avoid model instability. Fan and Li (2001) also derived the oracle property of the SCAD in the sense that the proposed estimators perform as well as if the true model were known in advance.

Several efficient algorithms have been proposed to find the penalized maximum likelihood estimate. Fan and Li (2001) presented the local quadratic approximation (LQA) for the penalty function. More precisely, given a initial value $\beta^{(0)}$, if $\beta_j^{(0)}$ is very close to 0, then set $\beta_j^{(0)} = 0$; otherwise, the penalty function is locally approximated by a quadratic function

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_j^{(0)}|) + \frac{\{\beta_j^2 - (\beta_j^{(0)})^2\}P'_\lambda(|\beta_j^{(0)}|+)}{2|\beta_j^{(0)}|} \tag{2.24}$$

for $\beta_j \approx \beta_j^{(0)} \neq 0$. A modified Newton-Raphson algorithm then can be used to find the penalized maximum likelihood estimates. As pointed out in Fan and Li (2001), one potential weakness of the LQA is that once a coefficient estimate is set to be zero, it will be excluded from the final model.

Hunter and Li (2005) showed that the LQA is an example of the Minorize-Maximize (MM) algorithm, which is an extension of the well known EM algorithm. Define

$$\Phi_{\beta^0} = P_\lambda(|\beta^0|) + \frac{\{\beta_j^2 - (\beta_j^0)^2\}P'_\lambda(|\beta_j^0|+)}{2|\beta_j^0|},\ \beta^0 \neq 0. \tag{2.25}$$

Note that

$$\Phi_{\beta^0}(\beta) - \Phi_{\beta^0}(\beta^0) \geq P_\lambda(|\beta|) - P_\lambda(|\beta^0|), \tag{2.26}$$

which gives the *descent property*

$$\Phi_{\beta^0}(\beta) < \Phi_{\beta^0}(\beta^0) \text{ implies } P_\lambda(|\beta|) < P_\lambda(|\beta_0|). \tag{2.27}$$

This property guarantees that a decrease in the value of $\Phi_{\beta^0}(\beta)$ leads to a decrease in the value of $P_\lambda(|\beta|)$. Hunter and Li (2005) studied the convergence properties of the LQA by using techniques applicable to the MM algorithm in general. In order to eliminate the potential drawback of the LQA, they improved the LQA by using small perturbed versions of P_λ and Φ_{β^0}, i.e.

$$P_{\lambda,\epsilon}(|\beta|) = P_\lambda(|\beta|) - \epsilon \int_0^{|\beta|} \frac{P'_\lambda(t)}{\epsilon + t} dt, \tag{2.28}$$

$$\Phi_{\beta^0} = P_\lambda(|\beta^0|) + \frac{\{\beta_j^2 - (\beta_j^0)^2\} P'_\lambda(|\beta_j^0|+)}{2(|\beta_j^0| + \epsilon)}. \tag{2.29}$$

They showed that the maximizer of the perturbed version is close to the maximizer of the original penalized likelihood function as long as ϵ is small and the original penalized likelihood function is not too fat near the maximizer.

Efron et al. (2004) proposed an efficient path algorithm, the least angel regression algorithm (LARS), to compute the entire solution path of the LASSO as a function of λ. The computational cost of LARS is $O(np^2)$, which is the same to that of the ordinary least squares. It turns out that with a squared-error loss and L_1 penalty, the entire solution path is piecewise linear. Rosset and Zhu (2007) generalized this idea, and derived a algorithm which gives piecewise linear coefficient paths for a class of loss functions with L_1 penalty.

Both the LARS and the piecewise linear algorithm can be viewed as direct path seeking methods. Instead of repeatedly solving numerical optimization problems for different tuning parameters, direct path seeking methods construct the entire

path directly in the parameter space. This dramatically reduces the computation burden. Denote $\hat{\beta}(v)$ the solution points on the entire path with path length v. Starting from $\hat{\beta}(0)$, the successive step is constructed by

$$\hat{\beta}(v + \Delta v) = \hat{\beta}(v) + d(v) \cdot \Delta v, \tag{2.30}$$

where $d(v)$ is the direction from the current path point to the successive path point and Δv is a specified distance along that direction. The algorithm proceeds until the path point arrives at the un-penalized likelihood estimate. Different combinations of loss functions and penalty functions lead to different $d(v)$ and Δv. LARS is designed for squared-error loss and L_1 penalty. The piecewise linear algorithm creates the entire path for the piecewise quadratic loss function and L_1 penalty. Friedman (2008) generalized the direct path seeking idea and proposed a unified algorithm called the generalized path seeking (GPS) approach. Recall that a regularization problem can also be expressed as

$$\hat{\boldsymbol{\beta}}(t) = \arg\min_{\boldsymbol{\beta}} \hat{R}(\boldsymbol{\beta}) \text{ s.t. } P(\boldsymbol{\beta}) \leq t, \tag{2.31}$$

which is equivalent to

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \Big\{ \hat{R}(\boldsymbol{\beta}) + \lambda \cdot P(\boldsymbol{\beta}) \Big\}. \tag{2.32}$$

Here

$$\hat{R}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} L\Big(y_i, \mathbf{x}_i{}'\boldsymbol{\beta}\Big),$$

and $L(\cdot, \cdot)$ is a loss function. Note that in (2.32) λ is excluded from the penalty

function. The GPS can be used to approximate the entire path of any (differentiable) convex loss function with any penalty that satisfies

$$\frac{\partial P(\boldsymbol{\beta})}{\partial |\beta_j|} > 0, \ j = 1, \cdots, p, \tag{2.33}$$

for all values of $\boldsymbol{\beta}$. The motivation of the GPS is from the linear approximation to optimization problem (2.31). When all exact coefficient paths $\hat{\beta}_j$ are continuous, the GPS paths coincide with the exact ones provided that all $\hat{\beta}_j(t)$ are monotonic functions of t. When the exact paths are not continuous, the GPS provides continuous approximations to the exact paths. As pointed out in Friedman (2008), the main principal advantages of GPS are simplicity, generality and speed.

The development of efficient algorithms has attracted much attention of researchers, and various LASSO type penalty functions have been proposed recently. Zou (2006) studied the consistency of the LASSO, and showed that the LASSO does not enjoy the oracle property. An intuitive explanation for this result can be found in Fan and Li (2001). Instead of the equal penalty to all coefficients in the ordinary LASSO, Zou (2006) assigned different penalties to different coefficients, and called it the adaptive LASSO. Suppose that $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}$, then the adaptive LASSO estimate $\hat{\boldsymbol{\beta}}^*$ is the solution to the following

$$-\sum_{i=1}^{n} \ell_i\Big(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i\Big) + n\lambda \sum_{j=1}^{p} \hat{\omega}_j \cdot |\beta_j|, \tag{2.34}$$

where $r > 0$ is a tuning parameter for the weight $\hat{\omega}_j = 1/|\hat{\beta}_j|^r$. One can use the ordinary least squares estimator as a candidate in the initial value. It has been showed that under mild regularity conditions and with a proper choice of λ, the adaptive LASSO enjoys the oracle property. Note that the adaptive LASSO is

essentially an L_1 penalty function, thus current efficient algorithms for solving the LASSO can be used to compute the adaptive LASSO estimates.

Zou and Li (2008) proposed a new unified algorithm called the local linear approximation (LLA) approach. It maximizes the penalized likelihood for a broad class of concave penalty functions via a local linear approximation. Given the initial value $\beta^{(0)}$, unlike (2.24), the penalty function is locally approximated by a linear function

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_j^{(0)}|) + P'_\lambda(|\beta_j^{(0)}|)(|\beta_j - \beta_j^{(0)}|) \tag{2.35}$$

for $\beta_j \approx \beta_j^{(0)}$. The LLA does not require $\beta_j^{(0)} \neq 0$, thus eliminates the potential weakness of the LQA. Using the un-penalized likelihood estimate as the initial value, the algorithm proceeds through repeatedly solving

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} \Big\{ -\sum_{i=1}^{n} \ell_i\Big(g(\mathbf{x}_i'\boldsymbol{\beta}), y_i\Big) + n\sum_{j=1}^{p} P'_\lambda(|\beta_j^{(k)}|)(\beta_j| - \beta_j^{(k)})\Big\}, \tag{2.36}$$

for $k = 0, 1, \cdots$.

The LLA enjoys some distinguished properties. First, the LLA is not only a majorization of the penalty function, but also the best convex majorization of P_λ in the sense that for any other convex majorization of P_λ (at $\boldsymbol{\beta}^{(k)}$), it is no smaller than the LLA for all $\boldsymbol{\beta}$. Second, given a good initial value, the one-step LLA estimator inherits the good feature of the LASSO of providing a sparse solution. Also, the one-step LLA has the oracle property, provided that the initial value is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}$ (e.g., un-penalized likelihood estimate).

Zhang (2007) introduced and studied a penalized variable selection for high-dimensional linear regression models, called MC+ methodology which consists of

two components: a *minimax concave penalty* (MCP) and a *penalized linear unbiased selection* (PLUS) algorithm. The penalty function MCP is defined as

$$P_\lambda(|\boldsymbol{\beta}|) = \lambda \int_0^{|\boldsymbol{\beta}|} \big(1 - \frac{t}{a\lambda}\big)^+ dt, \tag{2.37}$$

where a is a tuning parameter. Define a class of penalty functions that satisfy

$$P'_\lambda(t) = 0 \ \forall t \geq a\lambda, \ P'_\lambda(0+) = \lambda. \tag{2.38}$$

It has been shown that in this class, MCP is the minimizer of the maximum concavity

$$k(P;\lambda) \equiv \sup_{0<t_1<t_2} \frac{P'_\lambda(t_1) - P'_\lambda(t_2)}{t_1 - t_2}. \tag{2.39}$$

The PLUS computes a piecewise linear path of critical points for the possibly non-convex minimization problem. In each step, the PLUS computes one line segment in its path between two turning points, and its computational cost is the same as the LARS per step. As $a \to \infty$, the MC+ path converges to the LASSO path.

Under the unified framework of regularized least squares with concave penalties, Lv and Fan (2009) considered the properties of regularization methods for model selection and sparse recovery. For model selection, they established conditions under which a regularized least squares estimate has the nonasymptotic property. The nonasymptotic property is weaker than the oracle property in the sense that it does not require the estimate to be asymptotically normal but only consistent under L_∞ loss. Furthermore, the number of parameters can grow exponentially

with the sample size. Particularly, they considered a family of penalties defined as

$$P_a(|\beta|) = \frac{(a+1)|\beta|}{a+|\beta|} = \Big(\frac{\beta}{a+|\beta|}\Big) I(|\beta \neq 0|) + \Big(\frac{a}{a+|\beta|}\Big)|\beta|, \ a \in (0,\infty), \quad (2.40)$$

and

$$\lim_{a\to 0+} P_a(|\beta|) = I(|\beta| \neq 0), \ \lim_{a\to\infty} P_a(|\beta|) = |\beta|.$$

Note that this family of penalties gives a smooth homotopy between L_0 and L_1 penalties. As pointed out in Lv and Fan (2009), it would be interesting to extend these results to regularization methods for generalized linear models and more general models and loss functions.

2.2.2 Feature Screening for Ultrahigh Dimensional Data

The availability of massive data along with new scientific problems have challenged traditional statistical theories, and more innovative statistical procedures are required, especially when $p > n$ or even $p >> n$. Under this circumstance, one may want to reduce the original dimensionality p from a huge scale (e.g., $\exp\{O(n^{\xi})\}$ for some $\xi > 0$) to a relatively smaller scale (e.g., o(n)) via some efficient approaches, and then apply variable selection techniques to the reduced data to obtain a parsimonious model.

To achieve this goal, Fan and Lv (2008) proposed a sure independent screening (SIS) procedure, which is based on correlation learning and screens out features that have weak correlations with the response. The idea behind SIS is simple.

Consider the marginal linear regression for each predictor $j = 1, \cdots, p$

$$y = \alpha_{0j} + \alpha_j X_j + \epsilon_j.$$

If X_j and y are standardized, then the least squares estimate of α_j is exactly the Pearson correlation coefficient. This motivates us to consider the correlation coefficient between X_j and y as a marginal utility to screen predictors. Assume the true model is

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon,$$

and denote $\mathcal{M}_* = \{1 \leq j \leq p : \beta_j \neq 0\}$. Assume also that $\mathbf{x}$ is standardized, and define $\boldsymbol{\omega} = \mathbf{x}y$. For any given $\gamma \in (0, 1)$, we can sort the p components of the vector $\boldsymbol{\omega}$ in decreasing order and define a submodel

$$\mathcal{M}_\gamma = \{1 \leq j \leq p : \omega_j^2 \text{ is among the first } \gamma_n \text{ largest of all}\},$$

where γ_n is a prespecified threshold value. In other words, we shrink the full model $\{1, \cdots, p\}$ down to a submodel $\mathcal{M}_\gamma$ with size smaller than n. In particular, when y is binary, coded by 1 and -1, then $\boldsymbol{\omega}$ reduces to a two-sample t statistic, which has been used in gene selection (Storey and Tibshirani, 2003). Under some regularity conditions, Fan and Lv (2008) showed that the probability $P(\mathcal{M}_* \subset \mathcal{M}_\gamma)$ can tend to 1 even when p grows in an exponential rate with n.

Using a similar idea, Fan, Samworth and Wu (2009) proposed the generalized sure independent screening (GSIS) procedure for generalized linear models . They

suggested using the marginal utility of the jth feature as

$$\ell_j = \min_{\beta_0,\beta_j} n^{-1} \sum_{i=1}^{n} \ell(y_i, \beta_0 + X_{ij}\beta_j),$$

where $\ell(\cdot,\cdot)$ is the negative log-likelihood function. Similar to the idea of the SIS, we can compute the vector of marginal utilities $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_p)$ and rank the components in an ascending order. Note that the GSIS is equivalent to the SIS when the response follows a linear regression model. Fan and Song (2010) showed that the above screening procedure enjoys the sure screening property.

To address the issue that the marginal regression may be nonlinear even when the true model is linear, Fan, Feng and Song (2011) further extended correlation learning to a nonparametric independence screening (NIS) approach. They considered the following p marginal nonparametric regression problems:

$$\min_{f_j \in L_2(P)} E\Big(y - f_j(X_j)\Big)^2,$$

where P denotes the joint distribution of $(\mathbf{x}, y)$, and $L_2(P)$ is the class of square integrable functions under the measure P. The solution to this problem is the projection of y onto X_j: $f_j = E(y|X_j)$. This motivates ones to create the marginal utility based on $Ef_j^2(X_j)$. To obtain the sample version of the marginal nonparametric regression, they employed a B-Spline basis; the optimization problem can then be expressed as

$$\min_{\boldsymbol{\beta}_j \in \mathcal{R}^m} \sum_{i=1}^{n} \Big(y_i - \boldsymbol{\Phi}_{ij}^T \boldsymbol{\beta}_j\Big)^2,$$

where $\boldsymbol{\Phi}_{ij} = (\Phi_1(X_{ij}), \cdots, \Phi_m(X_{ij}))^T$ is the m dimensional B-Spline basis func-

tions. Then we can select a submodel

$$\mathcal{M}_\gamma = \{1 \le j \le p : ||\hat{f}_{nj}||_n^2 \text{ is among the first } \gamma_n \text{ largest of all}\},$$

where $||\hat{f}_{nj}||_n^2 = n^{-1}\sum_{i=1}^n \hat{f}_{nj}(X_{ij})^2$ and γ_n again is a predetermined threshold value. Note that $||\hat{f}_{nj}||_n^2 = ||y\hat{f}_{nj}||_n$, hence the NIS can also be viewed as ranking the marginal correlation of $\{\hat{f}_{nj}(X_{ij})\}_{i=1}^n$ with the response $\{y_i\}_{i=1}^n$, a version of the correlation learning proposed by Fan and Lv (2008). Under some regularity conditions, it is shown that the NIS enjoys the sure screening property.

Simulation results illustrate that the SIS, NIS and GSIS work well with moderate sample sizes and large dimensions. Their good performances are all based upon the belief that the imposed working model is close to the underlying true model. However, these model assumptions are typically regarded as quite restrictive, particularly in an ultrahigh dimensional setting, because it is difficult to test the validity of these model assumptions. It is desirable to develop screening procedures that work well for a very general model setting, and have comparable performance to the SIS and NIS under linear regression models, and the GSIS under generalized linear models. Motivated by this goal, Zhu, Li, Li and Zhu (2010) proposed a model free sure independence ranking and screening (SIRS) procedure. Assume that $F(y^*|\mathbf{x})$ is the conditional distribution function of y given $\mathbf{x}$, and denote two index sets:

$$\mathcal{A} = \{j : F(y^*|\mathbf{x}) \text{ functionally depends on } X_j \text{ for some } y^* \in \Psi_{y^*}\},$$
$$\mathcal{I} = \{j : F(y^*|\mathbf{x}) \text{ does not functionally depend on } X_j \text{ for any } y^* \in \Psi_{y^*}\},$$

where Ψ_{y^*} is the support of the response variable. Without loss of generality, we

assume that $E(X_j) = 0$ and $Var(X_j) = 1$ for $j = 1, \cdots, p$. Denote $\mathbf{\Omega}(y^*) = E(\mathbf{x}F(y^*|\mathbf{x}))$, then

$$\mathbf{\Omega}(y^*) = E\big[\mathbf{x}E\{I(y < y^*)|\mathbf{x}\}\big] = cov\{\mathbf{x}, I(y < y^*)\}.$$

The marginal utility is defined to be $\lambda_j = E\{\Omega_j^2(y^*)\}$, $j = 1, \cdots, p$, where $\Omega_j(y^*)$ is the jth element of $\mathbf{\Omega}(y^*)$. The reason why $\mathbf{\Omega}(y^*)$ is adopted here as the marginal utility is motivated by the following fact. When $\mathbf{x} = (X_1, \cdots, X_p)'$ comes from a standard normal distribution with identity covariance matrix, then

$$\mathbf{\Omega}(y^*) = E\{\mathbf{x}F(y^*|\mathbf{x}'\boldsymbol{\beta})\} = c(y^*)\boldsymbol{\beta},$$

where $c(y^*)$ is a constant which only depends on y^*. Then $\lambda_j = E(c^2(y^*))\beta_j^2$, and if $E(c^2(y^*)) > 0$

$$\lambda_j = 0 \Leftrightarrow j \in \mathcal{I}.$$

In the sample version, λ_j can be estimated by

$$\hat{\lambda}_j = n^{-1}\sum_{i=1}^{n} r_j^2(y_i), \text{ where } r_j(y_i) = \sum_{k=1}^{n} X_{kj} I(y_k < y_i).$$

Under some regularity conditions, it can be shown that (the ranking consistency property):

$$\max_{j\in\mathcal{I}} \lambda_j < \min_{j\in\mathcal{A}} \lambda_j.$$

Furthermore, with probability tending to 1,

$$\max_{j\in\mathcal{I}} \hat{\lambda}_j < \min_{j\in\mathcal{A}} \hat{\lambda}_j.$$

Then we can select the submodel again by

$$\mathcal{M}_\gamma = \{1 \le j \le p : \hat{\lambda}_j \text{ is among the first } \gamma_n \text{ largest of all}\}.$$

Like all the other screening procedures introduced above, SIRS has the sure screening property.

After ranking predictors based on different marginal utilities, we still need a benchmark to separate active predictors and inactive predictors; that is, we need to determine what value γ_n should be in practise. Fan and Lv (2008) suggested to use the first $[n/\log(n)]$ marginal utilities, and this is called a hard-threshold cutoff. Zhu, Li, Li and Zhu (2010) proposed a soft-threshold cutoff by adding auxiliary variables. More specifically, one may generate a s-dimensional auxiliary variable $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_s)$ such that $\mathbf{z}$ is independent of both $\mathbf{x}$ and y. Regard the combined vector $(\mathbf{x}', \boldsymbol{z}')'$ as the predictors and y as the response, then the marginal utility $\hat{\lambda}_j$ can be calculated for $j = 1, \cdots, p+s$. Since it is known that $\mathbf{z}$ is inactive, then $C_s = \max_{l=1,\cdots,s} \hat{\lambda}_{p+l}$ can be treated as a benchmark that separates the active predictors from the inactive ones. In other words, the index set can be chosen as

$$\hat{\mathcal{A}} = \{j : \lambda_j > C_s\}.$$

Note that using marginal utilities may encounter some potential issues (Fan and Lv, 2008):

1. Some unimportant predictors can obtain high priority mainly due to the high correlation with important predictors.

2. Important predictors that are marginally uncorrelated but jointly correlated with the response will be missed.

3. The issue of collinearity makes the problem of variable selection more difficult.

To overcome these problems, Fan and Lv (2008) proposed an iterative procedure to more fully use the joint information of the covariates rather than just the marginal information. In particular:

Step 1: Select d_1 predictors ($d_1 < n$), and denote the set of indices by $\mathcal{A}_1$. Here, one may also apply a moderate scale variable selection method such as the SCAD, the LASSO, or the Dantzig selector. That is, we use SIS-SCAD or SIS-LASSO instead of SIS.

Step 2: Regress y over those d_1 predictors, and use residuals as the new responses. Apply the same method to the new response variable and the remaining $p - d_1$ predictors, and select d_2 predictors $\mathcal{A}_2$.

Step 3: Repeat above two steps until the union of l disjoint sets $\cup_{i=1}^{l}\mathcal{A}_i$ has a size exceeding a predetermined value.

Residuals from each iterative step have much weaker correlations with the remaining unimportant predictors that are highly correlated with important predictors. Also, those important predictors that are marginally weakly correlated with y mainly due to the presence of other important predictors should now be correlated with the residuals.

Zhu, Li, Li and Zhu (2011) also suggested an iterative procedure to enhance their methodology. The basic idea is similar to the iterative procedure of the

SIS. However, they adopted a slightly different way to obtain "residuals" since their method is model free. In Step 2, instead of regressing the response over d_1 predictors, they regressed the remaining predictors over d_1 predictors; that is, the predictor residual matrix is defined by

$$\mathbf{x}_r = \left\{ I_n - \mathbf{x}_{\mathcal{A}_1} \left(\mathbf{x}_{\mathcal{A}_1}^T \mathbf{x}_{\mathcal{A}_1} \right)^{-1} \mathbf{x}_{\mathcal{A}_1}^T \right\} \mathbf{x}_{\mathcal{I}_1},$$

where $\mathcal{I}_1$ is the complement of $\mathcal{A}_1$. Then the proposed screening approach can be applied to y and $\mathbf{x}_r$. The iterative procedure proceeds until the union of l disjoint sets $\cup_{i=1}^{l} \mathcal{A}_i$ has a size exceeding a predetermined value.

Chapter 3

Bias and Variance of Estimation for Covariate-adjusted and -unadjusted Approaches

We consider a study whose major objective is to assess the treatment effect. Suppose y is a response variable, $\mathbf{x}$ and $\mathbf{z}$ are two sets of covariates whose roles will be clarified shortly, and T is a treatment indicator variable that takes values 1 or -1 with probabilities p and $1-p$ respectively. We assume that the response variable y follows a generalized linear model (McCullagh and Nelder, 1983). Specifically, y has an exponential family probability density function

$$\exp[\{y\theta - b(\theta)\}/a(\phi) + c(u,\phi)]$$

for some functions $a(), b(), c()$, and a scalar ϕ. The mean of y, $\mu = b'(\theta)$ depends on the covariates through:

$$\mu = h(\eta), \quad \eta = \alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma} \cdot T + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2,$$

where $h^{-1}(\cdot)$ is the link function and ϕ is the dispersion parameter. Note that, although $\mathbf{x}$ and $\mathbf{z}$ are both vectors of covariates, $\mathbf{x}$ interacts with T while $\mathbf{z}$ does not. In this framework, the variance, $b''(\theta)a(\phi)$, can be defined as a function of the mean

$$Var(y) = V_\phi(\mu).$$

The above generalized linear model with

$$\mu = h(\alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma} \cdot T + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2), \tag{3.1}$$

is often called a covariate-adjusted model because it examines the treatment effect adjusting the impact (e.g. confounding and interaction effects) of the covariates. When $\boldsymbol{\gamma} = 0$ (i.e. there is no treatment and covariate interaction), then this model reduces to the model considered by Gail and Simon (1984), among others.

In contrast, the following model estimates a crude treatment effect which does not adjust the impact of covariates,

$$\mu^* = h(\eta^*), \quad \eta^* = \alpha^* + T\beta^*. \tag{3.2}$$

This model is often called the covariate-unadjusted generalized linar model.

The focus of the paper will be on comparing β, the treatment effect estimated from covariate-adjusted model (3.1) with β^*, the treatment effect estimated from the covariate-unadjusted model (3.2). In order to make sure β and β^* represent the same overall (average) treatment effect, throughout this paper, we assume that $\mathbf{x}$ and $\mathbf{z}$ are all centralized so that $E(\mathbf{x})$=0 and $E(\mathbf{z}) = 0$, similar to the assumption made in Gail and Simon (1984). In practice, when $E(\mathbf{x})$ and $E(\mathbf{z})$ are not all zero,

we transform $\mathbf{x}$ to $\mathbf{x} - E(\mathbf{x})$, $\mathbf{z}$ to $\mathbf{z} - E(\mathbf{z})$. Consequently, we re-parameterize α to $\alpha + E(\mathbf{x})'\boldsymbol{\xi}_1 + E(\mathbf{z})'\boldsymbol{\xi}_2$ and β to $\beta + E(\mathbf{x})'\gamma$.

Comparing model (3.1) with model (3.2) is not straightforward because we do not know the true underlying model. In some situations, covariates do not influence the response variable except through the treatment; in other words, covariates are causal intermediaries. In this case, model (3.2) is correct and (3.1) is theoretically correct if all $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are equal to 0. Therefore, the estimated treatment effects obtained from both models (3.1) and (3.2) are asymptotically unbiased.

If model (3.1) is the true model, then covariates do impact the response variable and are not causal intermediaries. When $\boldsymbol{\gamma} = 0$ (i.e. none of the covariates interact with treatment), the results by Gail et al. (1984) indicate that the estimated treatment effect obtained from model (3.2) can be asymptotically biased in the sense that the expected values of the estimated β and β^* differ.

The first goal of this chapter is to examine the bias when $\boldsymbol{\gamma} \neq 0$ under the assumption that covariates are independent of the treatment, a condition held in randomized studies. The second goal is to compare the variances of $\hat{\beta}^*$ and $\hat{\beta}$ in Section 3.2. Here, $\hat{\beta}^*$ and $\hat{\beta}$ are maximum likelihood estimation of β^* and β, respectively. In contrast to Section 3.1, the assumption that covariates are independent of the treatment is not a general requirement here. In particular, given the treatment and covariates, if y follows a linear model, then the variance of $\hat{\beta}^*$ is not less than that of $\hat{\beta}$. On the other hand, if y follows a logistic model, then the variance of $\hat{\beta}^*$ is not greater than that of $\hat{\beta}$; see Robinson and Jewell (1991). The comparison of the variances of $\hat{\beta}^*$ and $\hat{\beta}$ are unknown for generalized linear models other than linear or logistic regression models even when $\boldsymbol{\gamma} = 0$ (i.e. none of the covariates interact with treatment). Section 3.2 will reveal the comparison in generalized linear models.

3.1 Bias for covariate-adjusted and -unadjusted approaches

Following Gail et al. (1984), we let

$$\zeta_1 = E(y|T=1) = E_{\mathbf{x},\mathbf{z}}\Big\{h\Big(\alpha+\beta+\mathbf{x}'\boldsymbol{\gamma}+\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2\Big)\Big\},$$
$$\zeta_2 = E(y|T=-1) = E_{\mathbf{x},\mathbf{z}}\Big\{h\Big(\alpha-\beta-\mathbf{x}'\boldsymbol{\gamma}+\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2\Big)\Big\}.$$

Here $E_{\mathbf{x},\mathbf{z}}$ is the expectation with respect to $\mathbf{x}$ and $\mathbf{z}$. Let $\kappa(\eta) = \frac{\partial\mu}{\partial\eta}\cdot\frac{1}{V_\phi(\mu)}$ and recall that $\eta = \alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma}\cdot T + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2$.

Similar to Gail et al (1984), the maximum likelihood equations divided by the sample size converge to

$$E\big[\kappa(\eta)\cdot\big\{y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot T\cdot\big\{y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot\mathbf{x}'\cdot T\cdot\big\{y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot\mathbf{x}'\cdot\big\{y-h(\eta)\big\}\big]=0,$$
$$E\big[\kappa(\eta)\cdot\mathbf{z}'\cdot\big\{y-h(\eta)\big\}\big]=0.$$

Hereafter this thesis, we denote $\dot{h}(\cdot)$ and $\ddot{h}(\cdot)$ to be the first and second derivatives of function $h(\cdot)$, respectively. The relationship between β and β^* is given in the following result.

Theorem 3.1. *Under the following assumptions*

(a) T is independent of all covariate variables;

(b) $E\big[\kappa(\eta)\cdot\big\{y-h(\eta)\big\}\big]$, $E\big[\kappa(\eta)\cdot T\cdot\big\{y-h(\eta)\big\}\big]$, ζ_1, and ζ_2 exist;

(c) if $h(\cdot)$ has a unique inverse $h^{-1}(\cdot)$ which is well defined at ζ_1 and ζ_2; $\dot{h}$ and $\ddot{h}$

exist;

(d) $h^{-1}(\cdot)$ *is nonsingular at* $h\big(\alpha+\beta\big)$ *and* $h\big(\alpha-\beta\big)$*;*

(e) $\kappa(\alpha^*+\beta^*)$ *and* $\kappa(\alpha^*-\beta^*)$ *do not vanish,*

we have the following results

$$\alpha^* = \frac{1}{2}\{h^{-1}(\zeta_1) + h^{-1}(\zeta_2)\}, \tag{3.3}$$

$$\beta^* = \frac{1}{2}\{h^{-1}(\zeta_1) - h^{-1}(\zeta_2)\}, \tag{3.4}$$

and for small $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$,

$$\beta^* \doteq \beta + \left\{ \frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 + \mathbf{x}'\boldsymbol{\gamma})^2}{4} \frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)} - \frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 - \mathbf{x}'\boldsymbol{\gamma})^2}{4} \frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)} \right\}. \tag{3.5}$$

Results (3.3)-(3.5) are extensions of (2.5), (2.6) and (2.9) in Gail et al. (1984), which correspond to the case $\boldsymbol{\gamma} = 0$. As in Gail et al. (1984), the result (3.5) is derived for small $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$.

Corollary 3.1. *Under conditions of Theorem 3.1,*

- *If* $h(\eta) = \eta$ *(e.g. the linear model and the Poisson model with identity link),* $\beta^* = \beta$.

- *If* $h(\eta) = \exp(\eta)$ *(e.g. the Poisson model with canonical link),*

$$\beta^* = \beta + \frac{1}{2}\log\Big[E_{\mathbf{x},\mathbf{z}}\Big\{\exp(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 + \mathbf{x}'\boldsymbol{\gamma})\Big\}\Big] - \frac{1}{2}\log\Big[E_{\mathbf{x},\mathbf{z}}\Big\{\exp(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 - \mathbf{x}'\boldsymbol{\gamma})\Big\}\Big].$$

Particularly, with small $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ *and* $\boldsymbol{\xi}_2$,

$$\beta^* \doteq \beta + E(\mathbf{x}'\boldsymbol{\xi}_1 \cdot \mathbf{x}'\boldsymbol{\gamma}) + E_{\mathbf{x},\mathbf{z}}(\mathbf{z}'\boldsymbol{\xi}_2 \cdot \mathbf{x}'\boldsymbol{\gamma}).$$

- *If* $h(\eta) = \exp(\eta)/\{1 + \exp(\eta)\}$ *(e.g. the binary model with canonical link), and if* $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ *and* $\boldsymbol{\xi}_2$ *are small, then*

$$\beta^* \doteq \beta + \left\{ \frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 + \mathbf{x}'\boldsymbol{\gamma})^2}{4} \frac{1 - \exp\big(\alpha + \beta\big)}{1 + \exp\big(\alpha + \beta\big)} - \frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 - \mathbf{x}'\boldsymbol{\gamma})^2}{4} \frac{1 - \exp\big(\alpha - \beta\big)}{1 + \exp\big(\alpha - \beta\big)} \right\}. \tag{3.6}$$

The approximation in Theorem 3.1 is derived based on Taylor's expansion at $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = \boldsymbol{\gamma} = 0$. This is similar to that in Gail and Simon (1984). A different approximation is also derived based on Taylor's expansion at $\beta = 0$ below, as in Neuhaus and Jewell (1993). Let $k = 1, -1$, and

$$\eta_k = \alpha + \beta k + \mathbf{x}'\boldsymbol{\gamma} k + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2;$$
$$\eta_k^* = \alpha^* + \beta^* k.$$

Assume model (3.1) is the correct model; we know that

$$E(y|T = k, \mathbf{x}, \mathbf{z}) = h(\eta_k).$$

Denote by $f(\mathbf{x}, \mathbf{z}|T = k)$ the probability density function of $\mathbf{x}$ and $\mathbf{z}$ given k; then

$$E(y|T = k) = \int E(y|T = k, \mathbf{x}, \mathbf{z}) f(\mathbf{x}, \mathbf{z}|T = k)\ d\mathbf{x}d\mathbf{z}$$

$$= \int E(y|T=k,\mathbf{x},\mathbf{z}) f(\mathbf{x},\mathbf{z})\ d\mathbf{x}d\mathbf{z}. \tag{3.7}$$

On the other hand, if model (3.2) is used to make inferences, we would have misspecified the expectation of y

$$E(y|T=k) = h(\eta_k^*). \tag{3.8}$$

It follows by (3.7) and (3.8) that

$$\alpha^* + \beta^* k = h^{-1}\Big(\int h\big(\alpha + \beta k + \mathbf{x}'\boldsymbol{\gamma} k + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2\big) f(\mathbf{x},\mathbf{z})\ d\mathbf{x}d\mathbf{z}\Big),$$

and

$$\begin{aligned}
\beta^* &= \frac{1}{2}\Big\{\alpha^* + \beta^*\Big\} - \frac{1}{2}\Big\{\alpha^* - \beta^*\Big\} \\
&= \frac{1}{2}h^{-1}\Big(\int h\big(\alpha + \beta + \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2\big) f(\mathbf{x},\mathbf{z})\ d\mathbf{x}d\mathbf{z}\Big) \\
&- \frac{1}{2}h^{-1}\Big(\int h\big(\alpha - \beta - \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2\big) f(\mathbf{x},\mathbf{z})\ d\mathbf{x}d\mathbf{z}\Big).
\end{aligned}$$

We denote the function in the right side of the last equation as $H(\beta)$ for simplicity. By expanding $H(\beta)$ in Taylor series around $\beta = 0$, we have

$$\beta^* = H(\beta) \approx H(0) + \beta \cdot \dot{H}(0), \tag{3.9}$$

where

$$\begin{aligned}
H(0) = \frac{1}{2}h^{-1}\Big[E_{\mathbf{x},\mathbf{z}}\{h(\alpha + \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2)\}\Big] \\
- \frac{1}{2}h^{-1}\Big[E_{\mathbf{x},\mathbf{z}}\{h(\alpha - \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2)\}\Big],
\end{aligned}$$

and

$$\dot{H}(0) =$$
$$\frac{1}{2}\dot{h}^{-1}\Big[E\{h(\alpha+\mathbf{x}'\boldsymbol{\gamma}+\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2)\}\Big]\times E\Big[1/\dot{h}^{-1}\{h(\alpha+\mathbf{x}'\boldsymbol{\gamma}+\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2)\}\Big]$$
$$+\frac{1}{2}\dot{h}^{-1}\Big[E\{h(\alpha-\mathbf{x}'\boldsymbol{\gamma}+\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2)\}\Big]\times E\Big[1/\dot{h}^{-1}\{h(\alpha-\mathbf{x}'\boldsymbol{\gamma}+\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2)\}\Big].$$

Note that $\dot{h}^{-1}$ is the first derivative of h^{-1}.

When $\boldsymbol{\gamma}=0$, $H(0)=0$. However $H(0)$ may be nonzero when $\boldsymbol{\gamma}\neq 0$. Therefore, the existence of treatment and covariate interaction further complicates the relationship between β^* and β.

We now illustrate the results through two simple examples. The first example suggests that the covariate-unadjusted linear regression model provides an unbiased estimate despite the existence of treatment and covariate interaction. The second example, on the other hand, illustrates the important role of interactions between treatment and covariates in a nonlinear regression model.

Example 3.1. Suppose that the true model is $\theta=\eta=\alpha+T\beta+\mathbf{x}'\boldsymbol{\gamma}\cdot T+\mathbf{x}'\boldsymbol{\xi}_1$ and $h(.)$ is the identity function. In this case, $\beta=\beta^*$; therefore the covariate-unadjusted model, even omitting true interaction between covariate and treatment, still provides an unbiased treatment estimate. We also note that $H(0)=0$ and $H'(0)=1$.

Example 3.2. Let us consider a logistic regression model, in which the exponential family probability density function is

$$\exp[\{y\theta-\log(1+\exp(\theta))\}+c].$$

Here c is a constant. The mean of y, $\mu = b'(\theta) = \exp(\theta)\{1 + \exp(\theta)\}$ and

$$\mu = h(\theta) = \exp(\theta)\{1 + \exp(\theta)\}, \quad \theta = \alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma} \cdot T + \mathbf{x}'\boldsymbol{\xi}_1.$$

We consider a few different choices for $\boldsymbol{\xi}_1$ and $\boldsymbol{\gamma}$. The treatment indicator $T(1, -1)$ and $\mathbf{x}(1, -1)$ have independent binomial distribution Binomial(1,1/2). It can be shown that

$$\zeta_1 = 1/2\Big\{h\Big(\alpha + \beta + \boldsymbol{\gamma} + \boldsymbol{\xi}_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha + \beta - \boldsymbol{\gamma} - \boldsymbol{\xi}_1\Big)\Big\}$$
$$\zeta_2 = 1/2\Big\{h\Big(\alpha - \beta - \boldsymbol{\gamma} + \boldsymbol{\xi}_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha - \beta + \boldsymbol{\gamma} - \boldsymbol{\xi}_1\Big)\Big\}.$$

First we consider a special case with $\beta = 0$,

$$\zeta_1 = 1/2\Big\{h\Big(\alpha + \boldsymbol{\gamma} + \boldsymbol{\xi}_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha - \boldsymbol{\gamma} - \boldsymbol{\xi}_1\Big)\Big\}$$
$$\zeta_2 = 1/2\Big\{h\Big(\alpha - \boldsymbol{\gamma} + \boldsymbol{\xi}_1\Big)\Big\} + 1/2\Big\{h\Big(\alpha + \boldsymbol{\gamma} - \boldsymbol{\xi}_1\Big)\Big\}.$$

It becomes straightforward from equation (3.4) that $\beta^* = 0$ when $\boldsymbol{\gamma} = 0$; that is there is no interaction. On the other hand, if there is treatment and covariate interaction, β^* is generally not 0 unless $\boldsymbol{\xi}_1 = 0$. For example, $\beta^* = -0.035$ when $\alpha = 1$, $\boldsymbol{\gamma} = 0.2$ and $\boldsymbol{\xi}_1 = 0.4$.

Next, we consider a different case with $\alpha = \beta = 1$ and $\boldsymbol{\gamma} = 0$ and $\boldsymbol{\xi}_1 = 1$. According to Neuhause and Jewell (1993), $0 < \beta^* < 1$. Actually, we obtain $\beta^* = 0.836$. However, if $\boldsymbol{\gamma} \neq 0$, we no longer expect $0 < \beta^* < 1$. In fact, let $\alpha = \beta = \boldsymbol{\xi}_1 = 1$ and $\boldsymbol{\gamma} = -1$; then $\beta^* = 1$. In this case, the interaction $\boldsymbol{\gamma} = -1$ seems to correct the difference of -0.164 between β and β^* when $\boldsymbol{\gamma} = 0$ and $\boldsymbol{\xi}_1 = 1$.

Now, we consider another case in which $\alpha = \beta = 1$. It can be shown that

$\beta^* = 0.954$ if $\boldsymbol{\gamma} = 0$ and $\boldsymbol{\xi}_1 = 0.5$, and $\beta^* = 0.836$ if $\boldsymbol{\gamma} = 0.5$ and $\boldsymbol{\xi}_1 = 05$. In this case, the interaction shifted the difference -0.046 between β and β^* when $\boldsymbol{\gamma} = 0$ and $\boldsymbol{\xi}_1 = 0.5$ even further to -0.164.

Finally, we let $\alpha = 2$, $\beta = \boldsymbol{\xi} = 1$, and $\boldsymbol{\gamma} = -1$. It can be shown that $\beta^* = 1.275$ which is greater than $\beta = 1$. This is not consistent with the results described in Neuhause and Jewell (1993), indicating a difference occurs when treatment and covariate interact.

3.2 Variance of Estimation for Covariate-adjusted and -unadjusted Approaches

To present general results in a simple form, we first consider the following two models:

$$\eta = \alpha + T\beta + \mathbf{x}'\boldsymbol{\gamma}, \tag{3.10a}$$

and

$$\eta = \alpha^* + T\beta^*, \tag{3.10b}$$

where $\mathbf{x}$ and $\boldsymbol{\gamma}$ are $p \times 1$ vectors. Note that, here $\mathbf{x}$ is not necessarily independent of T. In other words, we allow components of $\mathbf{x}$ to contain interactions between T and covariates; therefore model (3.10a) and model (3.1) are similar.

Let t_i, x_i and y_i be the observed values of T, $\mathbf{x}$ and y, respectively, $i = 1, \cdots, n$. Denote $\widetilde{1}$ the $n \times 1$-vector with all components equal to 1, t the $n \times 1$-vector $[t_1, \cdots, t_n]'$ and x' the $p \times n$-matrix $[x_1, \cdots, x_n]$. Recall the procedure of fitting

generalized linear models via the Fisher scoring algorithm. The estimated covariance matrix is the inverse of the expected Hessian matrix $M'WM$, where M is the corresponding design matrix and W is the $n \times n$ diagonal matrix of iterative weights with diagonal terms $w_i = \{V_\phi(\mu_i)(\frac{\partial \eta_i}{\partial \mu_i})^2\}^{-1}$, $i = 1, \cdots, n$.

Let $\mu(T) = E(y|T)$, $\mu(T, \mathbf{x}) = E(y|T, \mathbf{x})$ and $g(\mu) = \left\{V_\phi(\mu)\left(\frac{\partial \eta}{\partial \mu}\right)^2\right\}^{-1}$. We also let W and W^* be the iterative weights matrix for models (3.10a) and (3.10b), respectively.

Lemma 3.1. *Suppose* $\lim_{n\to\infty} \frac{1}{n}(\widetilde{1}, T, \mathbf{x})'W(\widetilde{1}, T, \mathbf{x})$ *and* $\lim_{n\to\infty} \frac{1}{n}(\widetilde{1}, t)'W^*(\widetilde{1}, t)$ *exist, and* $E_T\Big[g\big\{\mu(T)\big\}\Big] < \infty$, $E_{T,\mathbf{x}}\Big[g\big\{\mu(T, \mathbf{x})\big\}\Big] < \infty$. *Further assume that the dispersion parameter* ϕ *is fixed and known; we have the following results:*

- *If* $g(\mu)$ *is a strictly concave (convex) function of* μ*, then* Δ*, defined as*

$$\lim_{n\to\infty} \frac{1}{n} \begin{pmatrix} \widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t \\ t'W^*\widetilde{1} & t'W^*t \end{pmatrix} - \lim_{n\to\infty} \frac{1}{n} \begin{pmatrix} \widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt \\ t'W\widetilde{1} & t'Wt \end{pmatrix}$$

  is either a positive (negative) definite matrix or a zero matrix. It is a zero matrix if and only if $\boldsymbol{\gamma} = 0$.

- *If* $g(\mu)$ *is a linear function of* μ*, then* Δ *is a zero matrix.*

Remark 3.1. Lemma 3.1 imposes an important assumption: the dispersion parameter ϕ is the same in both covariate-adjusted and covariate-unadjusted models. Consequently, Lemma 3.1 does not apply to some models, including linear regression models, where the dispersion parameter ϕ may differ in (3.10a) and (3.10b). In particular, for linear regression models, the dispersion parameter ϕ is exactly the conditional variance of the response variable, which has different values in the true model and the misspecified model.

Denote $I_1 = \lim_{n\to\infty} \frac{1}{n}(\widetilde{1}, T, \mathbf{x})' W (\widetilde{1}, T, \mathbf{x})$, $I_0 = \lim_{n\to\infty} \frac{1}{n}(\widetilde{1}, t)' W^* (\widetilde{1}, t)$. Let $\hat{\alpha}^*$, $\hat{\beta}^*$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{\boldsymbol{\gamma}}$ be MLE estimators of α^*, β^*, α, β, $\boldsymbol{\gamma}$ in (3.10a) and (3.10b), respectively. If (3.10a) is the true model, then

$$\sqrt{n}\left\{ \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\gamma} \end{pmatrix} - \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \right\} \xrightarrow{D} N\left(0, I_1^{-1}\right), \ \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, I_{1,22}^{-1}). \tag{3.11a}$$

If (3.10b) is the true model, then

$$\sqrt{n}\left\{ \begin{pmatrix} \hat{\alpha}^* \\ \hat{\beta}^* \end{pmatrix} - \begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix} \right\} \xrightarrow{D} N\left(0, I_0^{-1}\right), \ \sqrt{n}(\hat{\beta}^* - \beta^*) \xrightarrow{D} N(0, I_{0,22}^{-1}). \tag{3.11b}$$

Here $I_{i,22}^{-1}$ are the $(2,2)$ entries of I_i^{-1}, $i = 0, 1$.

We note that with misspecification, the asymptotic covariance matrix of the maximum likelihood estimation may not be the inverse of Fisher's information matrix. White (1982) showed that the covariance matrix is given by a sandwich form of matrix products. However, people often assume the model that they select to make inferences is the correct one, and statistical packages automatically construct the covariance matrix by the inverse of the information matrix. Therefore, it is interesting to compare $I_{1,22}^{-1}$ and $I_{0,22}^{-1}$. Furthermore, as shown in Neuhaus (1998), for binary regression models, the information matrix of the misspecified model provides the correct value of the covariance matrix when using the results of White (1982).

Theorem 3.2. *Under conditions of Lemma 3.1,*

- *If $g(\mu)$ is a strictly concave function of μ, then*

$$\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} > 1 \text{ if } \gamma \neq 0, \text{ and } \frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} \geq 1 \text{ if } \boldsymbol{\gamma} = 0.$$

  In addition, $I_{1,22}^{-1} = I_{0,22}^{-1}$ if $\mathbf{x}$ is independent of (y, T).

- *If $g(\mu)$ is constant, then*

  $\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} \geq 1$. *The equality occurs if and only if $\mathbf{x}$ and T are uncorrelated.*

In general, the relative efficiency function $I_{1,22}^{-1}/I_{0,22}^{-1}$ is complicated. To gain insights into Theorem 3.2, we compute this function for a Gamma distribution $\Gamma(\phi, \mu/\phi)$ $(\phi > 0)$ with log link.

Example 3.3. Assume y follows a Gamma distribution $\Gamma(\phi, \mu/\phi))$ with log link, where $\phi > 0$. For simplicity, we also assume $\mathbf{x}$ is a scalar in (3.10a). If ϕ is the same in (3.10a) and (3.10b), then $g(\mu) = \phi$. Following the proof of Theorem 3.2,

$$\begin{aligned}\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} =& 1 + \frac{E(\mathbf{x}T) - E(\mathbf{x})E(T)}{Var(T)} \times \Upsilon \times \frac{E(\mathbf{x}T) - E(\mathbf{x})E(T)}{Var(T)} \\ =& 1 + \frac{\rho_{T,\mathbf{x}}^2}{1 - \rho_{T,\mathbf{x}}^2},\end{aligned}$$

where

$$\Upsilon = \frac{Var(T)}{E(\mathbf{x}^2) - \frac{1}{Var(T)}\left[\{E(\mathbf{x}T) - E(\mathbf{x})E(T)\}^2 + E^2(\mathbf{x})Var(T)\right]}$$

and $\rho_{T,\mathbf{x}}$ is the simple correlation between T and $\mathbf{x}$.

For logistic regression models, the first part of Theorem 3.2 is consistent with and extends results in Robinson and Jewell (1991) and Robinson et al.(1998) where

treatment and covariates do not interact. Our results are applicable to many commonly used models for which $g(\mu)$ is strictly concave, such as binomial distribution with logit link, probit link, log-log link. The second part of Theorem 3.2 applies to some distributions (e.g. Gamma distribution with log link) in which $g(\mu)$ is a constant.

For Poisson regression with the log link, $g(\mu) = \mu$, we have the following interesting result.

Corollary 3.2. *If the underlying model is Poisson regression with the log link, and if T is independent of $\mathbf{x}$, then under conditions of Lemma 3.1, $I_{1,22}^{-1}/I_{0,22}^{-1} = 1$.*

Now let us revisit models (3.1) and (3.2). Again, let $\hat{\beta}$, $\hat{\gamma}$, $\hat{\beta}^*$ be maximum likelihood estimates of β, $\boldsymbol{\gamma}$, β^* in (3.1) and (3.2). The relative magnitude between $Var(\hat{\beta})$ and $Var(\hat{\beta}^*)$ can be obtained directly through Theorem 3.2. Let $I_{1,22}^{-1}$ be $\lim_{n\to\infty} n \cdot Var(\hat{\beta})$ under model (3.1). Let $I_{0,22}^{-1}$ be $\lim_{n\to\infty} n \cdot Var(\hat{\beta}^*)$ under model (3.2). Similar to the proof of Theorem 3.2, we can show the following results.

Theorem 3.3. *Suppose that T is independent of all covariate variables, and the conditions of Lemma 3.1 hold.*

- *If $g(\mu)$ is a strictly concave function of μ, then $I_{1,22}^{-1}/I_{0,22}^{-1} \geq 1$.*
- *If $g(\mu)$ is a constant function of μ, then $I_{1,22}^{-1}/I_{0,22}^{-1} = 1$.*

The results suggest that the variance of $\hat{\beta}$ is no smaller than that of $\hat{\beta}^*$. That is, the covariate-adjusted model provides an estimate of treatment effect which usually has a larger variance compared to that obtained from a covariate-unadjusted model.

3.3 Simulation Studies

The main purpose of these simulation studies are to confirm the main results developed in this chapter. In our simulation studies, we focus on the logistic regression model:

$$y \sim \text{Binomial}(1, p),\ \text{logit}(p) = \alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T.$$

Assuming $\alpha = 1$ and $\beta = 1$, we consider three different scenarios: (1) $\boldsymbol{\xi} = 0$, $\boldsymbol{\gamma} = 0$; (2) $\boldsymbol{\xi} = 0.5$, $\boldsymbol{\gamma} = 0$; and (3) $\boldsymbol{\xi} = 0.5$, $\boldsymbol{\gamma} = 0.5$. In each scenario, 1000 data sets of 200 per treatment group were generated to mimic a moderate clinical trial, in which T and $\mathbf{x}$ are independently generated. Specifically, each treatment ($T = 1, -1$) and covariate ($\mathbf{x} = 1, -1$) combination will have the same number of patients to achieve balance. Note that balance is not a necessary condition here to get the result. We use a balanced case mainly for simplifing the simulation. Since $P(T = i, \mathbf{x} = j) = .25$, for $i, j = 1, -1$, $\mathbf{x}$ and T are independent. Each of the 1000 data sets was fitted through three different models, either correct or misspecified:

$$\text{logit}(p) = \alpha + \beta T,\ \text{logit}(p) = \alpha + \beta T + \mathbf{x}\boldsymbol{\xi},\ \text{and logit}(p) = \alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T.$$

The parameters $\alpha = 1$, $\beta = 1$, and $\boldsymbol{\xi} = 0.5$ are the same as they were in Gail et al. (1984). Estimates of β and $\boldsymbol{\gamma}$, when applicable, are reported in Table 3.1.

In Table 3.1, the true model refers to the model that we used to generate the simulated datasets. The fitted models refer to the models that were used to fit the simulated datasets and the estimates of β and $\boldsymbol{\gamma}$ were results of model fitting. In some setups, the true model is simpler than the fitted model; that is, the fitted models included some variables which are not necessary. In this case, as we discussed in Section 2.1, the fitted model still provides an unbiased estimate. However, the estimates have larger variances. For example, the first true model in

Table 3.1 is $\eta = 1 + T$. When the true model is fitted, the standard error of $\hat{\beta}$ is 0.131. When the model $\eta = \alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$ is fitted, the standard error of $\hat{\beta}$ is 0.134.

In some other setups, the fitted models are simpler than the true model; therefore the fitted models are misspecified. The second true model in Table 3.1 is $\eta = 1 + T + 0.5\mathbf{x}$. When the model $\eta = \alpha + \beta T$ (more formally, $\eta = \alpha^* + \beta^* T$) is fitted, the estimated β^* is 0.963 with standard error 0.128 when the sample size is 200 per treatment group. Table 3.1 also provides the prediction of β, which is computed through equation (3.4) (i.e. $\beta^* = \frac{1}{2}\{h^{-1}(\zeta_1) - h^{-1}(\zeta_2)\}$). The predicted value is 0.954, which is close to 0.963 the sample mean. When the true model is fitted, $\hat{\beta}$ does not bear bias, but it has a larger standard error 0.148.

Table 3.1 also presents results for a smaller sample size, where 1000 data sets of 100 per treatment group were generated. Each treatment ($T = 1, -1$) and covariate ($\mathbf{x} = 1, -1$) combination will have the same number of patients, which gives similar conclusions.

Table 3.1 demonstrates that an unadjusted model could lead to results that are difficult to interpret if we ignore a true interaction. When treatment and covariate do not interact, our simulation results are consistent with what Gail et al. (1984)'s result suggested, namely that treatment effect bears small bias when the covariate effect is small. For example, with $(\boldsymbol{\xi}, \boldsymbol{\gamma}) = (0.5, 0)$ in the true model and 200 subjects per treatment group, the estimate of β through the unadjusted model is 0.963.

On the other hand, when treatment and covariate do interact, our results suggest that smaller covariate effects may lead to large treatment effect, different than the phenomenon described in Gail et al. (1984). For example, with $(\boldsymbol{\xi}, \boldsymbol{\gamma}) = (0.5, 0.5)$ in the true model and 200 subjects per treatment group, the

estimate of β through the unadjusted model is 0.843. For the same case when the sample is 100 subjects per treatment group, the estimate of β through the unadjusted model is 0.842. We notice the magnitude of bias is not very small despite the small effects of $(\boldsymbol{\xi}, \boldsymbol{\gamma}) = (0.5, 0.5)$. When $(\boldsymbol{\xi}, \boldsymbol{\gamma}) = (1, 0)$ in the true model, the estimate of β through the unadjusted model is 0.842, in which the bias increases considerably from the one that appeared in the case where $(\boldsymbol{\xi}, \boldsymbol{\gamma}) = (0.5, 0)$. Provided that $(\boldsymbol{\xi}, \boldsymbol{\gamma}) = (1, 1)$ in the true model, the estimate of β through the unadjusted model is 0.527, which is again very different from the case in which $(\boldsymbol{\xi}, \boldsymbol{\gamma}) = (0.5, 0.5)$.

The results are generally consistent when $\boldsymbol{\xi} = \boldsymbol{\gamma} = 1$ (i.e. $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$ have larger effects). However, we note that $P(Y = 0) = 1/(\exp(4)) = 2\%$ for $\mathbf{x} = T = 1$ when data are generated through model $\text{logit}(Y) = 1 + T + \mathbf{x} + \mathbf{x}T$. Therefore, estimation may not realizable when sample size is small. Consequently, when the sample size is 100 per group, the estimation appeared to be poor. This phenomenon could happen in clinical trials when the primary endpoint is a rare event.

Table 3.1 also confirms the theoretical results in Theorem 3.2. When the unadjusted model is correct, the estimate of β through an adjusted model is valid with almost the same standard error. However, when an adjusted model is correct with $\boldsymbol{\xi} \neq 0$, the standard error of the treatment effect estimate obtained from the unadjusted model is always smaller.

3.4 Real Data Illustration

A Phase III randomized, double-blind, placebo-controlled clinical trial was conducted from 1996 to 1997 to evaluate the safety and efficacy of prophylaxis with Palivizumab in reduction of respiratory syncytial virus (RSV) infection in high-risk

Table 3.1. Logistic regression model with $E(\mathbf{x}) = 0$

True Model	Fitted Model	Estimate of β	Prediction of β
Sample Size: 200 per treatment group			
$1 + T + 0\mathbf{x} + 0\mathbf{x}T$	$\alpha + \beta T$	1.010(0.131)	1.000
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	1.012(0.132)	1.000
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	1.021(0.134)	1.000
$1 + T + 0.5\mathbf{x} + 0\mathbf{x}T$	$\alpha + \beta T$	0.963(0.128)	0.954
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	1.011(0.135)	1.000
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	1.024(0.148)	1.000
$1 + T + 0.5\mathbf{x} + 0.5\mathbf{x}T$	$\alpha + \beta T$	0.843(0.121)	0.836
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	0.859(0.121)	.
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	1.030(0.159)	1.000
$1 + T + \mathbf{x} + \mathbf{x}T$	$\alpha + \beta T$	0.527(0.096)	0.526
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	0.566(0.100)	.
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	0.996(0.157)	1.000
Sample Size: 100 per treatment group			
$1 + T + 0\mathbf{x} + 0\mathbf{x}T$	$\alpha + \beta T$	1.018(0.182)	1.000
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	1.024(0.183)	1.000
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	1.042(0.192)	1.000
$1 + T + 0.5\mathbf{x} + 0\mathbf{x}T$	$\alpha + \beta T$	0.967(0.183)	0.954
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	1.019(0.191)	1.000
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	1.036(0.203)	1.000
$1 + T + 0.5\mathbf{x} + 0.5\mathbf{x}T$	$\alpha + \beta T$	0.842(0.170)	0.826
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	0.860(0.170)	.
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	1.019(0.195)	1.000
$1 + T + \mathbf{x} + \mathbf{x}T$	$\alpha + \beta T$	0.515(0.136)	0.526
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi}$	0.552(0.142)	.
	$\alpha + \beta T + \mathbf{x}\boldsymbol{\xi} + \mathbf{x}\boldsymbol{\gamma}T$	0.891(0.162)	1.000

infants. A total of 1502 children with prematurity or bronchopulmonary dysplasia (BPD) were randomized to receive either palivizumab or placebo intramuscularly. The primary endpoint was RSV-related hospitalization within 150 days after administration of the first dose of treatment. For more information about this trial please refer to IMpact-RSV Study Group (1998).

Among the 500 subjects who received placebo, 53 (10.6%) had an RSV-related hospitalization; among the 1002 subjects who received palivizumab, 48 (4.8%) had an RSV-related hospitalization.

Without considering any confounders, we use model (3.1) to estimate the odds ratio of Placebo vs Palivizumab. The status of RSV-related hospitalization for the i^{th} subject, y_i, is modeled through the following logistic regression model:

$$y_i \sim \text{Binomial}(1, p_i),\ \text{logit}(p_i) = \alpha^* + \beta^* T_i,\ \ i = 1, \cdots, n$$

with $n = 1502$. Here $T_i = 1$ or -1 if the ith subject took Palivizumab or Placebo, respectively. Note that the parameter β^* is half of the log odds ratio.

Maximum likelihood estimates and standard errors of the parameters are $\hat{\alpha}^* = -2.561(0.104)$ and $\hat{\beta}^* = -0.429(0.104)$. Therefore, the response rates for experimental and standard treatment differ significantly with a two sided p-value smaller than 0.0001.

It worth noting that the population of this trial included exclusively two disjointed subgroups: 1) children 24 months old or younger with a clinical diagnosis of BPD requiring ongoing medical treatment; and 2) children with 35 weeks gestation or less and 6 months old or younger, who did not have a clinical diagnosis of BPD. Among patients enrolled with a diagnosis of BPD, the incidence rate of RSV-related hospitalization is 12.8% (34/266) in the placebo arm and 7.9% (39/496) in the Palivizumab arm. Among patients enrolled without a diagnosis of BPD, the

the incidence rate of RSV-related hospitalization is 8.1% (19/234) in the placebo arm and 1.8% (9/506) in the Palivizumab arm.

Understanding the heterogeneity of treatment effect size plays important role in improving treatments and developing new targeted treatments. For this purpose, we continue to analyze the data to incorporate each subject's BPD status. Let $\mathbf{x}_i = 1$ or 0 if the i^{th} subject had a diagnosis of BPD or not, respectively. As we explained before, in order to make β and β^* have the same meaning, we centralize $\mathbf{x}$ so that $E(\mathbf{x}) = 0$. In this example, we therefore subtract $\mathbf{x}_i$ from 0.507, the sample mean of $\mathbf{x}$. Let $\mathbf{x}_i^* = 0.493$ or -0.507 if the i^{th} subject had a diagnosis of BPD or not, respectively. The following model is therefore fitted:

$$y_i \sim \text{Binomial}(1, p_i),\ \text{logit}(p_i) = \alpha + \beta T_i + \mathbf{x}^*\boldsymbol{\xi}_i + \mathbf{x}^*\boldsymbol{\gamma}_i T_i,\ i = 1, \cdots, n.$$

Maximum likelihood estimates and standard errors of the parameters are $\hat{\alpha} = -2.697(0.120)$, $\hat{\beta} = -.528(0.120)$, $\hat{\boldsymbol{\xi}} = 1.028(0.241)$ and $\hat{\boldsymbol{\gamma}} = 0.522(0.241)$. The interaction between $\mathbf{x}$ and T is statistically significant with the p-value 0.03.

In addition to the logistic model, we also used log and probit link functions for the binomial distribution. The results are given in Table 3.2. From the results, we notice that the covariate adjusted and unadjusted approaches usually give different point estimates. In addition, the results confirm that $Var(\hat{\beta}^*)$ is smaller than $Var(\hat{\beta})$, which is always true whether the covariate does or does not interact with treatment.

Table 3.2. Treatment effects under different models and link functions for Palivizumab study

Model	Log Link	Logit Link	Probit Link
$\eta = \alpha^* + \beta^* T_i$	-.397 (.096)	-.429(.104)	.-209(.051)
$\eta = \alpha + \beta T_i + \mathbf{x}\boldsymbol{\xi}_i$	-.375 (.095)	-.418(.104)	-.212(.052)
$\eta = \alpha + \beta T_i + \mathbf{x}\boldsymbol{\xi}_i^* + \mathbf{x}\boldsymbol{\gamma}_i^* T_i$	-.497 (.113)	-.528(.120)	-.244(.055)

3.5 Theoretical Proofs

3.5.1 Proof of Theorem 3.1.

Parameters in the misspecified model satisfy the following equations:

$$\begin{aligned}
&E\Big[\kappa(\eta^*)\cdot\Big\{y-h(\eta^*)\Big\}\Big]\\
=&p\cdot\kappa(\alpha^*+\beta^*)\cdot\zeta_1+(1-p)\cdot\kappa(\alpha^*-\beta^*)\cdot\zeta_2\\
&-p\cdot\kappa(\alpha^*+\beta^*)\cdot h\big(\alpha^*+\beta^*\big)-(1-p)\cdot\kappa(\alpha^*-\beta^*)\cdot h\big(\alpha^*-\beta^*\big)=0, \quad (3.12a)
\end{aligned}$$

$$\begin{aligned}
&E\Big[\kappa(\eta^*)\cdot T\cdot\Big\{y-h(\eta^*)\Big\}\Big]\\
=&p\cdot\kappa(\alpha^*+\beta^*)\cdot\zeta_1-(1-p)\cdot\kappa(\alpha^*-\beta^*)\cdot\zeta_2\\
&-p\cdot\kappa(\alpha^*+\beta^*)\cdot h\big(\alpha^*+\beta^*\big)+(1-p)\cdot\kappa(\alpha^*-\beta^*)\cdot h\big(\alpha^*-\beta^*\big)=0. \quad (3.12b)
\end{aligned}$$

Since $\kappa(\alpha^*+\beta^*)$ and $\kappa(\alpha^*-\beta^*)$ do not vanish, solutions to (3.12a) and (3.12b) are given by

$$\begin{cases} h(\alpha^*+\beta^*)=\zeta_1 \\ \\ h(\alpha^*-\beta^*)=\zeta_2 \end{cases}, \text{ or equivalently } \begin{cases} \beta^*=\frac{1}{2}\{h^{-1}(\zeta_1)-h^{-1}(\zeta_2)\} \\ \\ \alpha^*=\frac{1}{2}\{h^{-1}(\zeta_1)+h^{-1}(\zeta_2)\} \end{cases}.$$

By Taylor's expansion, for small $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, ζ_1 and ζ_2 can be approximated by

$$\begin{aligned}
\zeta_1 \doteq\, & h\big(\alpha+\beta\big)+\dot{h}\big(\alpha+\beta\big)E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2+\mathbf{x}'\boldsymbol{\gamma})\\
&+\frac{1}{2}\ddot{h}\big(\alpha+\beta\big)E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2+\mathbf{x}'\boldsymbol{\gamma})^2,\\
\zeta_2 \doteq\, & h\big(\alpha-\beta\big)+\dot{h}\big(\alpha-\beta\big)E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2-\mathbf{x}'\boldsymbol{\gamma})
\end{aligned}$$

$$+\frac{1}{2}\ddot{h}\big(\alpha-\beta\big)E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2-\mathbf{x}'\boldsymbol{\gamma})^2.$$

Note that

$$\begin{aligned}h^{-1}(\zeta_1)\doteq& h^{-1}\big\{h\big(\alpha+\beta\big)\big\}+\dot{h}^{-1}\big\{h(\alpha+\beta)\big\}\Big\{E_{\mathbf{x},\mathbf{z}}\big(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2+\mathbf{x}'\boldsymbol{\gamma}\big)\dot{h}\big(\alpha+\beta\big)\\&+\frac{E_{\mathbf{x},\mathbf{z}}\big(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2+\mathbf{x}'\boldsymbol{\gamma}\big)^2}{2}\ddot{h}\big(\alpha+\beta\big)\Big\}\\=&\alpha+\beta+\Bigg\{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2+\mathbf{x}'\boldsymbol{\gamma})\\&+\frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2+\mathbf{x}'\boldsymbol{\gamma})^2}{2}\frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)}\Bigg\}.\end{aligned}$$

Similarly,

$$\begin{aligned}h^{-1}(\zeta_2)\doteq\alpha-\beta+&\Bigg\{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2-\mathbf{x}'\boldsymbol{\gamma})\\&+\frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2-\mathbf{x}'\boldsymbol{\gamma})^2}{2}\frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)}\Bigg\}.\end{aligned}$$

Therefore for small $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, the asymptotic bias is

$$\begin{aligned}E(\mathbf{x}'\boldsymbol{\gamma})+&\Bigg\{\frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2+\mathbf{x}'\boldsymbol{\gamma})^2}{4}\frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)}\\&-\frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1+\mathbf{z}'\boldsymbol{\xi}_2-\mathbf{x}'\boldsymbol{\gamma})^2}{4}\frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)}\Bigg\}.\end{aligned}$$

This completes the proof. □

3.5.2 Proof of Corollary 3.1

(1). If $h(\eta) = \eta$, then

$$h^{-1}(\zeta_1) = E_{\mathbf{x},\mathbf{z}}(\alpha + \beta + \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2),$$
$$h^{-1}(\zeta_2) = E_{\mathbf{x},\mathbf{z}}(\alpha - \beta - \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2).$$

Substituting those values into (3.3), hence $\beta^* = \beta + E(\mathbf{x}'\boldsymbol{\gamma})$.

(2). If $h(\eta) = \exp(\eta)$, then

$$h^{-1}(\zeta_1) = \log\Big[E_{\mathbf{x},\mathbf{z}}\Big\{\exp(\alpha + \beta + \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2)\Big\}\Big],$$
$$h^{-1}(\zeta_2) = \log\Big[E_{\mathbf{x},\mathbf{z}}\Big\{\exp(\alpha - \beta - \mathbf{x}'\boldsymbol{\gamma} + \mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2)\Big\}\Big].$$

Substituting those values into (3.3), it follows that

$$\beta^* = \beta + \log\Big[E_{\mathbf{x},\mathbf{z}}\Big\{\exp(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 + \mathbf{x}'\boldsymbol{\gamma})\Big\}\Big] - \log\Big[E_{\mathbf{x},\mathbf{z}}\Big\{\exp(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 - \mathbf{x}'\boldsymbol{\gamma})\Big\}\Big].$$

Note that $\ddot{h} = \dot{h} = h$, then

$$\Big\{\frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 + \mathbf{x}'\boldsymbol{\gamma})^2}{4}\frac{\ddot{h}\big(\alpha+\beta\big)}{\dot{h}\big(\alpha+\beta\big)} - \frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 - \mathbf{x}'\boldsymbol{\gamma})^2}{4}\frac{\ddot{h}\big(\alpha-\beta\big)}{\dot{h}\big(\alpha-\beta\big)}\Big\}$$
$$= E(\mathbf{x}'\boldsymbol{\xi}_1 \cdot \mathbf{x}'\boldsymbol{\gamma}) + E_{\mathbf{x},\mathbf{z}}(\mathbf{z}'\boldsymbol{\xi}_2 \cdot \mathbf{x}'\boldsymbol{\gamma}).$$

Therefore, when $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are small, the approximate bias in (3.5) is

$$E(\mathbf{x}'\boldsymbol{\gamma}) + E(\mathbf{x}'\boldsymbol{\xi}_1 \cdot \mathbf{x}'\boldsymbol{\gamma}) + E_{\mathbf{x},\mathbf{z}}(\mathbf{z}'\boldsymbol{\xi}_2 \cdot \mathbf{x}'\boldsymbol{\gamma}).$$

(3). If $h(\eta) = \exp(\eta)/\{1+\exp(\eta)\}$, then $\dot{h} = h(1-h)$, $\ddot{h} = h(1-h)(1-2h)$ and $0 < h(.) < 1$. Therefore, when $\boldsymbol{\gamma}$, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are small, $|\beta^* - \beta - E(\mathbf{x}'\boldsymbol{\gamma})|$ is bounded by

$$\frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 + \mathbf{x}'\boldsymbol{\gamma})^2}{4} + \frac{E_{\mathbf{x},\mathbf{z}}(\mathbf{x}'\boldsymbol{\xi}_1 + \mathbf{z}'\boldsymbol{\xi}_2 - \mathbf{x}'\boldsymbol{\gamma})^2}{4}.$$

This completes the proof. □

3.5.3 Proof of Lemma 3.1.

By the weak law of large numbers, it follows that

$$\lim_{n\to\infty} \frac{1}{n}\widetilde{1}'W^*\widetilde{1} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g\{\mu(t_i)\} = E_T\Big[g\big\{\mu(T)\big\}\Big].$$

Similarly, it follows that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\widetilde{1}'W\widetilde{1} = E_{T,\mathbf{x}}\Big[g\{\mu(T,\mathbf{x})\}\Big], && \lim_{n\to\infty} \frac{1}{n}\widetilde{1}'W^*t = E_T\Big[T\cdot g\big\{\mu(T)\big\}\Big],\\
&\lim_{n\to\infty} \frac{1}{n}\widetilde{1}'Wt = E_{T,\mathbf{x}}\Big[T\cdot g\{\mu(T,\mathbf{x})\}\Big], && \lim_{n\to\infty} \frac{1}{n}t'W^*t = E_T\Big[T^2\cdot g\big\{\mu(T)\big\}\Big],\\
&\lim_{n\to\infty} \frac{1}{n}t'Wt = E_{T,\mathbf{x}}\Big[T^2\cdot g\{\mu(T,\mathbf{x})\}\Big].
\end{aligned}$$

(1). If $g(\mu)$ is a strictly concave function of μ. By Jensen's inequality,

$$\begin{aligned}
E_{T,\mathbf{x}}\big[g\big\{\mu(T,\mathbf{x})\big\}\big] &= E_T\Big[E_X\Big\{g\Big(\mu(T,\mathbf{x})\Big)\Big|T\Big\}\Big]\\
&\le E_T\Big[g\Big\{E_X\Big(\mu(T,\mathbf{x})\Big|T\Big)\Big\}\Big] = E_T\big[g\big\{\mu(T)\big\}\big].
\end{aligned}$$

The equality occurs if and only if $\mathbf{x}$ is independent of y given T (i.e., $\boldsymbol{\gamma} = 0$).

Since $E_T\big[g\{\mu(T)\}\big] = E_{T,\mathbf{x}}\big[g\{\mu(T)\}\big]$,

$$\begin{aligned}
\Delta &= \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix} \widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t \\ t'W^*\widetilde{1} & t'W^*t \end{pmatrix} - \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix} \widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt \\ t'W\widetilde{1} & t'Wt \end{pmatrix} \\
&= \begin{pmatrix} E_T\big[g\{\mu(T)\}\big] & E_T\big[T\cdot g\{\mu(T)\}\big] \\ E_T\big[T\cdot g\{\mu(T)\}\big] & E_T\big[T^2\cdot g\{\mu(T)\}\big] \end{pmatrix} \\
&\qquad - \begin{pmatrix} E_{T,\mathbf{x}}\big[g\{\mu(T,\mathbf{x})\}\big] & E_{T,\mathbf{x}}\big[T\cdot g\{\mu(T,\mathbf{x})\}\big] \\ E_{T,\mathbf{x}}\big[T\cdot g\{\mu(T,\mathbf{x})\}\big] & E_{T,\mathbf{x}}\big[T^2\cdot g\{\mu(T,\mathbf{x})\}\big] \end{pmatrix} \\
&= E_{T,\mathbf{x}}\Big[\begin{pmatrix} 1 \\ T \end{pmatrix}\cdot\Big\{g\Big(\mu(T)\Big) - g\Big(\mu(T,\mathbf{x})\Big)\Big\}\cdot\begin{pmatrix} 1 & T \end{pmatrix}\Big],
\end{aligned}$$

which is either a positive definite matrix or a zero matrix. The latter occurs if and only if $\boldsymbol{\gamma} = 0$.

(2). If $g(\mu)$ is a strictly convex function of μ, then following a proof similar to the one above,

$$E_{T,\mathbf{x}}\big[g\{\mu(T,\mathbf{x})\}\big] \geq E_T\big[g\{\mu(T)\}\big].$$

Therefore,

$$\begin{aligned}
\Delta &= \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix} \widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t \\ t'W^*\widetilde{1} & t'W^*t \end{pmatrix} - \lim_{n\to\infty}\frac{1}{n}\begin{pmatrix} \widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt \\ t'W\widetilde{1} & t'Wt \end{pmatrix} \\
&= E_{T,\mathbf{x}}\Big[\begin{pmatrix} 1 \\ T \end{pmatrix}\cdot\Big\{g\Big(\mu(T)\Big) - g\Big(\mu(T,\mathbf{x})\Big)\Big\}\cdot\begin{pmatrix} 1 & T \end{pmatrix}\Big],
\end{aligned}$$

which is either a negative definite matrix or a zero matrix. It is a zero matrix if

and only if $\boldsymbol{\gamma} = 0$.

(3). If $g(\mu)$ is a linear function of μ, then

$$E_{T,\mathbf{x}}\Big[T^i \cdot g\Big\{\mu(T,\mathbf{x})\Big\}\Big] = E_T\Big[T^i \cdot g\Big\{\mu(T)\Big\}\Big],\ i = 0, 1, 2.$$

Thus,

$$\Delta = \lim_{n\to\infty} \frac{1}{n} \begin{pmatrix} \widetilde{1}'W^*\widetilde{1} & \widetilde{1}'W^*t \\ t'W^*\widetilde{1} & t'W^*t \end{pmatrix} - \lim_{n\to\infty} \frac{1}{n} \begin{pmatrix} \widetilde{1}'W\widetilde{1} & \widetilde{1}'Wt \\ t'W\widetilde{1} & t'Wt \end{pmatrix} = 0.$$

This completes the proof. □

3.5.4 Proof of Theorem 3.2.

Partition I_1 as

$$I_1 = \begin{pmatrix} A & b' \\ b & c \end{pmatrix},$$

where

$$A = \begin{pmatrix} E_{T,\mathbf{x}}\big[g\{\mu(T,\mathbf{x})\}\big] & E_{T,\mathbf{x}}\big[T \cdot g\{\mu(T,\mathbf{x})\}\big] \\ E_{T,\mathbf{x}}\big[T \cdot g\{\mu(T,\mathbf{x})\}\big] & E_{T,\mathbf{x}}\big[T^2 \cdot g\{\mu(T,\mathbf{x})\}\big] \end{pmatrix} \text{ is } 2 \times 2,$$

$b = \Big(E_{T,\mathbf{x}}\big[X \cdot g\{\mu(T,\mathbf{x})\}\big] \quad E_{T,\mathbf{x}}\big[X \cdot T \cdot g\{\mu(T,\mathbf{x})\}\big]\Big)$ is a $p \times 2$ matrix and c is a $p \times p$ matrix. Let $\Delta = I_0 - A$.

(1). First consider the case where $g(\mu)$ is a strictly concave function of μ. If $\boldsymbol{\gamma} \neq 0$,

by Lemma 1, A, I_0 and Δ are all positive definite matrices. Thus,

$$\begin{aligned}\Big(A - b'c^{-1}b\Big)^{-1} &= A^{-1} + A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1} \\ &= (I_0 - \Delta)^{-1} + A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1} \\ &= I_0^{-1} + I_0^{-1}\big(\Delta^{-1} - I_0^{-1}\big)^{-1}I_0^{-1} + A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1}.\end{aligned}$$

Note that $\big(\Delta^{-1} - I_0^{-1}\big)^{-1} > 0$. This results in

$$\begin{aligned}I_{1,22}^{-1} - I_{0,22}^{-1} = \begin{pmatrix}0 & 1\end{pmatrix} I_0^{-1}\big(\Delta^{-1} - I_0^{-1}\big)^{-1}I_0^{-1}\begin{pmatrix}0\\1\end{pmatrix} \\ + \begin{pmatrix}0 & 1\end{pmatrix} A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1}\begin{pmatrix}0\\1\end{pmatrix} > 0.\end{aligned}$$

If $\boldsymbol{\gamma} = 0$, by Lemma 3.1, $I_0 = A$. This leads to

$$I_{1,22}^{-1} - I_{0,22}^{-1} = \begin{pmatrix}0 & 1\end{pmatrix} A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1}bA^{-1}\begin{pmatrix}0\\1\end{pmatrix} \geq 0,$$

or equivalently,

$$\frac{I_{1,22}^{-1}}{I_{0,22}^{-1}} \geq 1$$

Note that under the condition of $\boldsymbol{\gamma} = 0$, $g\big\{\mu(T, \mathbf{x})\big\} = g\big\{\mu(T)\big\}$ almost surely; then

$$\begin{pmatrix}0 & 1\end{pmatrix} A^{-1}b' = 0$$

$$\Leftrightarrow$$

$$\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} E_T\big[T^2 \cdot g\{\mu(T)\}\big] & -E_T\big[T \cdot g\{\mu(T)\}\big] \\ -E_T\big[T \cdot g\{\mu(T)\}\big] & E_T\big[g\{\mu(T)\}\big] \end{pmatrix} \times \begin{pmatrix} E_{T,\mathbf{x}}\big[X' \cdot g\{\mu(T)\}\big] \\ E_{T,\mathbf{x}}\big[X' \cdot T \cdot g\{\mu(T)\}\big] \end{pmatrix} = 0$$

$\Leftrightarrow$

$$E_T\big[g\{\mu(T)\}\big] \cdot E_{T,\mathbf{x}}\big[X' \cdot T \cdot g\{\mu(T)\}\big] = E_T\big[T \cdot g\{\mu(T)\}\big] \cdot E_{T,\mathbf{x}}\big[X' \cdot g\{\mu(T)\}\big]. \quad (3.13)$$

The last identity (3.13) holds if $\mathbf{x}$ and T are independent. Hence, $I^{-1}_{1,22} = I^{-1}_{0,22}$ if $\boldsymbol{\gamma} = 0$ and $\mathbf{x}$ is independent of T.

(2). We then consider the case where $g(\mu)$ is constant with respect to μ, by Lemma 3.1, $I_0 = A$. Following the proof above, $I^{-1}_{1,22} \geq I^{-1}_{0,22}$. The equality occurs if and only if $\begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1} b' = 0$. Since $g(\mu)$ is constant with respect to μ,

$$\begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1} b' = 0$$

$$\Leftrightarrow \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} E(T^2) & -E(T) \\ -E(T) & 1 \end{pmatrix} \begin{pmatrix} E(X') \\ E_{T,\mathbf{x}}(X' \cdot T) \end{pmatrix} = 0$$

$$\Leftrightarrow E_{T,\mathbf{x}}(X' \cdot T) = E(T) \cdot E(X') \Leftrightarrow \ X \text{ and } T \text{ are uncorrelated.}$$

This completes the proof. $\square$

3.5.5 Proof of Corollary 3.2.

Partition I_1 in the same way as in the proof of Theorem 3.2,

$$I_1 = \begin{pmatrix} A & b' \\ b & c \end{pmatrix}.$$

Since $g(\mu) = \mu$, by Lemma 3.1, $I_0 = A$. Hence

$$I_{1,22}^{-1} - I_{0,22}^{-1} = \begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1}b'\Big(c - bA^{-1}b'\Big)^{-1} bA^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

However

$$\begin{pmatrix} 0 & 1 \end{pmatrix} A^{-1}b' = 0$$

$$\Leftrightarrow$$

$$\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} E_{T,\mathbf{x}}\big\{T^2 \cdot \mu(T,\mathbf{x})\big\} & -E_{T,\mathbf{x}}\big\{T \cdot \mu(T,\mathbf{x})\big\} \\ -E_{T,\mathbf{x}}\big\{T \cdot \mu(T,\mathbf{x})\big\} & E_{T,\mathbf{x}}\big\{\mu(T,\mathbf{x})\big\} \end{pmatrix} \begin{pmatrix} E_{T,\mathbf{x}}\big\{X' \cdot \mu(T,\mathbf{x})\big\} \\ E_{T,\mathbf{x}}\big\{X' \cdot T \cdot \mu(T,\mathbf{x})\big\} \end{pmatrix} = 0$$

$$\Leftrightarrow$$

$$E_{T,\mathbf{x}}\big\{\mu(T,\mathbf{x})\big\} \cdot E_{T,\mathbf{x}}\big\{X' \cdot T \cdot \mu(T,\mathbf{x})\big\} = E_{T,\mathbf{x}}\big\{T \cdot \mu(T,\mathbf{x})\big\} \cdot E_{T,\mathbf{x}}\big\{X' \cdot \mu(T,\mathbf{x})\big\}. \quad (3.14)$$

Recall that $\mu(T,\mathbf{x}) = \exp(\alpha) \cdot \exp(T\beta) \cdot \exp(\mathbf{x}'\boldsymbol{\gamma})$, hence the left part of (3.14) is

$$e^{2\alpha} \cdot E\big(e^{T\beta}\big) \cdot E\big(T \cdot e^{T\beta}\big) \cdot E\big(e^{\mathbf{x}'\boldsymbol{\gamma}}\big) \cdot E\big(X' \cdot e^{\mathbf{x}'\boldsymbol{\gamma}}\big),$$

and the right part of (3.14) is

$$e^{2\alpha} \cdot E\big(e^{T\beta}\big) \cdot E\big(T \cdot e^{T\beta}\big) \cdot E\big(e^{\mathbf{x}'\boldsymbol{\gamma}}\big) \cdot E\big(X' \cdot e^{\mathbf{x}'\boldsymbol{\gamma}}\big).$$

This completes the proof. □

Chapter 4

A Unified Sure Independence Ranking and Screening Procedure for Ultrahigh Dimensional Data

4.1 Introduction

Scientific data of unprecedented size and complexity arise in a large variety of fields such as computational biology, climatology, neurology, health science, economics and finance. The total number of features can be much larger than the available sample size. Fan, Samworth and Wu (2009) pointed out that statisticians are confronting simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability in ultrahigh dimensional statistical learning problems. Traditional variable selection procedures such as the AIC (Akaike, 1973), BIC (Schwartz, 1978) and many other existing procedures for high-dimensional data, including the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001; Kim, Choi and Oh, 2008; Zou and Li, 2008), adaptive LASSO (Zou, 2006) and Dantzig selector

(Candes and Tao, 2007; Candes, Wakin and Boyd, 2007) may not perform well due to the ultrahigh dimensionality. Fan and Lv (2008) emphasized the importance of feature screening in the ultrahigh dimensional data analysis, and proposed sure independence screening (SIS) to extract important features in the context of homoscedastic linear models. The SIS ranks the importance of candidate predictors using the Pearson correlation coefficients for univariate and continuous response variables. Fan and Song (2009) proposed to rank the importance of each feature using the magnitude of the marginal likelihood under the framework of generalized linear models. Fan, Feng and Song (2011) generalized the SIS to nonparametric additive models. The aforementioned procedures require data analysts to specify a regression model between the response and the predictors. Their performances rely heavily on how well the postulated working model approximates the underlying true model. We refer to these procedures as model-based screening procedures. The theoretic properties of the model-based procedures were established based on the assumption that the imposed working model is the underlying true model. It is well known that the model assumptions are typically quite restrictive, and it is difficult to test the validity of these model assumptions in an ultrahigh dimensional setting.

In this chapter, we propose a unified sure independence ranking and screening (USIRS, for short) procedure to simultaneously tackle the issues of unprecedented size and complexity. The USIRS selects the important predictors in a model-free fashion in the sense that it does not require the researchers to specify a functional regression relationship between the responses and the predictors. The proposed procedure enables us to identify important features on which the correlations among the response variables depend. The USIRS is readily applicable to the classical linear models, generalized linear models, index models, heteroscedas-

tic models, transformation models, additive models and many others. It can be readily applied to regressions with either univariate or multivariate responses, regardless of whether the response variables are continuous, discrete or categorical. Moreover, the USIRS is computationally efficient and simple to implement, in that it does not require an iterative numerical optimization algorithm for calculating the marginal utility. This is desired in ultrahigh dimensional data analysis. We conduct Monte Carlo simulations to examine the finite sample performance of the USIRS, and compare the performance of the USIRS with the SIS (Fan and Lv, 2008), the nonparametric SIS (Fan, Feng and Song, 2011) and the SIS for generalized linear models (Fan and Song, 2009). The numerical comparisons show that the USIRS is a dramatic improvement upon existing procedures under some statistical settings.

In this chapter we study the theoretical properties of the USIRS. First we prove that the USIRS possesses ranking consistency. That is, the USIRS can rank the important features above the unimportant ones with probability approaching one as the sample size $n \to \infty$. Thus, the USIRS offers a clear separation between the important and unimportant features in an asymptotic sense. We further establish the sure screening property for the USIRS. The sure screening property ensures that the exclusion of important features by the USIRS has a vanishing probability as $n \to \infty$. We show that both the ranking consistency and sure screening properties are valid under very general ultrahigh-dimensional settings.

This chapter is organized as follows. In Section 4.2, we first propose a new marginal utility for feature ranking and screening, then establish both the ranking consistency and sure screening properties for the USIRS. We illustrate the finite sample performance of the USIRS with Monte Carlo simulations in Section 4.3. All technical details are given in Section 4.4.

4.2 A New Screening Procedure

4.2.1 Some Preliminaries

Let $\mathbf{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ and $\mathbf{y} = (Y_1, Y_2, \ldots, Y_q)^{\mathrm{T}}$ be the predictor vector and the response vector, respectively. The response vector in this paper may be univariate (i.e., $q = 1$) and multivariate (i.e., $q > 1$, but is a fixed number). Moreover, the response vector can be discrete, categorical or continuous. It is assumed throughout this paper that $\mathbf{x}$ is marginally standardized; that is, $E(X_j) = 0$ and $\mathrm{var}(X_j) = 1$, for $j = 1, \ldots, p$. The dimension of $\mathbf{x}$ is very large in ultrahigh dimensional data analysis. However, it is commonplace to assume that the response vector is associated with only a small portion of predictors. In other words, the sparsity principle is frequently adopted and deemed useful in the analysis of ultrahigh dimensional data. Let $\Psi(\mathbf{t} \mid \mathbf{x}) = E\left\{\exp(i\mathbf{t}^{\mathrm{T}}\mathbf{y}) \mid \mathbf{x}\right\}$ for $\mathbf{t} \in \mathrm{R}^q$ be the conditional characteristic function of $\mathbf{y}$ given $\mathbf{x}$. To characterize the sparsity, define the active and inactive predictors by

$$\begin{aligned} \mathcal{D} &= \{j : \Psi(\mathbf{t} \mid \mathbf{x}) \text{ functionally depends on } X_j \text{ for some } \mathbf{t} \in \mathrm{R}^q\}, \\ \mathcal{I} &= \{j : \Psi(\mathbf{t} \mid \mathbf{x}) \text{ does not functionally depend on } X_j \text{ for any } \mathbf{t} \in \mathrm{R}^q\}. \end{aligned} \tag{4.1}$$

We further write $\mathbf{x}_{\mathcal{D}} = \{X_j : j \in \mathcal{D}\}$ and $\mathbf{x}_{\mathcal{I}} = \{X_j : j \in \mathcal{I}\}$, and refer to $\mathbf{x}_{\mathcal{D}}$ as an *active* predictor vector and its complement $\mathbf{x}_{\mathcal{I}}$ as an *inactive* predictor vector. Thus, the sparsity principle means that

$$\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{D}}, \tag{4.2}$$

where $\perp\!\!\!\perp$ stands for statistical independence. Model (4.2) indicates that, given the important predictors $\mathbf{x}_{\mathcal{D}}$, the response vector is statistically independent of the whole vector $\mathbf{x}$. The goal of feature screening is to identify the active set $\mathcal{D}$.

4.2.2 A New Marginal Utility

From the definitions of $\mathcal{I}$ and $\mathcal{D}$, $\partial\Psi(\mathbf{t} \mid \mathbf{x})/\partial\mathbf{x}_{\mathcal{I}} = \mathbf{0}$ for any fixed $\mathbf{t} \in \mathrm{R}^q$, and $\partial\Psi(\mathbf{t} \mid \mathbf{x})/\partial\mathbf{x}_{\mathcal{D}} \neq \mathbf{0}$ for some $\mathbf{t} \in \mathrm{R}^q$. Intuitively, the following statistic could serve as a marginal utility for feature ranking and screening:

$$\int_{\mathbf{t}\in\mathrm{R}^q} \left\| E\left\{\frac{\partial\Psi(\mathbf{t} \mid \mathbf{x})}{\partial X_j}\right\}\right\|^2 w(\mathbf{t})d\mathbf{t}, \tag{4.3}$$

where $w(\mathbf{t})$ is a nonnegative weight function. Throughout this article, we let $\|\mathbf{a}\| = (\mathbf{a}^{\mathrm{T}}\overline{\mathbf{a}})^{1/2}$, where $\overline{\mathbf{a}}$ is the complex conjugate of a complex vector $\mathbf{a}$. When $\mathbf{a}$ is a real vector, then $\|\mathbf{a}\|$ reduces to the Euclidean norm of $\mathbf{a}$. We expect the statistic in (4.3) can be used to quantify the importance of X_j.

To put (4.3) into practice, it is necessary to compute $E\{\partial\Psi(\mathbf{t} \mid \mathbf{x})/\partial X_j\}$ and choose an appropriate weight function $w(\mathbf{t})$ to simplify the calculation.

We discuss how to simplify $E\{\partial\Psi(\mathbf{t} \mid \mathbf{x})/\partial X_j\}$ first. We tentatively assume that the predictors are generated from a standard normal population with zero mean and identity covariance matrix. It is worth noting that the normality assumption is not necessary, and will be relaxed later. With integration by parts, we can obtain that

$$E\{\partial\Psi(\mathbf{t} \mid \mathbf{x})/\partial X_j\} = E\{\exp(i\mathbf{t}^{\mathrm{T}}\mathbf{y})X_j\}. \tag{4.4}$$

This implies that $E\{\partial\Psi(\mathbf{t} \mid \mathbf{x})/\partial X_j\}$ is easy to compute.

Next we consider how to choose a proper weight function $w(\mathbf{t})$ to further simplify the calculation of (4.3). Lemma 4.1 below states that we can derive a closed form for $\int_{\mathrm{R}^q} \|E\{\exp(i\mathbf{t}^{\mathrm{T}}\mathbf{y})X_j\}\|^2 w(\mathbf{t})d\mathbf{t}$ for a properly chosen weight function $w(\mathbf{t})$.

Lemma 4.1. *Let* $w(\mathbf{t}) = \Gamma\{(q+1)/2\}\,\|\mathbf{t}\|^{q+1}/\pi^{(q+1)/2}$. *Then*

$$\int_{\mathrm{R}^q} \|E\{\exp(i\mathbf{t}^T\mathbf{y})X_j\}\|^2 w(\mathbf{t})d\mathbf{t} = -E\left(\|\mathbf{y}-\widetilde{\mathbf{y}}\|X_j\widetilde{X}_j\right), \tag{4.5}$$

where $(\widetilde{X}_j, \widetilde{\mathbf{y}})$ *is an independent copy of* $(X_j, \mathbf{y})$, *and the integrals at 0 and* ∞ *are meant in the sense that* $\lim_{\varepsilon\to 0}\int_{\mathrm{R}^d_\epsilon}$, *where* $\mathrm{R}^q_\varepsilon = \mathrm{R}^d/\{\varepsilon B + \varepsilon^{-1}B^c\}$, *and* B *is the unit ball centered at 0 in* R^q.

Define

$$\lambda_j \stackrel{\text{def}}{=} -E\left(\|\mathbf{y}-\widetilde{\mathbf{y}}\|X_j\widetilde{X}_j\right). \tag{4.6}$$

Lemma 4.1, (4.3) and (4.4) imply that λ_j can be used for measuring the importance of X_j. This motivates us to use λ_j as a marginal utility for feature screening.

Although (4.4) is derived under the motivating normality assumption on $\mathbf{x}$, we next show that under some conditions,

$$\max_{j\in\mathcal{I}} \lambda_j < \min_{j\in\mathcal{D}} \lambda_j \tag{4.7}$$

holds uniformly in p without imposing the normality assumption on $\mathbf{x}$.

Denote by $\lambda_{\max}\{\mathbf{B}\}$ and $\lambda_{\min}\{\mathbf{B}\}$ the largest and the smallest eigenvalue of a matrix $\mathbf{B}$, respectively. Define $\mathbf{\Lambda}_{\mathcal{D}} = -E\left(\mathbf{x}_{\mathcal{D}}\widetilde{\mathbf{x}}_{\mathcal{D}}^{\mathrm{T}}\|\mathbf{y}-\widetilde{\mathbf{y}}\|\right)$. We have the following theorem.

Theorem 4.1. *Suppose that*

$$E(\mathbf{x} \mid \mathbf{x}_{\mathcal{D}}) = \mathrm{cov}(\mathbf{x}, \mathbf{x}_{\mathcal{D}}^{T}) \left\{ \mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{D}}^{T}) \right\}^{-1} \mathbf{x}_{\mathcal{D}}, \tag{4.8}$$

and

$$\limsup_{p \to \infty} \left\{ \frac{\lambda_{\max} \left\{ \mathrm{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{D}}^{T}) \mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{I}}) \right\}}{\lambda_{\min}^{2} \left\{ \mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{D}}^{T}) \right\}} - \frac{\min_{j \in \mathcal{D}} \{\lambda_j\}}{\lambda_{\max}\{\mathbf{\Lambda}_{\mathcal{D}}\}} \right\} < 0. \tag{4.9}$$

Then (4.7) holds uniformly in p.

The proof of Theorem 4.1 is given in Section 4.4. We remark here that λ_j for $j \in \mathcal{I}$ can be nonzero because we allow for correlations between $\mathbf{x}_{\mathcal{D}}$ and $\mathbf{x}_{\mathcal{I}}$. Actually, $\lambda_j = 0$ for $j \in \mathcal{I}$ if and only if $\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{I}}^{\mathrm{T}}) = \mathbf{0}$. This can be derived from (4.17) in the proof of Theorem 4.1. When $\mathbf{x}_{\mathcal{D}}$ and $\mathbf{x}_{\mathcal{I}}$ are uncorrelated, (4.7) and (4.9) follow logically.

Theorem 4.1 ensures that without the normality assumption on $\mathbf{x}$, the marginal utilities of $\mathbf{x}_{\mathcal{D}}$ are always larger than those of $\mathbf{x}_{\mathcal{I}}$ under conditions (4.8) and (4.9), which are mild. Condition (4.8) holds if $\mathbf{x}$ follows a normal or an elliptically symmetric distribution. Condition (4.9) is parallel to conditions (3)-(4) in Fan and Lv (2008). This is a key assumption to ensure the USIRS will work properly. The quantity $\min_{j \in \mathcal{D}} \lambda_j$ on the right hand side of (4.9) reflects the signal strength of an individual important predictor. This is similar to condition (3) of Fan and Lv (2008, page 870), which requires the contribution of an important predictor to be sufficiently large.

4.2.3 An Estimator and Its Sampling Properties

Suppose that $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$ is a random sample from $(\mathbf{x}, \mathbf{y})$. It is assumed henceforth that the sample predictors are marginally standardized. That is, $n^{-1}\sum_{i=1}^{n} X_{ij} = 0$ and $n^{-1}\sum_{i=1}^{n} X_{ij}^2 = 1$, for $j = 1, \ldots, p$. A natural estimator of λ_j defined in (4.6) is the following U-statistic

$$\hat{\lambda}_j = -\frac{1}{n(n-1)}\sum_{i \neq k} \|\mathbf{y}_i - \mathbf{y}_k\| X_{ij} X_{kj}. \tag{4.10}$$

We next establish the ranking consistency property for the $\hat{\lambda}_j$ under the following sub-exponential tail probability condition:

(C1) Both $\mathbf{x}$ and $\mathbf{y}$ satisfy sub-exponential tail probabilities. That is, there exists a positive constant s_0 such that

$$\max_{1 \leq j \leq p} E\left\{\exp\left(sX_j^2\right)\right\} < \infty, \text{ and } E\left\{\exp\left(s\|\mathbf{y}\|^2\right)\right\} < \infty, \text{ for all } 0 < s \leq s_0.$$

Theorem 4.2. (Ranking Consistency Property) *Suppose that Condition (C1) holds. Under the conditions in Theorem 4.1, it follows that for any $\varepsilon > 0$, there exists a sufficiently small constant $s_\varepsilon > 0$ such that*

$$Pr\left(\sup_{j=1,\ldots,p} |\widehat{\lambda}_j - \lambda_j| > \varepsilon\right) \leq 2p \exp\left\{n \log(1 - \varepsilon s_\varepsilon/2)/2\right\}. \tag{4.11}$$

In addition, if we write $\delta = \min_{j \in \mathcal{D}} \lambda_j - \max_{j \in \mathcal{I}} \lambda_j$, then there exists a sufficiently small constant $s_\delta > 0$ such that

$$Pr\left(\max_{j \in \mathcal{I}} \widehat{\lambda}_j < \min_{j \in \mathcal{D}} \widehat{\lambda}_j\right) \leq 1 - 4p \exp\left\{n \log(1 - \delta s_\delta/4)/2\right\}. \tag{4.12}$$

Theorem 4.2 ensures that the sample marginal utilities of important predictors will be ranked at the top with probability tending to 1 as $n \to \infty$ when $p = O\{\exp(\xi n)\}$ for some $\xi > 0$. Thus, to identify the important predictors, we can choose the predictors associated with the largest N marginal utilities $\widehat{\lambda}_j$s, where N is a user-specified number. Fan and Lv (2008) suggested choosing $N = O\{n/\log(n)\}$. In summary,

$$\widehat{\mathcal{D}}^{\star} = \left\{ j : \widehat{\lambda}_j > \widehat{\lambda}_{(N)} \right\}$$

is a natural estimator of $\mathcal{D}$, where $\widehat{\lambda}_{(N)}$ is the N-th largest value among all $\widehat{\lambda}_j$s.

To establish the sure screening property of $\hat{\lambda}_j$, we need the following conditions:

(C2) Both $\mathbf{x}$ and $\mathbf{y}$ are uniformly bounded. That is, there exist positive constants a and b such that $\max_{1 \le j \le p} |X_j| \le a$ and $\max_{1 \le k \le q} |Y_k| \le b$.

(C3) The minimum marginal utility of active predictors satisfies

$$\min_{j \in \mathcal{D}} \lambda_j \ge 2cn^{-\kappa}, \text{ for some constants } c \text{ and } 0 \le \kappa < 1/2.$$

Condition (C2) is more stringent than the sub-exponential tail probability (C1). It is a technical condition to facilitate the proof, but it may not be the weakest one. It has been observed from our numerical studies that, even when this condition is violated, the newly proposed approach still performs quite well. In addition, we can consider transforming the predictors to ensure that they are bounded, because any monotonic transformation will not change the role of important predictors defined in (4.2). When (C2) holds, we can expect our screening procedure to be valid for an even larger p value than when (C1) holds true. Condition (C3) is milder than (4.7) in that we allow the quantity $\min_{j \in \mathcal{D}} \lambda_j$ to shrink to zero as $n \to \infty$. Again, we select a set of important predictors with large magnitudes. That is, we

define

$$\widehat{\mathcal{D}} = \left\{ j : \widehat{\lambda}_j \geq cn^{-\kappa} \text{ for } 1 \leq j \leq p \right\},$$

where c and κ are pre-specified threshold values which are defined in (C3).

The following theorem establishes the sure screening property for our proposed independence ranking and screening procedure. This is a desired property for ultrahigh dimensional statistical learning problems (Fan and Lv, 2008). In this theorem, we explicitly quantify the extent to which the predictor dimension can be reduced by the independence screening, which depends on the number of truly important parameters.

Theorem 4.3. (Sure Screening Property) *If Condition (C2) holds, then*

$$Pr\left(\max_{1 \leq j \leq p} |\widehat{\lambda}_j - \lambda_j| > cn^{-\kappa} \right) \leq O\left(p \exp\left\{ -n^{1-2\kappa}/(16a^4b^2) \right\} \right). \tag{4.13}$$

If, in addition, Condition (C3) holds, then

$$Pr\left(\mathcal{D} \subseteq \widehat{\mathcal{D}} \right) \geq 1 - O\left(s_n \exp\left\{ -n^{1-2\kappa}/(16a^4b^2) \right\} \right), \tag{4.14}$$

where s_n is the cardinality of $\mathcal{D}$.

4.3 Numerical Studies

4.3.1 Simulations

In this section, we conduct Monte Carlo simulations to assess the performance of the proposed procedure and compare its performance with existing independence screening procedures. All simulation studies were conducted with Matlab. In our

simulation, we set the dimension p of $\mathbf{x}$ to be 2000, and generate the predictor vector $\mathbf{x}$ from a p-dimensional normal distribution with mean zero and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = 0.8^{|i-j|}$. We conduct 1000 replications for each Monte Carlo experiment.

We adopt the following two criteria to assess the performance.

1. The minimal model size (denoted by $\mathcal{S}$) which is required to include all truly important predictors. We report the quintuplet consisting of the minimum, the first quartile, the median, the third quartile and the maximum number of $\mathcal{S}$ out of 1000 replications. This criteria is to assess the ranking consistency property. It is expected that $\mathcal{S}$ is close to the number of truly important predictors when the ranking consistency property holds.

2. The frequency (denoted by $\mathcal{F}$) that a truly important predictor is correctly identified in 1000 replications. Following Fan and Lv (2008), we retain the predictors associated with the $[n/\log n]$ largest $\hat{\lambda}_j$s in our simulations, where $[a]$ stands for the integer part of a. This criterion measures the sure screening property. It is expected that the $\mathcal{F}$ values for all truly important predictors are close to one when the sure screening property is valid.

Example 4.1. This example is designed to compare the performance of the proposed USIRS for linear regression models with the sure independent screening (SIS) procedure proposed in Fan and Lv (2008) and the nonparametric independent screening (NIS) procedure proposed in Fan, Feng and Song (2011). We generate 1000 data sets each consisting of $n = 200$ random samples from the following linear regression model:

$$Y = X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5 + \sigma_0\sigma(\mathbf{x})\epsilon, \tag{4.15}$$

where $\epsilon \sim N(0,1)$. We consider the following two scenarios for $\sigma(\mathbf{x})$:

S1: Let $\sigma^2(\mathbf{x}) = 6.8285$. In this case, (4.15) is homoscedastic.

S2: Let $\sigma^2(\mathbf{x}) = \exp(1.0727X_{20})$. Consequently, (4.15) is heteroscedastic.

Thus, $X_1, \cdots, X_5$ are important predictors in S1, while $X_1, \cdots, X_5$ and X_{20} are important predictors in S2. Therefore, the number of truly important predictors is 5 in S1 and 6 in S2. In this example, we set $\sigma_0 = 0.5, 1$ and 2 so that the corresponding multiple correlation coefficient (R^2) approximately equals 80%, 50% and 20%, respectively, for both scenarios.

Table 4.1. The minimum model size required to ensure the inclusion of all truly active predictors. The quintuplet in each parenthesis consists of the minimum, the first quartile, the median, the third quartile and the maximum value of $\mathcal{S}$ of 1000 replications.

σ_0		Scenario S1					Scenario S2				
0.5	USIRS	(5	5	5	5	6)	(7	80	218	553	1960)
	NIS	(5	5	5	5	6)	(6	25.5	245	859	1998)
	SIS	(5	5	5	5	6)	(7	236	828.5	1544.5	2000)
1	USIRS	(5	5	5	5	6)	(6	10	16	43	1458)
	NIS	(5	5	5	5	8)	(6	22	83	603	1994)
	SIS	(5	5	5	5	6)	(6	159.5	888	1703.5	2000)
2	USIRS	(5	5	5	6	1235)	(6	6	7	8	101)
	NIS	(5	5	6	13	1272)	(6	51	152	578	1997)
	SIS	(5	5	5	6	401)	(6	159	976.5	1814.5	2000)

Tables 4.1 and 4.2 depict the quintuplet of the minimal model size $\mathcal{S}$ and the frequency $\mathcal{F}$, respectively. As expected, all three methods perform quite well for the linear model with homogenous error. For scenario S1, the left panel of Table 4.1 indicates that all three methods deteriorate when the R^2 value becomes smaller. However, the right panel of Table 4.1 reveals a reversed phenomenon for the USIRS and NIS in scenario S2. That is, when R^2 is large (i.e., the noise level is comparatively small), it is difficult for both the USIRS and NSIS to identify X_{20};

Table 4.2. The frequency $\mathcal{F}$ of each active predictor out of 1000 replications.

σ_0		Scenario S1					Scenario S2					
		X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_{20}
0.5	USIRS	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	140
	NIS	1000	1000	1000	1000	1000	1000	999	998	997	993	300
	SIS	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	87
1	USIRS	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	731
	NIS	1000	1000	1000	1000	1000	982	988	984	974	937	417
	SIS	1000	1000	1000	1000	1000	1000	1000	1000	999	997	129
2	USIRS	999	1000	999	997	963	1000	1000	1000	1000	1000	998
	NIS	998	1000	997	979	880	794	835	811	721	547	482
	SIS	999	1000	1000	998	983	978	987	988	967	921	160

when R^2 is small (e.g. $R^2 = 20\%$), the random error contributes more variation to the response variable. Therefore, the USIRS can identify X_{20} successfully. In this case, our procedure works quite well, as the maximum value of $\mathcal{S}$ is 101. A close inspection of Table 4.2 reveals that the minimal model size $\mathcal{S}$ was greater than $[n/\log(n)] = 37$ only twice in the 1000 replications. Since both the SIS and NIS are proposed for screening features contained in regression functions, they are expected to perform poorly in identifying the active variable X_{20} contained in the variance function. This is consistent with the simulation results in the right panel of Table 4.2. In sum, the USIRS performs as well as the SIS and NIS for linear models with homogeneous errors. The USIRS outperforms both the SIS and NIS for linear models with heteroscedastic errors, in that the SIS fails to identify X_{20} as an active variable, while the performance of the NIS in terms of identifying active predictors in the regression function deteriorate quickly as the noise level increases.

Example 4.2. In this example, we assess the performance of the proposed procedure for ultrahigh dimensional nonlinear regression models. Simulation data sets,

each of size $n = 200$, were generated from the following four nonlinear models:

$$\text{(M1)} \qquad Y = \exp(X_1) + \sin(X_2 + X_3) + \exp(2^{1/2} X_{20})\epsilon;$$

$$\text{(M2)} \qquad Y = (X_1 + X_2 + X_3) \,/\, \left\{0.5 + (X_{20} - 1)^2\right\} + 0.2\epsilon;$$

$$\text{(M3)} \qquad Y = (X_1 + X_2 + X_3)\,(X_{20} + 1)^2 + \epsilon;$$

$$\text{(M4)} \qquad Y = \text{sign}\,(X_1 + X_2 + X_3) \log\left(|X_{20} + 2|\right) + 0.2\epsilon.$$

The number of active predictors is 4 for these nonlinear models. Model (M1) is heteroscedastic, and models (M2)-(M4) are homoscedastic and yet have nonlinear regression functions. To improve the SIS, Fan, Feng and Song (2011) proposed the NIS for an additive model with homogeneous errors.

Table 4.3. Simulation results for Example 4.2.

Model	Method	Minimum Model Size $\mathcal{S}$					Frequency $\mathcal{F}$			
							X_1	X_2	X_3	X_{20}
M1	USIRS	(4	4	5	5	53)	1000	1000	1000	999
	NIS	(4	63.5	206	650.5	1985)	705	556	397	553
	SIS	(4	172.5	983.5	1920	2000)	907	852	776	184
M2	USIRS	(5	6	7	8	12)	1000	1000	1000	1000
	NIS	(15	906.5	1494	1853	2000)	1000	1000	1000	6
	SIS	(16	420	872.5	1351	2000)	1000	1000	1000	18
M3	USIRS	(4	4	5	5	8)	1000	1000	1000	1000
	NIS	(4	15	77	444	2000)	1000	1000	1000	410
	SIS	(5	137	725	1620	2000)	1000	1000	1000	148
M4	USIRS	(4	5	6	7	53)	1000	1000	1000	999
	NIS	(4	97	341.5	928.5	1998)	998	999	991	158
	SIS	(5	338.5	1008	1639.5	2000)	1000	1000	1000	54

The minimal model size $\mathcal{S}$ and the frequency $\mathcal{F}$ out of 1000 simulations are summarized in the left and the right panel of Table 4.3, respectively. For models M1-M4, the USIRS certainly outperforms both the NIS and SIS. The minimum model size for the USIRS is close to 4 and the frequency $\mathcal{F}$ is close to the number

of simulations. This is consistent with our theoretic analysis because the USIRS is a unified independence ranking and screening procedure, and does not require model specification for the regression function. It is expected that for both the NIS and SIS it is difficult to identify the active variable X_{20} in model M1 because X_{20} is contained in the conditional variance function. The NIS and SIS perform poorly in identifying the active predictor X_{20} in models M2-M4, partially because the signal strength of X_{20} is very weak.

Table 4.4. Simulation results for Example 4.3.

Model	n	Method	Minimum Model Size $\mathcal{S}$	Frequency $\mathcal{F}$			
				X_1	X_2	X_3	X_{20}
Poisson	200	USIRS	(4 5 6 7 258)	1000	1000	1000	990
		GSIS	(4 6 8 16 1895)	994	998	1000	861
	400	USIRS	(4 5 6 6 26)	1000	1000	1000	1000
		GSIS	(4 5 6 8 430)	1000	1000	1000	980
	800	USIRS	(4 6 6 6 8)	1000	1000	1000	1000
		GSIS	(4 5 6 7 272)	1000	1000	1000	994
Logistic	200	USIRS	(4 5 6 8 998)	1000	1000	1000	967
		GSIS	(4 5 6 8 997)	1000	1000	1000	967
	400	USIRS	(4 5 6 7 24)	1000	1000	1000	1000
		GSIS	(4 5 6 7 24)	1000	1000	1000	1000
	800	USIRS	(4 5 6 6 9)	1000	1000	1000	1000
		GSIS	(4 5 6 6 9)	1000	1000	1000	1000

Example 4.3. In this example, we examine the performance of the USIRS for the generalized linear models. Fan and Song (2009) proposed a sure independence screening procedure for generalized linear models based on the likelihood function. Their procedure is referred to as GSIS in this example. It is interesting to compare the performance of the USIRS and GSIS under the framework of generalized linear models. To this end, we generate data from a Poisson log-linear regression model:

$$Y \sim \text{Poisson}\left\{\lambda(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\}$$

with $\log\{\lambda(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} = 1+0.5X_1+0.5X_2+X_3+X_{20}$, and a logistic regression model:

$$Y \sim \text{Bernoulli}\{\mu(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\},$$

where $\text{logit}\{\mu(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} = 1 + 0.5X_1 + 0.5X_2 + X_3 + X_{20}$. The number of active predictors is 4 in both models.

The minimal model size $\mathcal{S}$ and the frequency $\mathcal{F}$ are reported in the left panel and the right panel of Tables 4.4, respectively. Both the GSIS and the USIRS perform very well in these two models. The performance of the USIRS and the GSIS is almost the same for the logistic regression, and the USIRS performs slightly better than the GSIS in terms of the maximum value of the minimum model size $\mathcal{S}$.

Example 4.4. In this example, we investigate the performance of the proposed USIRS for multiple responses. We generated a bivariate response $\mathbf{y} = (Y_1, Y_2)^{\mathrm{T}}$ from a normal population with mean zero and covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \begin{pmatrix} 1 & \sin(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \\ \sin(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) & 1 \end{pmatrix},$$

where $\boldsymbol{\beta} = (0.8, 0.6, 0, \cdots, 0, 0)^{\mathrm{T}}$. In this example, $\text{cov}(\mathbf{x}, \mathbf{y}) = E(\mathbf{x}\mathbf{y}^{\mathrm{T}}) = \mathbf{0}$, which makes the identification of important predictors very challenging. We set the sample size $n = 200$, 400 and 800. The results are summarized in Table 4.5, from which it can be seen that the USIRS performs reasonably well even when $n = 200$, and performs very well when $n = 400$. When $n = 800$, the minimal model size is close to the number of active predictors with an overwhelming probability.

In the following example, we consider regressions with multivariate responses.

Table 4.5. Simulation results for Example 4.4 with bivariate response.

n	Minimum Model Size $\mathcal{S}$					Frequency $\mathcal{F}$	
						X_1	X_2
200	(2	7	16	39	377)	841	791
400	(2	2	2	3	43)	1000	1000
800	(2	2	2	2	3)	1000	1000

We generate the response vector $\mathbf{y} = (Y_1, \ldots, Y_5)^{\mathrm{T}}$ from the following model:

$$\begin{aligned} Y_1 &= 1 + X_1 + \sin(X_2 + X_3) + \epsilon_1; \\ Y_2 &= (X_2 + X_3) \,/\, \left\{0.5 + (X_1 + 1)^2\right\} + \epsilon_2; \\ Y_3 &= |\exp(X_{20} + 1)|\epsilon_3; \\ Y_4 &= \epsilon_4; \\ Y_5 &= \epsilon_5, \end{aligned}$$

where ϵ_is are generated from a normal population with mean zero and covariance matrix

$$\mathbf{\Sigma}_\epsilon = \begin{pmatrix} 1 & -1/2 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We started with $n = 200$. The simulation results with $n = 200$ shown in Table 4.6 are very encouraging. The minimum $\mathcal{S}$ value out of 1000 repetitions is 4 and the maximum $\mathcal{S}$ value is 7. In addition, the frequency $\mathcal{F}$ for the truly important predictors X_1, X_2, X_3 and X_{20} are all 1000 out of 1000 simulations. Then we

reduce the sample size to $n = 100$ and $n = 50$. Results for $n = 100$ are still very good, and results for $n = 50$ are reasonable.

Table 4.6. Simulation results for Example 4.4 with a six-dimensional response vector.

n	Minimum Model Size $\mathcal{S}$					Frequency $\mathcal{F}$			
						X_1	X_2	X_3	X_{20}
50	(4	10	21	48	1081)	689	799	735	702
100	(4	4	5	6	73)	993	999	997	1000
200	(4	4	4	5	7)	1000	1000	1000	1000

4.3.2 An Application

In this section we illustrate the USIRS by an empirical analysis of the data set reported in Scheetz (2006). In this data set, 120 rats were selected for tissue harvesting, microarray analysis and genotyping. Microarrays from the eyes of these rats were collected to analyze the RNA. These microarrays contain more than 30,000 different probes, among which 18,976 probes were found at a level sufficient to be considered "expressed" with sufficient variation. We study how the gene expression levels at the probes 1389163_at (Y_1, a gene identified to be involved in human hereditary diseases of the retina) and 1382223_at (Y_2, the probe which is the most correlated with Y_1 among all probes) depend on expression levels at other probes ($\mathbf{x}$). The dimension $p = 18,974$ is fairly large compared with the sample size $n = 120$. To analyze this data set, feature screening seems a necessary initial step as a prelude to any other sophisticated statistical modelings that do not cope well with such a high dimensionality.

We demonstrate the performance of USIRS with different response variables $\mathbf{y}_i$s, denoted with USIRS($\mathbf{y}_i$)s, where $\mathbf{y}_1 = (Y_1, Y_2)^{\mathrm{T}}$, $\mathbf{y}_2 = Y_1$, $\mathbf{y}_3 = Y_2$ and $\mathbf{y}_4 = Y_1Y_2$. USIRS($\mathbf{y}_1$) stands for the USIRS with a bivariate response, while USIRS($\mathbf{y}_2$),

USIRS($\mathbf{y}_3$) and USIRS($\mathbf{y}_4$) for the USIRS with a univariate response variable. In particular, USIRS($\mathbf{y}_4$) is designed to detect important predictors associated with their correlations between Y_1 and Y_2. For each of these four analyses, we apply USIRS($\mathbf{y}_i$) to the data set $(\mathbf{x}, \mathbf{y}_i)$, and select $[n/\log n] = 25$ probes which are associated with the largest marginal utilities λ_js. We compare both similarities and differences of these probes selected by USIRS($\mathbf{y}_i$)s. The results are summarized in Table 4.7, from which it can be easily seen that the USIRS($\mathbf{y}_1$) selects 25 probes, 11 of which are also selected by USIRS($\mathbf{y}_2$), 12 of which are selected by USIRS($\mathbf{y}_3$), and 24 of which are selected by USIRS($\mathbf{y}_4$). The analysis of USIRS($\mathbf{y}_4$) verifies the existence of associations between Y_1 and Y_2. This is expected because these two probes are most correlated among all candidate probes. This observation implies that the USIRS with a bivariate response $\mathbf{y}_1$ identifies some relevant probes which are missed by the USIRS with univariate responses $\mathbf{y}_2$ and $\mathbf{y}_3$.

We further demonstrate the performance through regression modelings. We fit four linear regression models. The first model uses 25 probes selected by USIRS($\mathbf{y}_2$) as predictors; the second uses the 11 common probes selected by both USIRS($\mathbf{y}_1$) and USIRS($\mathbf{y}_2$); the third uses 25 probes selected by USIRS($\mathbf{y}_3$) and the fourth uses the 12 common probes selected by USIRS($\mathbf{y}_1$) and USIRS($\mathbf{y}_3$). We randomly divide the original data set into two sets, a training set consisting of 100 observations and a test set consisting of 20 observations. We fit four linear regressions with different predictors using the training data set and calculate the prediction error using the test data set. We repeat this experiment 100 times. The average of the squared prediction error of these four models are, respectively, 1.0827, 1.0833, 1.0996 and 0.8878. These numbers are very close to each other, indicating that these 11 (or 12) common probes are sufficient to describe the variation in the response variable $\mathbf{y}_2$ (or $\mathbf{y}_3$).

Table 4.7. Probes selected by USIRS with different response variables. Common probes are marked with $\surd$.

	common probes			
Probes	USIRS($\mathbf{y}_1$)	USIRS($\mathbf{y}_2$)	USIRS($\mathbf{y}_3$)	USIRS($\mathbf{y}_4$)
1373165_at	$\surd$			$\surd$
1373534_at	$\surd$	$\surd$		$\surd$
1373944_at	$\surd$		$\surd$	$\surd$
1374647_at	$\surd$		$\surd$	$\surd$
1376374_at	$\surd$		$\surd$	$\surd$
1376747_at	$\surd$	$\surd$		$\surd$
1376773_at	$\surd$		$\surd$	$\surd$
1377005_at	$\surd$		$\surd$	$\surd$
1379580_at	$\surd$			$\surd$
1380466_at	$\surd$	$\surd$	$\surd$	$\surd$
1388656_at	$\surd$	$\surd$		$\surd$
1389539_at	$\surd$			$\surd$
1390272_at	$\surd$	$\surd$		$\surd$
1390323_at	$\surd$	$\surd$		$\surd$
1390539_at	$\surd$	$\surd$		$\surd$
1398340_at	$\surd$		$\surd$	$\surd$
1399134_at	$\surd$	$\surd$		$\surd$
1377745_at	$\surd$		$\surd$	$\surd$
1377791_at	$\surd$	$\surd$	$\surd$	$\surd$
1378590_at	$\surd$	$\surd$		$\surd$
1379728_at	$\surd$		$\surd$	$\surd$
1382154_at	$\surd$		$\surd$	$\surd$
1383783_at	$\surd$	$\surd$		
1391183_at	$\surd$			$\surd$
1392470_at	$\surd$		$\surd$	$\surd$
Num.	25	11	12	24

Both USIRS($\mathbf{y}_1$) and USIRS($\mathbf{y}_4$) select 24 probes, which contain 14 probes missed by USIRS($\mathbf{y}_2$) and 13 probes missed by USIRS($\mathbf{y}_3$). This phenomenon verifies that USIRS($\mathbf{y}_1$) automatically takes into account the covariance structure between the response variables. This observation demonstrates that an intuitive combination of USIRS($\mathbf{y}_2$) and USIRS($\mathbf{y}_3$) cannot identify the predictors which

are only associated with the correlations between the response variables, while the joint analysis of USIRS($\mathbf{y}_1$) does not suffer from this problem. This echoes the phenomenon we observed in Example 4.4.

4.4 Theoretical Proofs

4.4.1 Proof of Lemma4.1

For notational clarity, we denote by $\mathrm{supp}(X_j, \mathbf{y})$ the support of $(X_j, \mathbf{y})$, write $\Lambda_j(\mathbf{t}) = E\left\{\exp(i\mathbf{t}^{\mathrm{T}}\mathbf{y})X_j\right\}$ and $\lambda_j = \int_{\mathrm{R}^q} \|\Lambda_j(\mathbf{t})\|^2 w(\mathbf{t})d\mathbf{t}$. Recall that $\|\Lambda_j(\mathbf{t})\|^2 = \Lambda_j(\mathbf{t})\overline{\Lambda}_j(\mathbf{t})$, where $\overline{\Lambda}_j(\mathbf{t}) = E\{\exp(-i\mathbf{t}^{\mathrm{T}}\widetilde{\mathbf{y}})\widetilde{X}_j\}$ is the complex conjugate of $\Lambda_j(\mathbf{t}) = E\left\{\exp(i\mathbf{t}^{\mathrm{T}}\mathbf{y})X_j\right\}$. These facts, together with the Fubini theorem, imply that

$$
\begin{aligned}
& \int_{\mathrm{R}^q} \|\Lambda_j(\mathbf{t})\|^2 w(\mathbf{t})d\mathbf{t} = \int_{\mathrm{R}^q} \Lambda_j(\mathbf{t})\overline{\Lambda}_j(\mathbf{t})w(\mathbf{t})d\mathbf{t} \\
= & \int_{\mathrm{supp}(X_j,\mathbf{y})} \int_{\mathrm{supp}(\widetilde{X}_j,\widetilde{\mathbf{y}})} \int_{\mathrm{R}^q} \exp\left\{i\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(\mathbf{t})d\mathbf{t}X_j\widetilde{X}_j dF(X_j,\mathbf{y})dF(\widetilde{X}_j,\widetilde{\mathbf{y}}) \\
= & \int_{\mathrm{supp}(X_j,\mathbf{y})} \int_{\mathrm{supp}(\widetilde{X}_j,\widetilde{\mathbf{y}})} \int_{\mathrm{R}^q} \cos\left\{\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(\mathbf{t})d\mathbf{t}X_j\widetilde{X}_j dF(X_j,\mathbf{y})dF(\widetilde{X}_j,\widetilde{\mathbf{y}}) \\
+ & \int_{\mathrm{supp}(X_j,\mathbf{y})} \int_{\mathrm{supp}(\widetilde{X}_j,\widetilde{\mathbf{y}})} \int_{\mathrm{R}^q} i\sin\left\{\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(\mathbf{t})d\mathbf{t}X_j\widetilde{X}_j dF(X_j,\mathbf{y})dF(\widetilde{X}_j,\widetilde{\mathbf{y}}).
\end{aligned}
$$

Apply Lemma 1 of Szekely, Rizzo and Bakirov (2007) with $w(\mathbf{t}) = c_q\|\mathbf{t}\|^{q+1}$ and $c_q = \Gamma\left\{(q+1)/2\right\}/\pi^{(q+1)/2}$. It follows that

$$
\int_{\mathrm{R}^q} \left[1-\cos\left\{\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\}\right] w(\mathbf{t})d\mathbf{t} = \|\mathbf{y}-\widetilde{\mathbf{y}}\|,
$$

which entails that

$$
\int_{\mathrm{supp}(X_j,\mathbf{y})} \int_{\mathrm{supp}(\widetilde{X}_j,\widetilde{\mathbf{y}})} \int_{\mathrm{R}^q} \cos\left\{\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(\mathbf{t})d\mathbf{t}X_j\widetilde{X}_j dF(X_j,\mathbf{y})dF(\widetilde{X}_j,\widetilde{\mathbf{y}})
$$

$$
\begin{aligned}
&= \int_{\mathrm{supp}(X_j,\mathbf{y})} \int_{\mathrm{supp}(\widetilde{X}_j,\widetilde{\mathbf{y}})} \int_{\mathrm{R}^q} \left[\cos\left\{\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} - 1\right] w(\mathbf{t}) d\mathbf{t} X_j \widetilde{X}_j dF(X_j,\mathbf{y}) dF(\widetilde{X}_j,\widetilde{\mathbf{y}}) \\
&+ \int_{\mathrm{supp}(X_j,\mathbf{y})} X_j dF(X_j,\mathbf{y}) \int_{\mathrm{supp}(\widetilde{X}_j,\widetilde{\mathbf{y}})} \widetilde{X}_j dF(\widetilde{X}_j,\widetilde{\mathbf{y}}) \int_{\mathrm{R}^q} w(\mathbf{t}) d\mathbf{t} \\
&= -E\left(\|\mathbf{y}-\widetilde{\mathbf{y}}\| X_j \widetilde{X}_j\right).
\end{aligned}
$$

The last equality follows because (i) $E(X_j) = E(\widetilde{X}_j) = 0$, and (ii) the integrals at 0 and ∞ are meant in the sense that $\lim_{\varepsilon \to 0} \int_{\mathrm{R}^d_\epsilon}$, where $\mathrm{R}^q_\varepsilon = \mathrm{R}^d / \{\varepsilon B + \varepsilon^{-1} B^c\}$, and B is the unit ball centered at 0 in R^q.

For any given $\mathbf{y} - \widetilde{\mathbf{y}} \neq 0$, $\sin\left\{\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} w(\mathbf{t})$ is an odd function, we can easily obtain that

$$
\int_{\mathrm{supp}(X_j,\mathbf{y})} \int_{\mathrm{supp}(\widetilde{X}_j,\widetilde{\mathbf{y}})} \int_{\mathrm{R}^q} \sin\left\{\mathbf{t}^{\mathrm{T}}(\mathbf{y}-\widetilde{\mathbf{y}})\right\} \lambda(\mathbf{t}) d\mathbf{t} \left(X_j \widetilde{X}_j\right) dF(X_j,\mathbf{y}) dF(\widetilde{X}_j,\widetilde{\mathbf{y}}) = 0.
$$

The proof is completed by combining these two results. □

Proof of Theorem 4.1

Consider the left hand side of (4.7). The linearity condition (4.8), together with the definition of $\mathbf{x}_{\mathcal{D}}$ in model (4.2), entails that

$$
\begin{aligned}
E(X_j \mid \mathbf{y}) &= E\left\{E(X_j \mid \mathbf{x}_{\mathcal{D}}, \mathbf{y}) \mid \mathbf{y}\right\} = E\left\{E(X_j \mid \mathbf{x}_{\mathcal{D}}) \mid \mathbf{y}\right\} \\
&= \mathrm{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}}) \left\{\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\right\}^{-1} E\left(\mathbf{x}_{\mathcal{D}} \mid \mathbf{y}\right). \qquad (4.16)
\end{aligned}
$$

For notational clarity, we write $\mathbf{\Lambda}_{\mathcal{D}} = E\left(-\mathbf{x}_{\mathcal{D}} \widetilde{\mathbf{x}}_{\mathcal{D}}^{\mathrm{T}} \|\mathbf{y} - \widetilde{\mathbf{y}}\|\right)$ and $\mathbf{\Sigma}_{\mathcal{D}} = \mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})$. Since $(X_j, \mathbf{y}) \perp\!\!\!\perp (\widetilde{X}_j, \widetilde{\mathbf{y}})$, it follows that $X_j \perp\!\!\!\perp \widetilde{X}_j \mid (\mathbf{y}, \widetilde{\mathbf{y}})$, $\widetilde{X}_j \perp\!\!\!\perp \mathbf{y} \mid \widetilde{\mathbf{y}}$ and $X_j \perp\!\!\!\perp \widetilde{\mathbf{y}} \mid \mathbf{y}$. These three results, together with (4.16), yield that

$$
\begin{aligned}
\lambda_j &= E\left\{-E(X_j \widetilde{X}_j \mid \mathbf{y}, \widetilde{\mathbf{y}}) \|\mathbf{y} - \widetilde{\mathbf{y}}\|\right\} = E\left\{-E(X_j \mid \mathbf{y}) E(\widetilde{X}_j \mid \widetilde{\mathbf{y}}) \|\mathbf{y} - \widetilde{\mathbf{y}}\|\right\} \\
&= \mathrm{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}}) \mathbf{\Sigma}_{\mathcal{D}}^{-1} \mathbf{\Lambda}_{\mathcal{D}} \mathbf{\Sigma}_{\mathcal{D}}^{-1} \mathrm{cov}(\mathbf{x}_{\mathcal{D}}, X_j). \qquad (4.17)
\end{aligned}
$$

Therefore, it follows that

$$\max_{j\in\mathcal{I}} \lambda_j \leq \lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\boldsymbol{\Lambda}_{\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\right) \max_{j\in\mathcal{I}} \left\{\mathrm{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, X_j)\right\}.$$

Use the fact that $\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{B}\mathbf{C}) \leq \lambda_{\max}(\mathbf{B})\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{C})$ for any matrix $\mathbf{B} \geq 0$. We obtain

$$\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\boldsymbol{\Lambda}_{\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\right) \leq \lambda_{\max}\left(\boldsymbol{\Lambda}\right)/\lambda_{\min}^2\left(\boldsymbol{\Sigma}_{\mathcal{D}}\right).$$

In addition,

$$\max_{j\in\mathcal{I}} \left\{\mathrm{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, X_j)\right\} \leq \lambda_{\max}\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{I}})\right\}.$$

These two results imply that

$$\max_{j\in\mathcal{I}} \lambda_j \leq \lambda_{\max}\left(\boldsymbol{\Lambda}_{\mathcal{D}}\right)\lambda_{\max}\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{I}})\right\}/\lambda_{\min}^2\left(\boldsymbol{\Sigma}_{\mathcal{D}}\right),$$

which completes the proof of (4.7). □

4.4.2 Proof of Theorem 4.2

The basic idea of this proof is to use the U-statistic theory repetitively. See for example Serfling (1980, Section 5).

For notational clarity, we write $h(X_j, \mathbf{y}; \widetilde{X}_j, \widetilde{\mathbf{y}}) = -X_j\widetilde{X}_j\|\mathbf{y} - \widetilde{\mathbf{y}}\|$. Thus,

$$\lambda_j = E\left\{h(X_j, \mathbf{y}; \widetilde{X}_j, \widetilde{\mathbf{y}})\right\}, \text{ and } \widehat{\lambda}_j = \{n(n-1)\}^{-1}\sum_{i\neq k} h(X_{ij}, \mathbf{y}_i; X_{kj}, \mathbf{y}_k).$$

With Markov's inequality, we can easily obtain that for any $t > 0$

$$\Pr(\widehat{\lambda}_j - \lambda_j \geq \varepsilon) \leq \exp(-t\varepsilon)\exp(-t\lambda_j)E\{\exp(t\widehat{\lambda}_j)\}. \tag{4.18}$$

Serfling (1980, section 5.1.6) pointed out that the U-statistic $\widehat{\lambda}_j$ can be written as an average of independent and identically distributed variables; that is, $\widehat{\lambda}_j = (n!)^{-1}\sum_{\mathcal{P}}\lambda_1(X_{1j}, \mathbf{y}_1; \cdots, X_{nj}, \mathbf{y}_n)$, where each $\lambda_1(X_{1j}, \mathbf{y}_1; \cdots, X_{nj}, \mathbf{y}_n)$ is an average of $m = [n/2]$ independent and identically distributed random variables, and "$\sum_{\mathcal{P}}$" denotes summation over $n!$ permutations $(i_1, i_2, \ldots, i_n)$ of $(1, 2, \ldots, n)$. We denote $\psi(s) = E[\exp\{sh(X_{ij}, \mathbf{y}_i; X_{kj}, \mathbf{y}_k)\}]$. Since the exponential function is convex, it follows immediately from Jensen's inequality that

$$\begin{aligned} E\{\exp(t\widehat{\lambda}_j)\} &= E\left[\exp\left\{t(n!)^{-1}\sum_{\mathcal{P}}\lambda_1(X_{1j}, \mathbf{y}_1; \cdots, X_{nj}, \mathbf{y}_n)\right\}\right] \\ \leq\ & (n!)^{-1}\sum_{\mathcal{P}} E\left[\exp\left\{t\lambda_1(X_{1j}, \mathbf{y}_1; \cdots, X_{nj}, \mathbf{y}_n)\right\}\right] = \psi^m(t/m). \end{aligned} \tag{4.19}$$

Let $s = t/m$. Combining the results of (4.18) and (4.19), we obtain that

$$\begin{aligned} \Pr(\widehat{\lambda}_j - \lambda_j \geq \varepsilon) &\leq \exp(-t\varepsilon)\left\{\exp(-t\lambda_j/m)\psi(t/m)\right\}^m \\ &= \left\{\exp(-s\varepsilon)\exp(-s\lambda_j)\psi(s)\right\}^m. \end{aligned} \tag{4.20}$$

Recall that $E\left\{h(X_{ij}, \mathbf{y}_i; X_{kj}, \mathbf{y}_k)\right\} = \lambda_j$. With Taylor expansion $\exp\{s(Y-\mu)\} = 1 + s(Y-\mu) + s^2Z/2$, where $0 < Z < (Y-\mu)^2\exp\left\{s_0(Y-\mu)\right\}$ for any random variable Y such that $E(Y) = \mu$, we can obtain that

$$\exp(-s\lambda_j)\psi_1(s) \leq 1 + \frac{s^2}{2}\sqrt{E\left\{h(X_{ij}, \mathbf{y}_i; X_{kj}, \mathbf{y}_k)\right\}^4 E\left[\exp\left\{2s_0(h-\lambda_j)\right\}\right]}. \tag{4.21}$$

Using that fact that $(|a|+|b|)^2 \geq (a^2+b^2) \geq (a+b)^2/2$ and after some simple calculations, we obtain that

$$\begin{aligned}
& E\left[\exp\left\{2s_0 h(X_{ij},\mathbf{y}_i;X_{kj},\mathbf{y}_k)\right\}\right] \leq E\exp\left[2s_0|X_{ij}X_{kj}|\,\|\mathbf{y}_i-\mathbf{y}_k\|\right] \\
\leq\ & E\exp\left[2s_0|X_{ij}X_{kj}|\left\{2\left(\|\mathbf{y}_i\|^2+\|\mathbf{y}_k\|^2\right)\right\}^{1/2}\right] \\
\leq\ & E\left[\exp\left\{2\sqrt{2}s_0|X_{ij}X_{kj}|\left(\|\mathbf{y}_i\|+\|\mathbf{y}_k\|\right)\right\}\right] \leq E\left\{\exp\left(4\sqrt{2}s_0|X_{ij}X_{kj}|\|\mathbf{y}_i\|\right)\right\} \\
=\ & E\left\{\exp\left(2\sqrt{2s_0}|X_{ij}|\|\mathbf{y}_i\|\right)\right\}E\left\{\exp\left(2\sqrt{s_0}|X_{kj}|\right)\right\} \\
\leq\ & E\left[\exp\left\{\sqrt{2s_0}\left(|X_{ij}|^2+\|\mathbf{y}_i\|^2\right)\right\}\right]E\left\{\exp\left(2\sqrt{s_0}|X_{kj}|\right)\right\}.
\end{aligned}$$

By invoking (4.21), the Cauchy-Schwartz inequality and the sub-exponential tail probability condition (C1), it follows that there exists C (independent of n and p) such that

$$\max_{1\leq j\leq p}\exp(-s\lambda_j)\psi(s) < 1+Cs^2.$$

Consequently, for sufficiently small s, we have

$$\max_{1\leq j\leq p}\left\{\exp(-s\varepsilon)\exp(-s\lambda_j)\psi(s)\right\} = 1-\varepsilon s+O(s^2) < 1-\varepsilon s/2. \tag{4.22}$$

Combining the results of (4.20) and (4.22), we show that, for every $\varepsilon>0$, there exists a sufficiently small s such that $\max_{1\leq j\leq p}\Pr(\widehat{\lambda}_j-\lambda_j\geq\varepsilon) \leq (1-\varepsilon s/2)^{n/2}$. Similarly, we can prove that $\max_{1\leq j\leq p}\Pr(\widehat{\lambda}_j-\lambda_j\leq-\varepsilon) \leq (1-\varepsilon s/2)^{n/2}$. Therefore,

$$\begin{aligned}
\Pr\left(\max_{1\leq j\leq p}|\widehat{\lambda}_j-\lambda_j|\geq\varepsilon\right) &\leq p\max_{1\leq j\leq p}\left\{\Pr(\widehat{\lambda}_j-\lambda_j\geq\varepsilon)\right\} \\
&= 2p\exp\left\{n\log\left(1-\varepsilon s/2\right)/2\right\}.
\end{aligned} \tag{4.23}$$

which completes the proof of (4.11).

Let $\delta = \min_{j\in\mathcal{D}}\lambda_j - \max_{j\in\mathcal{I}}\lambda_j$. By the Bonferroni inequality, it follows immediately that

$$\begin{aligned}
& \Pr\left(\min_{j\in\mathcal{D}}\widehat{\lambda}_j \le \max_{j\in\mathcal{I}}\widehat{\lambda}_j\right) = \Pr\left(\min_{j\in\mathcal{D}}\widehat{\lambda}_j - \min_{j\in\mathcal{D}}\lambda_j + \delta \le \max_{j\in\mathcal{I}}\widehat{\lambda}_j - \max_{j\in\mathcal{I}}\lambda_j\right) \\
= & \Pr\left(\left|\min_{j\in\mathcal{D}}\widehat{\lambda}_j - \min_{j\in\mathcal{D}}\lambda_j\right| \ge \delta/2 \text{ or } \left|\max_{j\in\mathcal{I}}\widehat{\lambda}_j - \max_{j\in\mathcal{I}}\lambda_j\right| \ge \delta/2\right) \\
\le & \Pr\left(\left|\min_{j\in\mathcal{D}}\widehat{\lambda}_j - \min_{j\in\mathcal{D}}\lambda_j\right| \ge \delta/2\right) + \Pr\left(\left|\max_{j\in\mathcal{I}}\widehat{\lambda}_j - \max_{j\in\mathcal{I}}\lambda_j\right| \ge \delta/2\right).
\end{aligned}$$

The first term in the right-hand side of the above inequality is bounded by $\Pr\left(\sup_{j\in\mathcal{D}}|\widehat{\lambda}_j - \lambda_j| \ge \delta/2\right)$, and the second term is less than $\Pr\left(\sup_{j\in\mathcal{I}}|\widehat{\lambda}_j - \lambda_j| \ge \delta/2\right)$. Both converge to 0 in probability by using (4.23), which completes the proof of Theorem 4.2. □

4.4.3 Proof of Theorem 4.3

This proof is a slight modification of the proof of Theorem 4.2. In this proof we use condition (C2) instead of condition (C1) to obtain the probability bounds.

We first prove (4.13). It is sufficient to derive the consistency of $\widehat{\lambda}_j$. Since q is fixed, without loss of generality and for ease of illustration, it is assumed throughout this proof that $q = 1$. Let $h(X_j, \mathbf{y}; \widetilde{X}_j, \widetilde{\mathbf{y}}) = -X_j\widetilde{X}_j\|\mathbf{y} - \widetilde{\mathbf{y}}\|$ be the kernel of the U-statistic $\widehat{\lambda}_j$. It can be easily seen from the uniform bound condition (C2) that $-2a^2b \le h(X_j, \mathbf{y}; \widetilde{X}_j, \widetilde{\mathbf{y}}) \le 2a^2b$. Following similar arguments for obtaining (4.20), for any $s > 0$, we can have

$$\Pr(\widehat{\lambda}_j - \lambda_j \ge t) \le \exp(-ts)\left\{\exp(-\lambda_j s/\tau)\psi(s/\tau)\right\}^{\tau},$$

where $\tau = [n/2]$, which, together with the exponential inequality in Lemma 5.6.1.A

of Serfling (1980, page 200), entails that

$$\Pr(\widehat{\lambda}_j - \lambda_j \geq t) \;\leq\; \exp\left(-ts + 2a^4b^2s^2/\tau\right).$$

By choosing $s = t\tau/(4a^4b^2)$, the right hand side of the ineuqality attains its minimum $\exp\{-\tau t^2/(8a^4b^2)\}$, which together with the symmetry of U-statistic implies

$$\Pr(|\widehat{\lambda}_j - \lambda_j| \geq t) \;\leq\; 2\exp\left\{-\tau t^2/(8a^4b^2)\right\}. \tag{4.24}$$

We also remark here that (4.24) is also a direct result of Theorem 5.6.1.A of Serfling (1980, page 201). Therefore, the first part of Theorem 4.3 is proven.

The second part for proving (5.15) follows literally the arguments for proving the second part of Theorem 4 in Fan and Song (2009). We sketch the proof as follows.

We write the event $\mathcal{C}_n = \left\{\max_{j\in\mathcal{D}} \left|\widehat{\lambda}_j - \lambda_j\right| \leq cn^{-\kappa}\right\}$. On the event $\mathcal{C}_n$ and under the condition (C3), we have $\widehat{\lambda}_j \geq cn^{-\kappa}$, for all $j \in \mathcal{D}$. This implies immediately that $\Pr(\mathcal{D} \subseteq \widehat{\mathcal{D}}) \geq \Pr(\mathcal{C}_n) = 1 - \Pr(\mathcal{C}_n^c)$. Without much difficulty we can obtain that

$$\begin{aligned}
\Pr(\mathcal{C}_n^c) &= 1 - \Pr\left\{\min_{j\in\mathcal{D}} \left|\widehat{\lambda}_j - \lambda_j\right| \geq cn^{-\kappa}\right\} \\
&= s_n \Pr\left\{\left|\widehat{\lambda}_j - \lambda_j\right| \geq cn^{-\kappa}\right\} \leq O\left\{s_n \exp\left(-n^{1-2\kappa}/16a^4b^2\right)\right\}.
\end{aligned}$$

This completes the proof of this theorem.

Chapter 5

A Second-Moment Sure Independence Ranking and Screening Approach

5.1 Introduction

In ultrahigh dimensional regressions where the predictor dimension p is very large relative to the available sample size n, it is often assumed that only a small subset of predictors are necessary to explain the response variable. This is the so-called sparsity principle in the literature. How to identify important ones from a large amount of predictors plays a central role in the ultrahigh dimensional regression analysis nowadays.

Analysis of ultrahigh dimensional data calls for new statistical methodologies that are numerically stable and computationally feasible. Some penalized pseudo-likelihood approaches are proposed in the literature to identify important predictors relevant to the response, such as the bridge regression (Frank and Friedman,

1993), the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001), Dantzig selector (Candes and Tao, 2007), the penalized linear unbiased selector (Zhang, 2010, PLUS) and their variations such as the elastic net (Zou and Hastie, 2005), the adaptive LASSO (Zou, 2006) and nonnegative garrote (Yuan and Lin, 2007), etc. These approaches have some nice theoretical properties even when p is larger than n. However, in implementations they may not perform well due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (Fan, Samworth and Wu, 2009).

Fan and Lv (2008) emphasized the importance of the marginal regression in favor of its computational simplicity. When the response is univariate, they proposed a sure independence screening (SIS) procedure. The SIS measures the importance of each predictor using the marginal Pearson correlation coefficient. They also established the sure screening property for the SIS in the linear regression setup where p grows exponentially faster than the available sample size n. That is, with an overwhelming probability, all important predictors will survive after the screening procedure. However, some unimportant predictors, through their correlations with important predictors, will be retained as well. To reduce the false discovery rate, Fan and Lv (2008) suggested to conjunct the SIS with certain penalized pseudo-likelihood approaches. Fan and Song (2009) and Fan, Feng and Song (2009) further generalized the SIS to generalized linear models and nonparametric additive models. These model-based approaches have a nice performance if the working model is close to the underlying true one.

In ultrahigh dimensional data analysis where there is little information about the model structure, Zhu, Li, Li and Zhu (2010) proposed a sure independence ranking and screening (SIRS) procedure. The SIRS calculates the Pearson correlation coefficient marginally between each predictor and a transformed response.

It does not require specification of any model structure. Thus we view it as a model-free approach. In Chapter 4, we also introduced a unified sure independence ranking and screening (USIRS) approach. The USIRS retains the model-free flavor of the SIRS. It can also identify important predictors associated with correlations among multivariate response variables. It is thus more comprehensive than the SIRS. Both the SIRS and USIRS enjoy the ranking consistency property when p diverges exponentially faster than n; that is, the SIRS and USIRS can rank the important predictors above the unimportant ones with an overwhelming probability. This property ensures a clear separation between the important and unimportant predictors, thus offers a possibility to identify exactly all active predictors pending that an ideal cut-off is available.

Fan and Lv (2008) observed that the SIS may miss some important predictors which are marginally independent of the response. To fix this issue, Fan and Lv (2008) proposed an iterative procedure through calculating the correlation between the residuals of the response variable and the remaining predictors in the linear regression setup. This general concept has been extended to generalized linear models (Fan and Song, 2009) and nonparametric additive models (Fan, Feng and Song, 2009). Zhu, Li, Li and Zhu (2010) proposed a modified iterative procedure which computes the correlation between the original response and the residuals of the remaining predictors. This idea retains the model-free flavor, and can be readily applied to the USIRS (see Chapter 4). In their papers, comprehensive simulations demonstrate the good performance of the iterative procedures.

We observe that these existing procedures, as well as their iterative counterparts, fail to identify a class of important predictors which exhibit "symmetric patterns" to the response variable $\mathbf{y}$. In our context the symmetry pattern means that $E(X_j \mid \mathbf{y}) = 0$ for an important predictor X_j. To address this issue, in

this chapter we develop a complementary methodology through a second-moment approach. This proposed second-moment unified sure independence ranking and screening (USIRS-II, for short) approach can be regarded as an important member in the marginal regression family. It retains the model-free feature, which is a very appealing property in ultrahigh dimensional regression analysis, particularly when there is little information about the model structure. It also inherits the merits of the USIRS in that it allows different types of response variable (e.g., quantitative and categorical), no matter the response is univariate or multivariate. It thus can readily be used to capture important predictors which describe correlations among multivariate response variables. In parallel to the SIRS and USIRS, the new approach has both the sure screening and ranking consistency properties even when the dimension p grows at an exponential rate of n. Besides its nice properties in an asymptotic sense, the numerical studies demonstrate that it has a competent finite-sample performance in a wide range of regression models.

The rest of this chapter is organized as follows. In Section 5.2 we review systematically the sure independence screening family and illustrate their limitation through a motivating example. Then we propose our second-moment approach to avoid this limitation. We discuss the estimation at the sample level, and investigate the theoretical properties of the sample estimator. We also establish the sure screening and ranking consistency properties for the second-moment approach. In Section 5.3 we demonstrate its finite sample performance through comprehensive numerical examples. All technical derivations are given in Section 5.4.

5.2 Sure Independence Ranking and Screening

5.2.1 A Review of Marginal Regression Family

Before introducing our new approach, we first systematically review the independence screening family. For notational clarity, let $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathrm{R}^p$ be the predictor vector, and $\mathbf{y} = (Y_1, \ldots, Y_q) \in \mathrm{R}^q$ be the response vector for $q > 1$. When the response variable is univariate (i.e., $q = 1$), we use Y instead of $\mathbf{y}$ for clarification. To ease the subsequent illustration, we assume throughout that all the coordinates of $\mathbf{x}$ are standardized to have zero mean and unit variance, that is, $E(X_j) = 0$ and $\mathrm{var}(X_j) = 1$, for $1 \leq j \leq p$.

The general methodology of independence screening is to develop a marginal utility, say λ_j, which is derived from the marginal regression of $\mathbf{y}$ onto X_j, for $j = 1, \ldots, p$. The marginal utility λ_j can often be rapidly computed. Thus the independence screening procedures avoid the numerical instability in ultrahigh dimensional statistical learning problems. In usual practice, it is often reasonable to expect the magnitude of λ_j to be large if X_j is an important feature and small otherwise. Therefore, λ_j can be utilized to rank the importance of X_j, although the marginal regression of $\mathbf{y} \mid X_j$ misspecifies the joint regression of $\mathbf{y} \mid \mathbf{x}$. Instead of using λ_j itself, the independence screening procedures use its rank by choosing a subset of predictors, that is,

$$\mathcal{D}^\star = \{j : |\lambda_j| \geq \nu_n, j = 1, \ldots, p\},$$

where ν_n is a predefined threshold value. The index set $\mathcal{D}^\star$ is often of our primary interest because it contains the indices of all important predictors under mild conditions.

In the context of linear regression with a univariate response variable, Fan and Lv (2008) proposed the sure independence screening (SIS) procedure which ranks the importance of each predictor through the marginal Pearson correlation coefficient or equivalently, the marginal covariance in that each predictor coordinate is standardized to have zero mean and unit variance. They adopted the marginal utility

$$\lambda_j^{sis} \stackrel{\text{def}}{=} \text{cov}^2\left(X_j, Y\right) = \text{cov}^2\left\{E(X_j \mid Y), Y\right\}. \tag{5.1}$$

The last equality in (5.1) follows from the law of iterated expectations.

In some situations the univariate response variable may be discrete or categorical. Fan and Song (2009) considered a general sure independence screening (GSIS) procedure which applies to generalized linear models. The GSIS employes the marginal utility

$$\lambda_j^{gsis} \stackrel{\text{def}}{=} \left\{ \beta_{1j}^2 : \underset{\beta_{0j},\beta_{1j}}{\arg\max} E\left\{\ell\left(\beta_{0j} + \beta_{1j} X_j, Y\right)\right\} \right\}, \tag{5.2}$$

where $\ell(\cdot,\cdot)$ is the logarithm of the posited likelihood function of $Y \mid X_j$. If we posit normal likelihood when Y is continuous, then the GSIS is equivalent to the SIS. However, the GSIS has wider applications as it allows for discrete or categorical response variables.

Fan, Feng and Song (2009) proposed a nonparametric independence screening (NIS) procedure for nonparametric additive models, where they suggested the marginal utility

$$\lambda_j^{nis} \stackrel{\text{def}}{=} E\left\{E^2(Y \mid X_j)\right\}. \tag{5.3}$$

The performance of these three model-based screening procedures depends heavily upon whether the working model is close to the underlying true model.

When there is little information about the regression structure, which is often the case in ultrahigh dimensional data analysis, Zhu, Li, Li and Zhu (2010) proposed a model-free sure independent ranking and screening (SIRS) procedure which ranks the importance of X_j through the marginal utility

$$\lambda_j^{sirs} = E\big[\mathrm{cov}^2\big\{X_j, \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\big\}\big] = E\big[\mathrm{cov}^2\big\{E(X_j \mid Y), \mathbf{1}(Y < \widetilde{Y}) \mid \widetilde{Y}\big\}\big], \quad (5.4)$$

where $\widetilde{Y}$ is an independent copy of Y, and $\mathbf{1}(Y < \widetilde{Y})$ is an indicator function which takes value one if $Y < \widetilde{Y}$, and zero otherwise.

In Chapter 4, we proposed a unified sure independence ranking and screening (USIRS) procedure which allows for both univariate and multivariate responses. The USIRS uses

$$\lambda_j^{usis} \stackrel{\text{def}}{=} -E\big(X_j \widetilde{X}_j \, \|\mathbf{y} - \widetilde{\mathbf{y}}\|\big) = -E\big\{E(X_j \mid Y) E(\widetilde{X}_j \mid \widetilde{\mathbf{y}}) \, \|\mathbf{y} - \widetilde{\mathbf{y}}\|\big\}, \quad (5.5)$$

where $(\widetilde{X}_j, \widetilde{\mathbf{y}})$ is an independent copy of $(X_j, \mathbf{y})$, and $\|\cdot\|$ denotes the Euclidean norm. The last equalities of (5.4) and (5.5) follow from the law of iterated expectations.

These model-free independent screening procedures, together with those three model-based procedures, may fail to identify important predictors which exhibit "symmetric patterns" with the response variable. In the present context, the "symmetric pattern" means that X_j satisfies $E(X_j \mid \mathbf{y}) = 0$. To illustrate this issue, we consider the following toy example. Suppose the underlying true model

is of the form

$$Y = X_1^2 + |X_2|\varepsilon, \tag{5.6}$$

where X_js and ε are independent and normally distributed with zero mean and unit variance. Only the first two components of $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ are truly important, while the rest $p-2$ components are redundant. Note that $E(X_j \mid Y) = 0$ for $j = 1, \ldots, p$; that is, X_1 and X_2, which are of primary interest, as well as all other predictors, have a symmetric relationship with Y. In model (5.6), it can be easily verified that $\lambda_j^{sis} = \lambda_j^{gsis} = \lambda_j^{sirs} = \lambda_j^{usis} = 0$ for $j = 1, \ldots, p$, and $\lambda_1^{nis} = 4$ and $\lambda_j^{nis} = 1$ for $j = 2, \ldots, p$. Therefore, the SIS, GSIS, SIRS and USIRS cannot pick out any important predictors, and the NIS can only identify X_1 as an important one. The NIS fails because it cannot capture any information beyond the mean regression function.

One may argue that the posited normal likelihood function accounts for the failure of the GSIS. To get more insight into this phenomenon, we consider a logistic model where the binary response Y is a Bernoulli random variable with a success probability $p(\mathbf{x}) = 1/\left\{1 + \exp\left(-X_1^2\right)\right\}$. In this motivating example only X_1 is important, and X_1 satisfies that $E(X_1 \mid Y) = E(X_1 \mid X_1^2) = 0$. If we use the marginal logistic regression to recruit important predictors, then we will model the mean function by $\theta_j(X_j) = 1/\left\{1 + \exp\left(-\beta_{0j} - \beta_{1j}X_j\right)\right\}$, as suggested by the GSIS procedure. Clearly, the marginal mean function for X_1 is misspecified, and this is a special case of model misspecification we discussed in Chapter 3. Using the similar idea adopted in Chapter 3, we can find the relation between the coefficient of X_1^2 in the true model and the coefficient of X_1 in the misspecified model. Recall

that the marginal logistic regression has the form

$$\begin{aligned} E\left\{\ell\left(\beta_0+\beta_1 X_j, Y\right)\right\} &= E\left[p(\mathbf{x})\log\left\{\theta_j(X_j)\right\}+\left\{1-p(\mathbf{x})\right\}\log\left\{1-\theta_j(X_j)\right\}\right] \\ &= E\left[-\log\left\{1+\exp(\beta_{0j}+\beta_{1j}X_j)\right\}+p(\mathbf{x})\left(\beta_{0j}+\beta_{1j}X_j\right)\right]. \end{aligned}$$

For any $j=1,\ldots,p$, the maximizer of $E\left\{\ell\left(\beta_{0j}+\beta_{1j}X_j, Y\right)\right\}$ must satisfy

$$\partial E\left\{\ell\left(\beta_{0j}+\beta_{1j}X_j, Y\right)\right\}/\partial\beta_{kj}=0,\ k=0,1.$$

In other words, we have the following equations:

$$\frac{\partial E\left\{\ell\left(\beta_{0j}+\beta_{1j}X_j, Y\right)\right\}}{\partial\beta_{0j}} = E\left\{-\frac{\exp(\beta_{0j}+\beta_{1j}X_j)}{1+\exp(\beta_{0j}+\beta_{1j}X_j)}+p(\mathbf{x})\right\}=0,\ \text{and}$$

$$\frac{\partial E\left\{\ell\left(\beta_{0j}+\beta_{1j}X_j, Y\right)\right\}}{\partial\beta_{1j}} = E\left\{-\frac{\exp(\beta_{0j}+\beta_{1j}X_j)}{1+\exp(\beta_{0j}+\beta_{1j}X_j)}X_j+p(\mathbf{x})X_j\right\}=0.$$

Because of the fact that $p(\mathbf{x})$ is an even function of X_1 and the fact that $E(X_1 \mid Y)=E(X_1 \mid X_1^2)=0$, we obtain

$$E\left\{p(\mathbf{x})X_j\right\}=E\left\{p(\mathbf{x})E\left(X_j|X_1^2\right)\right\}=0.$$

Thus it follows immediately that $\beta_{0j}=\text{logit}\left(E\left\{p(\mathbf{x})\right\}\right)$ and $\beta_{1j}=0$ is the unique maximizer for $j=1,\ldots,p$; that is, the GSIS fails to identify the important feature X_1. In fact we also observe this phenomenon under other generalized linear models (e.g., Poisson model) and even more general situations. This confirms that the GSIS fails to identify the important predictor X_j if it satisfies $E(X_j \mid Y)=0$.

If we have the prior information that all symmetric important features are

contained only in the mean regression function, then one may argue that we can consider all possible interactions and quadratic terms. However, when these terms are considered, the dimensionality grows even more quickly; for example, considering possible interactions among thousands of predictors yields the number of parameters in the order of millions. Thus this strategy is typically computationally infeasible.

Note that $E(X_j \mid \mathbf{y}) = 0$ implies X_j is marginally uncorrelated with $\mathbf{y}$, as we can observe from (5.1). To identify important predictors which are marginally uncorrelated with the response variable, Fan and Lv (2008) and Zhu, Li, Li and Zhu (2010) proposed respectively different iterative procedures. However, following our previous discussions, we can conclude that those iterative algorithms cannot identify an important predictor X_j either if X_j has a symmetric relationship with $\mathbf{y}$ in the sense that $E(X_j \mid \mathbf{y}) = 0$.

Our observation reveals a limitation of the present independence screening family, and calls for new methodologies to fix this issue. We next discuss a possible solution.

5.2.2 A Second-Moment Approach

In this subsection we design a second-moment approach that can identify the important predictor X_j even if it satisfies $E(X_j \mid \mathbf{y}) = 0$. We illustrate our rationale through model (5.6). It can be seen that, although $E(X_j \mid Y) = 0$ for $j = 1, 2$, the second-moment $E(X_j^2 \mid Y) \neq 0$ and are still isotonic functions of Y, for $j = 1, 2$. The variation of the second-moment $E(X_j^2 \mid Y)$ indicates that X_js are relevant to Y. This observation motivates us to utilize $E(X_j^2 \mid Y)$ in place of $E(X_j \mid Y)$ in previous independent screening procedures. Thinking that the USIRS has the

model-free flavor and wide applications (e.g., it allows univariate and multivariate response variables), we choose to use a centralized quantity $E(X_j^2 \mid Y) - 1$ in place of $E(X_j \mid Y)$ in the USIRS method; that is, we use the following marginal utility λ_j in the second-moment approach,

$$\begin{aligned} \lambda_j &\stackrel{\text{def}}{=} -E\left[\left\{E(X_j^2 \mid \mathbf{y}) - 1\right\}\left\{E(\widetilde{X}_j^2 \mid \widetilde{\mathbf{y}}) - 1\right\}\|\mathbf{y} - \widetilde{\mathbf{y}}\|\right] \\ &= -E\left\{(X_j^2 - 1)(\widetilde{X}_j^2 - 1)\,\|\mathbf{y} - \widetilde{\mathbf{y}}\|\right\}. \end{aligned} \tag{5.7}$$

Without notational confusion we use λ_j to denote specifically the marginal utility of our second-moment approach in the subsequent development in order to avoid introducing a new notation. In parallel to what we did in Chapter 4, we define the active and inactive predictors in the following way to characterize the sparsity

$$\begin{aligned} \mathcal{D} &= \{j : \Psi(\mathbf{t} \mid \mathbf{x}) \text{ functionally depends on } X_j \text{ for some } \mathbf{t} \in \mathrm{R}^q\}, \\ \mathcal{I} &= \{j : \Psi(\mathbf{t} \mid \mathbf{x}) \text{ does not functionally depend on } X_j \text{ for any } \mathbf{t} \in \mathrm{R}^q\}, \end{aligned}$$

where $\Psi(\mathbf{t} \mid \mathbf{x}) = E\{\exp(i\mathbf{t}^{\mathrm{T}}\mathbf{y}) \mid \mathbf{x}\}$ for $\mathbf{t} \in \mathrm{R}^q$ is the conditional characteristic function of $\mathbf{y}$ given $\mathbf{x}$. We further write $\mathbf{x}_{\mathcal{D}} = \{X_j : j \in \mathcal{D}\}$ and $\mathbf{x}_{\mathcal{I}} = \{X_j : j \in \mathcal{I}\}$, and refer to $\mathbf{x}_{\mathcal{D}}$ as an *active* predictor vector and its complement $\mathbf{x}_{\mathcal{I}}$ as an *inactive* predictor vector. Recall that the sparsity principle implies that

$$\mathbf{y} \perp\!\!\!\perp \mathbf{x}_{\mathcal{I}} \mid \mathbf{x}_{\mathcal{D}}, \tag{5.8}$$

where $\perp\!\!\!\perp$ stands for statistical independence. In model (5.8), the variations of $\mathbf{y}$ are completely characterized by the changes of $\mathbf{x}_{\mathcal{D}}$, and $\mathbf{x}_{\mathcal{I}}$ does not provide any additional information for inferences about $\mathbf{y}$ when $\mathbf{x}_{\mathcal{D}}$ is known. Thus $\mathbf{x}_{\mathcal{I}}$ is

unimportant and can be excluded from subsequent regression analysis once $\mathbf{x}_{\mathcal{D}}$ is known.

Theorem 5.1 justifies at the population level that the second-moment approach using the marginal utility (5.7) can distinguish the important predictors from the unimportant ones under the following conditions.

(A.1) The inequality condition holds uniformly in p:

$$\frac{\lambda_{\max}^2\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{I}})\right\}}{\lambda_{\min}^4\left\{\boldsymbol{\Sigma}_{\mathcal{D}}\right\}} < \frac{\min\limits_{j\in\mathcal{D}}\left\{\lambda_j\right\}}{\lambda_{\max}\{\boldsymbol{\Lambda}_{\mathcal{D}}\}}, \tag{5.9}$$

where $\boldsymbol{\Lambda}_{\mathcal{D}} = -E\left\{\left(\mathbf{x}_{\mathcal{D}}\mathbf{x}_{\mathcal{D}}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\mathcal{D}}\right)\left(\widetilde{\mathbf{x}}_{\mathcal{D}}\widetilde{\mathbf{x}}_{\mathcal{D}}^{\mathrm{T}} - \boldsymbol{\Sigma}_{\mathcal{D}}\right)\|\mathbf{y} - \widetilde{\mathbf{y}}\|\right\}$, $\boldsymbol{\Sigma}_{\mathcal{D}} = \mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})$, and $\lambda_{\max}\{\mathbf{B}\}$ and $\lambda_{\min}\{\mathbf{B}\}$ denote respectively the largest and smallest eigenvalues of a matrix $\mathbf{B}$. Throughout this article, when we say that "$a_p < b_p$" holds uniformly in p, it means that $\limsup\limits_{p\to\infty}(a_p - b_p) < 0$.

(A.2) The linearity condition:

$$E(\mathbf{x} \mid \mathbf{x}_{\mathcal{D}}) = \mathrm{cov}(\mathbf{x}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\mathbf{x}_{\mathcal{D}}. \tag{5.10}$$

(A.3) The constant variance condition:

$$\mathrm{var}(\mathbf{x} \mid \mathbf{x}_{\mathcal{D}}) = \mathrm{cov}(\mathbf{x}, \mathbf{x}^{\mathrm{T}}) - \mathrm{cov}(\mathbf{x}, \mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}^{\mathrm{T}}). \tag{5.11}$$

Theorem 5.1. *Suppose that conditions (A.1)-(A.3) are true. Then*

$$\max_{j\in\mathcal{I}}\lambda_j < \min_{j\in\mathcal{D}}\lambda_j \tag{5.12}$$

holds uniformly in p.

Theorem 5.1 ensures that the marginal utilities of $\mathbf{x}_{\mathcal{D}}$ are always larger than those of $\mathbf{x}_{\mathcal{I}}$ when conditions (A.1)-(A.3) are true. Condition (A.1) is parallel to condition (C1) in Zhu, Li, Li and Zhu (2010) and conditions (3)-(4) in Fan and Lv (2008). This is a key assumption to guarantee our screening procedure to have the ranking consistency property. It also rules out the case where there is strong collinearity between $\mathbf{x}_{\mathcal{D}}$ and $\mathbf{x}_{\mathcal{I}}$, or among $\mathbf{x}_{\mathcal{D}}$ themselves. Note that the quantity $\min_{j\in\mathcal{D}}\lambda_j$ on the right hand side of (5.9) reflects the signal strength of the individual important predictor. This is similar to condition (3) of Fan and Lv (2008, page 870), which requires that the contribution of an important predictor is sufficiently large.

In Chapter 4, we assumed conditions (A.1) and (A.2) when we employed the first-moment $E(X_j \mid \mathbf{y})$. Because in the present context we consider the second-moment approach, we assume an additional condition (A.3) which requires the second-moment of $\mathbf{x}$ remains constant when $\mathbf{x}_{\mathcal{D}}$ is given. Conditions (A.2) and (A.3) are widely assumed in the dimension reduction literature. See for example Zhu, Zhu and Feng (2010). These two conditions follow immediately from the partial orthogonality condition (Fan and Song, 2009 corollary 1). In addition, conditions (A.2) and (A.3) are true when $\mathbf{x}$ is multivariate normal. The normality assumption is often assumed in ultrahigh dimensional data analysis. See for instance, Fan and Lv (2008), Bickel and Levina (2008) and Wang (2009), etc.

We remark here that λ_j is a nonnegative constant for $j = 1, \ldots, p$. We allow correlations between $\mathbf{x}_{\mathcal{D}}$ and $\mathbf{x}_{\mathcal{I}}$, hence λ_j can be nonzero for $j \in \mathcal{I}$. In fact, under conditions (A.2) and (A.3), $\lambda_j = 0$ for $j \in \mathcal{I}$ if and only if $\mathrm{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{I}}^{\mathrm{T}}) = \mathbf{0}$. This is a direct conclusion of (5.19) in the proof of Theorem 5.1. We can also see that, when $\mathbf{x}_{\mathcal{D}}$ and $\mathbf{x}_{\mathcal{I}}$ are independent, (5.9) and (5.12) follow immediately.

5.2.3 The Sample Properties

In this subsection we discuss the implementation of this second-moment approach at the sample level. Suppose a random sample $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$ is available. For the sake of simplicity, we assume hereafter that the sample predictors are marginally standardized; that is, $n^{-1}\sum_{i=1}^{n} X_{ij} = 0$ and $n^{-1}\sum_{i=1}^{n} X_{ij}^2 = 1$, for $j = 1, \ldots, p$. The U-statistic estimator (Serfling, 1980) of λ_j defined in (5.7) is given by

$$\widehat{\lambda}_j = -\left\{n(n-1)\right\}^{-1} \sum_{i \neq k} (X_{ij}^2 - 1)(X_{kj}^2 - 1)\|\mathbf{y}_i - \mathbf{y}_k\|.$$

We select a subset of predictors whose marginal utilities are greater than a specified threshold. For instance, we choose

$$\widehat{\mathcal{D}} = \left\{ j : \widehat{\lambda}_j \geq cn^{-\kappa}, j = 1, \ldots, p \right\},$$

where c and κ are certain pre-specified threshold values.

Next we discuss the sample properties of this second-moment approach. We assume the following regularity conditions.

(B.1) $E\left(X_j^8\right) < \infty$ and $E\left(\|\mathbf{y}\|^8\right) < \infty$;

(B.2) Let $\Omega_n\left(X_j, \mathbf{y}\right) = \{(X_j, \mathbf{y}) : |X_j| \leq K_n, \|\mathbf{y}\|_\infty \leq K_n^\star\}$ for some sufficiently large positive constants K_n and $K_n^\star$, where $\|\cdot\|_\infty$ is the supremum norm. Let $\Omega_n^c\left(X_j, \mathbf{y}\right)$ be the complement of $\Omega_n\left(X_j, \mathbf{y}\right)$. We choose K_n and $K_n^\star$ large enough to ensure that $\Pr\left\{(X_j, \mathbf{y}) \in \Omega_n^c\left(X_j, \mathbf{y}\right)\right\} \leq m_1 \exp\left\{-m_0(K_n K_n^\star)^\alpha\right\}$, for $j = 1, \ldots, p$, and some positive constants m_0, m_1 and α.

(B.3) The minimum marginal utility of important predictors satisfies

$$\min_{j\in\mathcal{D}} \lambda_j \geq 2cn^{-\kappa}, \text{ for some constants } c \text{ and } 0 \leq \kappa < 1/2.$$

For ease of presentation, we let

$$\alpha_n = 4\exp\left\{-c^2 n^{1-2\kappa}/\left\{128q\text{var}(X_j^2)\left(K_n K_n^\star\right)^2\right\}\right\} + nm_1\exp\left\{-m_0(K_n K_n^\star)^\alpha\right\}.$$

Theorem 5.2. *Under the conditions (B.1)-(B.2), if* $n^{1-2\kappa}/\left(K_n^2 K_n^{\star 2}\right) \to \infty$, *then*

$$\Pr\left(|\widehat{\lambda}_j - \lambda_j| \geq cn^{-\kappa}\right) \leq \alpha_n. \tag{5.13}$$

If we further assume the conditions of Theorem 5.1, then

$$\Pr\left(\max_{j\in\mathcal{I}} \widehat{\lambda}_j < \min_{j\in\mathcal{D}} \widehat{\lambda}_j\right) \;\leq\; 4p\alpha_n, \text{ as } n \to \infty. \tag{5.14}$$

If in addition, condition (B.3) holds, then

$$\Pr\left(\mathcal{D} \subseteq \widehat{\mathcal{D}}\right) \geq 1 - 2s_n\alpha_n, \text{ as } n \to \infty, \tag{5.15}$$

where s_n *is the cardinality of* $\mathcal{D}$.

(5.14) ensures that the second-moment approach using the marginal utility (5.7) has the ranking consistency property, and (5.15) implies that it also has the sure screening property. To balance the two terms in the upper bound of (5.13), the optimal order of $K_n K_n^\star$ is given by $K_n K_n^\star = n^{(1-2\kappa)/(\alpha+2)}$, and thus

$$\Pr\left(\max_{1\leq j\leq p} |\widehat{\lambda}_j - \lambda_j| \geq cn^{-\kappa}\right) \leq O\left\{p\exp\left(-c_0 n^{(1-2\kappa)(\alpha+1)/(\alpha+2)}\right)\right\},$$

for a positive constant c_0. When the predictors and the responses are uniformly bounded, then K_n and $K_n^\star$ can be taken as finite constants. In this case

$$\Pr\left(\max_{1\le j\le p}|\widehat{\lambda}_j-\lambda_j|\ge cn^{-\kappa}\right)\le O\left\{p\exp\left(-c_0n^{(1-2\kappa)}\right)\right\}.$$

In both aforementioned cases, the tail probability in Theorem 5.2 is exponentially small. In other words. we can handle the NP-dimensionality:

$$\log p_n=o\left(n^{(1-2\kappa)(\alpha+1)/(\alpha+2)}\right),$$

where $\alpha=\infty$ for cases of uniformly bounded covariates and responses.

5.3 Numerical Studies

5.3.1 Simulations

In this section, we conduct intensive simulation studies to demonstrate the performance of different screening procedures in the presence of the "symmetry patterns". For ease of illustration, we refer to our second-moment unified independence ranking and screening approach as the USIRS-II method. We compare the USIRS-II with several competitors such as the sure independent screening (SIS) approach and its iterative algorithm ISIS (Fan and Lv, 2008), the nonparametric independent screening (NIS) procedure in Fan, Feng and Song (2009), the sure independence ranking and screening (SIRS) approach in Zhu, Li, Li and Zhu (2010), the unified sure independence ranking and screening (USIRS) approach and its iterative algorithm IUSIRS (see Chapter 4 and Chapter 5, Section 5.2). For the iterative algorithms such as ISIS and IUSIRS, we choose five predictors

with largest marginal utilities at each iteration. The iteration proceeds until we choose $[n/\log n]$ predictors, where $[a]$ denotes the maximum integer not greater than a. Unless otherwise stated, we set the sample size $n = 200$ and the total number of predictors $p = 2000$. In addition, we generate the predictor vector $\mathbf{x}$ from a normal population with mean zero and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = 0.8^{|i-j|}$, and the error term ϵ from the standard normal population. Each experiment is repeated for 1000 times.

The following three different criteria are adopted to evaluate the performance of different independence screening procedures:

1. The minimal model size (denoted by $\mathcal{S}$) that measures the minimal number of predictors to ensure the inclusion of all truly active predictors. If a procedure has the ranking consistency property, we may expect $\mathcal{S}$ to be close enough to the number of truly important predictors. We report the minimum, the first quartile, the median, the third quartile and the maximum number of $\mathcal{S}$ for each independence screening method out of 1000 replications.

2. The proportion (denoted by $\mathcal{P}_A$) that all active predictors are ranked at the top $[n/\log n]$ positions out of 1000 replications. We follow Fan and Lv (2008) by using the hard-thresholding rule to select the $[n/\log n]$ predictors which are ranked at the top. We expect $\mathcal{P}_A$ to be close to 1 if an independence screening procedure has the sure screening property.

3. The proportion (denoted by $\mathcal{P}_S$) that an individual important predictor is ranked at the top $[n/\log n]$ positions out of 1000 repetitions. It can also be used to assess the sure screening property. Besides, it helps us understand which predictors are likely missed by a specific independence screening procedure. We expect the value of $\mathcal{P}_S$ to be close to 1 if an independent screening

procedure is able to identify the individual important predictor.

Example 5.1. We start with the following model

$$Y = k_1 (X_1 + X_2 + X_3 + X_4 + X_5) + k_2 X_{20}^2 + \varepsilon. \tag{5.16}$$

We choose $k_1 = 0.5$, 1.0 and 1.5 and $k_2 = 0$ and 1 to control the signal strength of linear and nonlinear components. When $k_2 = 0$, model (5.16) boils down to a linear model, and the number of truly active predictors is 5. When $k_2 = 1$, there are 6 important predictors, and the active predictor X_{20} is symmetrically relevant to Y because $E(X_{20} \mid Y) = 0$. This is similar to our toy example (5.6).

The simulation results are summarized in Tables 5.1-5.3. We first examine the scenario with $k_2 = 0$, which corresponds to the case of no symmetric pattern. It can be seen from Table 5.1 that, when $k_2 = 0$, the USIRS-II and iterative procedures such as the ISIS and IUSIRS perform slightly worse than other competitors in terms of the $\mathcal{S}$ values. When the linear signal is very weak, say, $k_1 = 0.5$, the minimum mode size $\mathcal{S}$ of the USIRS-II can be even larger than the sample size $n = 200$. However, Tables 5.2 and 5.3 both indicate that those large values of $\mathcal{S}$ occur with very small proportions. For example, when $(k_1, k_2) = (0.5, 0)$, the USIRS-II can pick out all five important predictors with a proportion 94.1%. This implies that, with a frequency less than 5.9%, the $\mathcal{S}$ value exceeds $[n/\log n] = 37$.

The simulation results of $k_2 = 1$ are quite different from $k_2 = 0$. Table 5.3 implies that the presence of X_{20}^2 in model (5.16) makes many methods such as the USIRS, SIRS and SIS lose their ranking consistency and sure screening properties. In contrast, both the USIRS-II and NIS perform quite well in terms of the $\mathcal{S}$ values reported in Table 5.1. This is particularly true when $k_1 = k_2 = 1$, in which case the linear and nonlinear components are almost comparable. Table 5.3 shows that

Table 5.1. The minimum model size $\mathcal{S}$ required to include all truly important predictors. The quintuplet in each parenthesis consists of the minimum, the first quartile, the median, the third quartile and the maximum value of $\mathcal{S}$ out of 1000 replications.

		k_2=0					$k_2 = 1$				
	USIRS-II	(5	5	7	13.5	279)	(6	15	31	70.5	919)
	USIRS	(5	5	5	5	6)	(8	225.5	691	1241.5	1997)
	SIRS	(5	5	5	5	7)	(8	362	935	1542	1999)
$k_1 = 0.5$	NIS	(5	5	5	5	6)	(6	6	6	7	26)
	SIS	(5	5	5	5	6)	(8	205	686.5	1292	1995)
	IUSIRS	(5	5	5	5	1741)	(6	164.5	593	1144.5	1997)
	ISIS	(5	5	5	5	1942)	(6	335	841	1430.5	1997)
	USIRS-II	(5	5	5	6	97)	(6	7	9	17	1301)
	USIRS	(5	5	5	5	6)	(11	342	840.5	1355.5	1998)
	SIRS	(5	5	5	5	6)	(13	391	946.5	1418	1999)
$k_1 = 1$	NIS	(5	5	5	5	6)	(6	9	10	15	1608)
	SIS	(5	5	5	5	6)	(13	319	832	1361.5	1999)
	IUSIRS	(5	5	5	5	12)	(6	241	686	1251.5	1991)
	ISIS	(5	5	5	5	7)	(6	313.5	823	1418.5	1997)
	USIRS-II	(5	5	5	6	80)	(6	8.5	19	80.5	1987)
	USIRS	(5	5	5	5	6)	(14	357.5	896.5	1426.5	2000)
	SIRS	(5	5	5	5	6)	(13	394.5	956	1458	1998)
$k_1 = 1.5$	NIS	(5	5	5	5	5)	(7	14	42	157	1955)
	SIS	(5	5	5	5	5)	(17	348.5	897	1419.5	1999)
	IUSIRS	(5	5	5	5	11)	(6	310	729.5	1284	1995)
	ISIS	(5	5	5	5	5)	(6	310	823	1421	1997)

Table 5.2. The proportion $\mathcal{P}_A$ that all truly important predictors can be ranked at the top $[n/\log n]$ positions during 1000 replications.

	$k_2 = 0$			$k_2 = 1$		
	$k_1 = 0.5$	$k_1 = 1$	$k_1 = 1.5$	$k_1 = 0.5$	$k_1 = 1$	$k_1 = 1.5$
USIRS-II	0.941	0.995	0.998	0.571	0.908	0.646
USIRS	1	1	1	0.059	0.031	0.019
SIRS	1	1	1	0.027	0.018	0.019
NIS	1	1	1	1	0.889	0.482
SIS	1	1	1	0.091	0.034	0.023
IUSIRS	0.996	1	1	0.117	0.084	0.041
ISIS	0.996	1	1	0.123	0.130	0.130

Table 5.3. The proportion $\mathcal{P}_S$ for a single truly important predictor that is ranked at the top $[n/\log n]$ positions during 1000 replications.

	$k_2 = 0$					$k_2 = 1$					
$k_1 = 0.5$	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_{20}
USIRS-II	0.968	0.999	1	1	0.971	0.740	0.959	0.989	0.966	0.767	1
USIRS	1	1	1	1	1	1	1	1	1	1	0.059
SIRS	1	1	1	1	1	1	1	1	1	1	0.027
NIS	1	1	1	1	1	1	1	1	1	1	1
SIS	1	1	1	1	1	1	1	1	1	1	0.091
IUSIRS	0.996	1	1	1	1	0.973	1	1	1	1	0.123
ISIS	0.996	1	1	1	1	0.967	1	1	1	1	0.130
$k_1 = 1$	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_{20}
USIRS-II	0.996	1	1	1	0.999	0.982	1	1	1	0.990	0.935
USIRS	1	1	1	1	1	1	1	1	1	1	0.031
SIRS	1	1	1	1	1	1	1	1	1	1	0.018
NIS	1	1	1	1	1	1	1	1	1	1	0.889
SIS	1	1	1	1	1	1	1	1	1	1	0.034
IUSIRS	1	1	1	1	1	0.997	1	1	1	1	0.084
ISIS	1	1	1	1	1	0.995	1	1	1	1	0.130
$k_1 = 1.5$	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_{20}
USIRS-II	0.998	1	1	1	1	0.995	1	1	1	0.996	0.653
USIRS	1	1	1	1	1	1	1	1	1	1	0.019
SIRS	1	1	1	1	1	1	1	1	1	1	0.019
NIS	1	1	1	1	1	1	1	1	1	1	0.482
SIS	1	1	1	1	1	1	1	1	1	1	0.023
IUSIRS	1	1	1	1	1	1	1	1	1	1	0.041
ISIS	1	1	1	1	1	1	1	1	1	1	0.130

both the USIRS-II and NIS can identify X_{20} efficiently. We also note that the iterative algorithms such as the ISIS and IUSIRS can only identify the predictor X_{20} by chance.

Example 5.2. Next we consider four nonlinear models:

$$(N1): \quad Y = (X_1 + X_2 + X_3 + X_4)\,(X_{20} + 1)^2 + \varepsilon;$$

$$(N2): \quad Y = \sqrt{2}X_{20}\{\exp(X_{20}) - 1\}/\{\exp(X_{20}) + 1\} + |X_1X_2X_3X_4|\varepsilon;$$

$$(N3): \quad Y = \{\exp(X_1) + \exp(-X_1)\}/2 + X_2^2 + X_3^2 + \sqrt{|X_4|} + 5|X_{20}|\varepsilon;$$

$$(N4): \quad Y = X_1X_2 + X_3X_4 + 4|X_{20}|\varepsilon.$$

We choose these models based on the following considerations. None of the important predictors are symmetrically related to Y in model $(N1)$. Therefore, all existing procedures are expected to perform well in this model. However, all important predictors in models $(N2)$, $(N3)$ and $(N4)$ are symmetrically related to Y. In addition, in models $(N3)$ and $(N4)$ the important predictor X_{20} represents the heteroscedasticity. In this example we choose $(n, p) = (200, 2000)$ and $(n, p) = (400, 10000)$ to evaluate the performance of these methods.

The simulation results are summarized in Tables 5.4-5.6. We can see that in model $(N1)$ where no symmetric pattern is present, the SIRS and USIRS perform the best, and our proposed approach USIRS-II follows. The other two competitors, the NIS and SIS do not behave well. This can be explained by the fact that the regularity conditions for these two model-based approaches are violated.

The results in Tables 5.4-5.6 also indicate that all methods except the USIRS-II fail to identify all truly active predictors in models $(N2)$-$(N4)$, when either the symmetric patterns or interactions of highly correlated predictors are present. Table 5.4 shows that in these three models, for all our competitors, even the median of the $\mathcal{S}$ values can be very large. Similarly, as shown in Table 5.5, our competitors can only simultaneously identify all truly important predictors with trivial probabilities. However, their failures can be interpreted by different reasons. The NIS fails because it cannot identify important predictors beyond mean regression functions. The SIS fails probably due to model misspecification. The USIRS and SIRS are not effective in identifying symmetric patterns, although they have the model-free feature; that is, they can only identify symmetric patterns or interactions of highly correlated covariates by chance. In contrast, the USIRS-II method performs quite well across all three models. When $(n, p) = (400, 10000)$, the $\mathcal{S}$ values of the USIRS-II are all very close to the number of truly important predictors,

and the proportions $\mathcal{P}_A$ and $\mathcal{P}_S$ are all close to 1.

Table 5.4. The minimum model size $\mathcal{S}$ for Example 5.2. See caption of Table 5.1.

(N1)	$(n,p)=(200,2000)$					$(n,p)=(400,10000)$				
USIRS-II	(5	19	44	107.5	1764)	(5	7	12	28	2113)
USIRS	(5	5	6	6	9)	(5	5	6	6	7)
SIRS	(5	7	8	8	16)	(6	7	8	8	11)
NIS	(5	16	77	461	1984)	(5	29	182	1539	9944)
SIS	(5	74	414.5	1102.5	2000)	(7	220	1697	5090	9997)
(N2)	$(n,p)=(200,2000)$					$(n,p)=(400,10000)$				
USIRS-II	(5	6	8	11	152)	(5	5	5	6	81)
USIRS	(6	399	767	1297	1992)	(9	1870.5	3913.5	6161.5	9976)
SIRS	(56	1178.5	1611.5	1846	2000)	(552	6193	8145	9274	10000)
NIS	(13	172.5	392.5	777.5	1999)	(24	450.5	1031.5	2571	9857)
SIS	(33	811	1373	1757.5	2000)	(15	3964	6823.5	8839.5	9999)
(N3)	$(n,p)=(200,2000)$					$(n,p)=(400,10000)$				
USIRS-II	(5	5	6	10	567)	(5	5	5	6	17)
USIRS	(18	828	1287.5	1653.5	1997)	(79	4212	6415.5	8240	9990)
SIRS	(46	1268.5	1630.5	1859.5	2000)	(94	6302	8185	9406	10000)
NIS	(5	30	131.5	567	1991)	(5	62	494.5	2205	9998)
SIS	(8	1061	1527	1816.5	2000)	(57	5060	7777	9155	9998)
(N4)	$(n,p)=(200,2000)$					$(n,p)=(400,10000)$				
USIRS-II	(5	5	6	9	444)	(5	5	5	6	21)
USIRS	(9	878	1354.5	1686.5	1998)	(33	4474	6760	8536.5	9999)
SIRS	(25	1275	1643.5	1873	2000)	(305	6495.5	8332	9401	9998)
NIS	(5	27	116.5	489.5	1994)	(5	43.5	341.5	1750.5	9986)
SIS	(7	1099.5	1553.5	1821	2000)	(60	5619.5	7938	9189	9993)

Table 5.5. The proportion $\mathcal{P}_A$ for Example 5.2. See caption of Table 5.2.

	$(n,p)=(200,2000)$				$(n,p)=(400,10000)$			
	(N1)	(N2)	(N3)	(N4)	(N1)	(N2)	(N3)	(N4)
USIRS-II	0.443	0.984	0.958	0.974	0.887	0.999	1	1
USIRS	1	0.014	0.002	0.001	1	0.008	0	0.001
SIRS	1	0	0	0.001	1	0	0	0
NIS	0.412	0.010	0.293	0.323	0.356	0.006	0.258	0.304
SIS	0.186	0.001	0.004	0.003	0.155	0.001	0.002	0.001

Example 5.3. In this example we consider a regression with a multivariate response. We generate $\mathbf{y}=(Y_1,Y_2)$ from a normal distribution with zero mean and unit variance. The correlation between Y_1 and Y_2 given $\mathbf{x}$ equals $\cos(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, where

Table 5.6. The proportion $\mathcal{P}_S$ for Example 5.2. See caption of Table 5.3.

	$(n,p)=(200,2000)$					$(n,p)=(400,10000)$				
(N1)	X_1	X_2	X_3	X_4	X_{20}	X_1	X_2	X_3	X_4	X_{20}
USIRS-II	0.679	0.882	0.883	0.714	0.946	0.946	0.996	0.997	0.933	1
USIRS	1	1	1	1	1	1	1	1	1	1
SIRS	1	1	1	1	1	1	1	1	1	1
NIS	1	1	1	0.999	0.412	1	1	1	1	0.356
SIS	1	1	1	1	0.186	1	1	1	1	0.155
(N2)	X_1	X_2	X_3	X_4	X_{20}	X_1	X_2	X_3	X_4	X_{20}
USIRS-II	0.991	1	1	0.996	0.997	1	1	1	1	0.999
USIRS	0.439	0.541	0.555	0.471	0.071	0.395	0.475	0.477	0.403	0.035
SIRS	0.047	0.047	0.047	0.044	0.044	0.024	0.024	0.015	0.017	0.024
NIS	0.521	0.677	0.680	0.550	0.133	0.456	0.606	0.617	0.458	0.071
SIS	0.326	0.375	0.364	0.323	0.015	0.263	0.296	0.281	0.258	0.012
(N3)	X_1	X_2	X_3	X_4	X_{20}	X_1	X_2	X_3	X_4	X_{20}
USIRS-II	0.999	1	1	0.985	0.974	1	1	1	1	1
USIRS	0.092	0.116	0.105	0.081	0.099	0.055	0.057	0.046	0.023	0.045
SIRS	0.037	0.045	0.042	0.031	0.033	0.025	0.020	0.020	0.010	0.020
NIS	0.982	1	0.999	0.913	0.339	1	1	1	1	0.258
SIS	0.100	0.122	0.108	0.081	0.117	0.061	0.065	0.053	0.029	0.066
(N4)	X_1	X_2	X_3	X_4	X_{20}	X_1	X_2	X_3	X_4	X_{20}
USIRS-II	0.991	1	1	0.995	0.988	1	1	1	1	1
USIRS	0.057	0.084	0.085	0.087	0.111	0.043	0.034	0.048	0.037	0.061
SIRS	0.034	0.036	0.047	0.038	0.042	0.026	0.015	0.023	0.013	0.022
NIS	0.913	0.982	0.981	0.923	0.395	0.999	1	1	0.997	0.305
SIS	0.060	0.075	0.084	0.080	0.131	0.043	0.046	0.047	0.039	0.071

$\boldsymbol{\beta} = (0.8, 0.6, 0, \cdots, 0, 0)^{\mathrm{T}}$. In this example, the important predictors X_1 and X_2 describe the correlation structure between the response variables Y_1 and Y_2. It is usually difficult for existing independence screening procedures to pick out such important predictors. We set $(n,p) = (200, 2000)$ and $(n,p) = (400, 2000)$ in this example.

We compare the performance of the USIRS-II with the USIRS and SIRS. To identify the important predictors X_1 and X_2, the USIRS-II and USIRS can be directly implemented. To implement the SIRS in regressions with multivariate

responses, Zhu, Li, Li and Zhu (2010) proposed to use the marginal utility

$$\lambda_j^{sirsm} = \sum_{k=1}^{q} E\left[\mathrm{cov}^2\left\{X_j, \mathbf{1}(Y_k < \widetilde{Y}_k) \mid \widetilde{Y}_k\right\}\right], \tag{5.17}$$

where $\widetilde{\mathbf{y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_q)^{\mathrm{T}}$ is an independent copy of $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$.

The simulation results are summarized in Tables 5.7. It can be seen that, when $n = 200$, all three procedures do not behave well, although the USIRS-II is slightly superior to the other two methods. In contrast, when $n = 400$, the USIRS-II performs quite well. The median of the minimum model size $\mathcal{S}$ is as small as 6, which is very close to the number 5 of truly important predictors. However, the USIRS and SIRS remain a poor performance, because they cannot identify the symmetric patterns in this example.

Table 5.7. The minimum model size $\mathcal{S}$, the proportion $\mathcal{P}_A$ and the proportion $\mathcal{P}_S$ for Example 5.3. See caption of Table 5.1, Table 5.2, Table 5.3.

	$\mathcal{S}$					$\mathcal{P}_A$	$\mathcal{P}_S$	
							X_1	X_2
$n = 200$								
USIRS-II	(2	35	70.5	137	701)	0.278	0.505	0.393
USIRS	(2	662.5	1163	1646	2000)	0.005	0.017	0.015
SIRS	(2	719	1229	1654	2000)	0.004	0.013	0.014
$n = 400$								
USIRS-II	(2	3	6	12	126)	0.990	1	0.990
USIRS	(2	713.5	1163.5	1608	2000)	0.014	0.030	0.028
SIRS	(9	778	1257.5	1630.5	1998)	0.009	0.029	0.020

Example 5.4. We use this example to demonstrate that the USIRS-II performs quite well even when the response variable is discrete or categorial. Towards this end, we consider the following two generalized linear models:

$$(G1) : Y \sim \text{Bernoulli}\left\{\mu(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\},$$

$$(G2) : Y \sim \text{Poisson}\left\{\lambda(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\};$$

where $\text{logit}\left\{\mu(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\} = \log\left\{\lambda(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\} = 1+X_1X_2+X_3X_4$. For comparison purpose, we also include two model-free procedures: the model-free sure independence ranking and screening (SIRS) approach in Zhu, Li, Li and Zhu (2010) and the unified sure independence ranking and screening (USIRS) approach proposed in Chapter 4. We also consider the sure independence screening procedure for generalized linear models (Fan and Song, 2009). Their model-based procedure is referred to as GSIS in this example. Similar to our previous example, we set $(n, p) = (200, 2000)$ and $(n, p) = (400, 2000)$.

We report the simulation results in Tables 5.8. In this example, we observe very similar phenomena as in our previous examples. The USIRS, SIRS and GSIS perform poorly in this case. They can only pick out the important predictors by chance. In contrast, the USIRS-II performs quite well, particularly when the sample size $n = 400$. We also observe that the USIRS-II performs slighter better in model $(G2)$ than in model $(G1)$. This happens possibly due to the fact that the response variable Y in the Poisson model is more informative than that in the logistic model.

5.3.2 An Application

In this section we apply our approach to the gene expression data collected by Scheetz (2006). They delicately chose 120 rats and collected microarrays from the eyes of these rats to analyze the RNA. These microarrays contain more than 30000 different probes, and only 18976 were identified at a level sufficient to be considered "expressed". The gene expression level at the probe 1389163_at (Y)

Table 5.8. The minimum model size $\mathcal{S}$, the proportion $\mathcal{P}_A$ and the proportion $\mathcal{P}_S$ for Example 5.4. See caption of Table 5.1, Table 5.2, Table 5.3.

	$\mathcal{S}$					$\mathcal{P}_A$	$\mathcal{P}_S$			
							X_1	X_2	X_3	X_4
(G1)										
$n = 200$										
USIRS-II	(4	66	122	236.5	1983)	0.116	0.347	0.570	0.551	0.370
USIRS	(284	1531.5	1765	1906	1999)	0	0	0.001	0	0.001
SIRS	(30	1176	1581.5	1826	2000)	0.001	0.007	0.019	0.012	0.008
GSIS	(276	1530.5	1765	1906	1999)	0	0	0.001	0	0.001
$n = 400$										
USIRS-II	(4	9	16	28	792)	0.925	0.955	0.999	0.999	0.966
USIRS	(317	1519	1755.5	1903.5	2000)	0	0.004	0	0.003	0.003
SIRS	(13	1269.5	1604.5	1837	2000)	0.005	0.028	0.031	0.030	0.027
GSIS	(318	1519	1755.5	1903.5	2000)	0	0.005	0	0.004	0.004
(G2)										
$n = 200$										
USIRS-II	(4	11	23	50.5	575)	0.663	0.852	0.950	0.955	0.839
USIRS	(4	23	68	195.5	1571)	0.358	0.594	0.693	0.705	0.590
SIRS	(4	883.5	1505	1851	2000)	0.015	0.061	0.064	0.065	0.055
GSIS	(4	44.5	179.5	1421.5	2000)	0.213	0.479	0.577	0.578	0.470
$n = 400$										
USIRS-II	(4	7	14	30.5	1032)	0.919	0.958	0.996	0.999	0.964
USIRS	(4	15	43	132.5	1469)	0.603	0.764	0.847	0.837	0.751
SIRS	(5	919.5	1556	1852	2000)	0.026	0.074	0.086	0.082	0.078
GSIS	(4	32.5	139.5	1381.5	2000)	0.365	0.629	0.692	0.658	0.608

was identified to be involved in human hereditary diseases of the retina. We are interested in studying how the gene at the probe 1389163_at depends on expression levels at other probes. This is an ultrahigh dimensional dataset in which the dimension $p = 18974$ is much larger than the sample size $n = 120$. We apply our proposed independence screening approach as an initial step to exclude a majority of irrelevant predictors before we obtain a meaningful insight into the dependence between the probe 1389163_at and other probes.

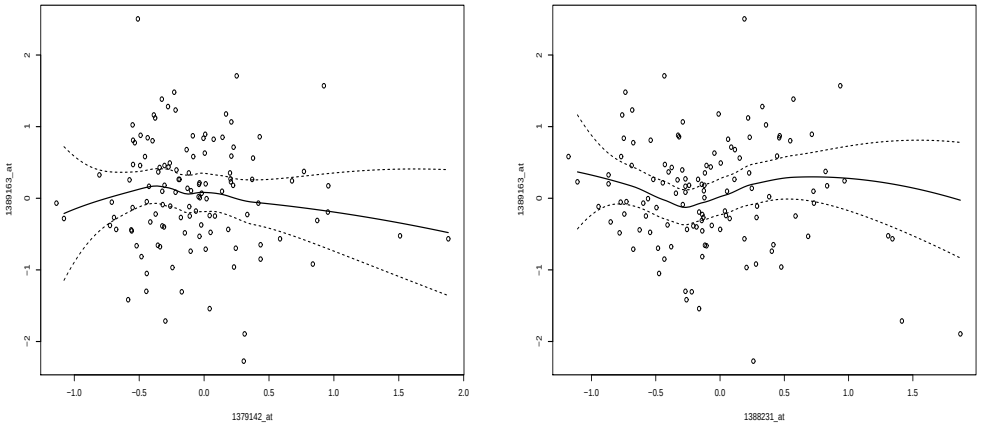
We first standardize all predictors marginally to have zero mean and unit variance during the stage of exploration data analysis. Then we apply the USIRS-

II method to pick out ten predictors with largest marginal utilities. The plots (I.1) and (I.2) in Figure 5.1 give the scatter plots of the gene expression levels at 1389163_at (Y) versus the first two selected predictors (probe 1379142_at and probe 1388231_at). To see how each predictor is related to the conditional mean function, we marginally fit a local polynomial regression over those two selected probes. More specifically, we select 100 evenly spaced grid points between the 0.5%-quantile and the 99.5%-quantile of each predictor, and apply local polynomial regression at each grid point. The solid lines in the plots (I.1) and (I.2) are the estimated mean functions, and the dashed lines are their pointwise 95% confidence intervals. It can be seen that these two regression functions are almost flat, and their 95% pointwise confidence intervals cover zero almost all the time. This indicates that these two selected predictors are not related to Y through the conditional mean function $E(Y \mid X_j)$. Therefore, we believe these two predictors do not have marginal effects on the mean function.

Next we marginally regress the squared residuals over these two predictors through local polynomial approximations, and the grid points are the same as we used in the mean regression functions. The plots (II.1) and (II.2) in Figure 5.1 show the estimated variance functions along with their 95% pointwise confidence intervals. It is clear that both estimated variance functions resemble certain patterns. For example, we can see that the variance function of Y is an increasing function with respect to the probe 1388231_at. This can also be visualized in the plot (I.2) of Figure 5.1 because Y resembles a triangle shape in the scatter plot. We believe both probes may only have effects on the conditional variance function of Y; that is, the two predictors represent the heteroscedasticity. However, these two probes are easily missed by some screening approaches like the SIS and NIS. For example, the SIS ranks the probe 1379142_at at the 1606-th position and the

probe 1388231_at at the 1715-th position.

(I.1) $E(Y \mid X_{(1)})$ (I.2) $E(Y \mid X_{(2)})$



(II.1) $\mathrm{var}(Y \mid X_{(1)})$ (II.2) $\mathrm{var}(Y \mid X_{(2)})$

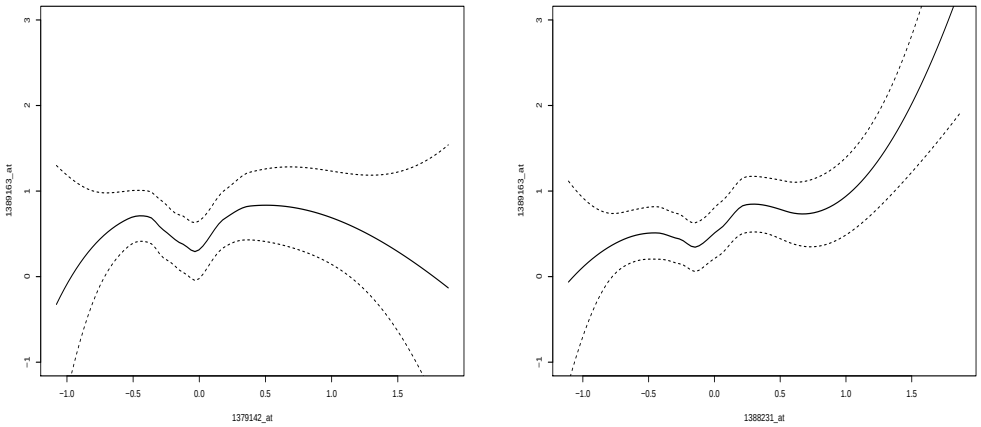


Figure 5.1. *Estimated mean functions and variance functions with 95% pointwise confidence intervals.*

5.4 Theoretical Proofs

5.4.1 Proof of Theorem 5.1

Let $\mathbf{\Sigma}_{\mathcal{D}} = \text{cov}(\mathbf{x}_{\mathcal{D}}, \mathbf{x}_{\mathcal{D}}^{\text{T}})$. We first note that the linearity condition (5.10) entails that

$$E(X_j \mid \mathbf{x}_{\mathcal{D}}) = \text{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\text{T}})\mathbf{\Sigma}_{\mathcal{D}}^{-1}\mathbf{x}_{\mathcal{D}},$$

and the constant variance condition (5.11) implies that

$$\text{var}(X_j \mid \mathbf{x}_{\mathcal{D}}) = 1 - \text{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\text{T}})\mathbf{\Sigma}_{\mathcal{D}}^{-1}\text{cov}(\mathbf{x}_{\mathcal{D}}, X_j).$$

In addition, (5.8) implies that X_j is independent of $\mathbf{y}$ when $\mathbf{x}_{\mathcal{D}}$ is given. Thus

$$\begin{aligned} E(X_j^2 \mid \mathbf{y}) &= E\left\{E(X_j^2 \mid \mathbf{x}_{\mathcal{D}}, \mathbf{y}) \mid \mathbf{y}\right\} = E\left\{E(X_j^2 \mid \mathbf{x}_{\mathcal{D}}) \mid \mathbf{y}\right\} \\ &= E\left\{\text{var}(X_j \mid \mathbf{x}_{\mathcal{D}}) + E^2(X_j \mid \mathbf{x}_{\mathcal{D}}) \mid \mathbf{y}\right\} \qquad (5.18) \\ &= 1 + \text{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\text{T}})\mathbf{\Sigma}_{\mathcal{D}}^{-1}\left\{E(\mathbf{x}_{\mathcal{D}}\mathbf{x}_{\mathcal{D}}^{\text{T}} \mid \mathbf{y}) - \mathbf{\Sigma}_{\mathcal{D}}\right\}\mathbf{\Sigma}_{\mathcal{D}}^{-1}\text{cov}(\mathbf{x}_{\mathcal{D}}, X_j). \end{aligned}$$

Let $\mathbf{\Lambda}_{\mathcal{D}} = -E\left\{(\mathbf{x}_{\mathcal{D}}\mathbf{x}_{\mathcal{D}}^{\text{T}} - \mathbf{\Sigma}_{\mathcal{D}})\left(\widetilde{\mathbf{x}}_{\mathcal{D}}\widetilde{\mathbf{x}}_{\mathcal{D}}^{\text{T}} - \mathbf{\Sigma}_{\mathcal{D}}\right)^{\text{T}}\|\mathbf{y} - \widetilde{\mathbf{y}}\|\right\}$. Since $(X_j, \mathbf{y}) \perp\!\!\!\perp (\widetilde{X}_j, \widetilde{\mathbf{y}})$, it follows that $X_j \perp\!\!\!\perp \widetilde{X}_j \mid (\mathbf{y}, \widetilde{\mathbf{y}})$, $\widetilde{X}_j \perp\!\!\!\perp \mathbf{y} \mid \widetilde{\mathbf{y}}$ and $X_j \perp\!\!\!\perp \widetilde{\mathbf{y}} \mid \mathbf{y}$. These three results, together with (5.18), yield that

$$\begin{aligned} \lambda_j &= -E\left[E\left\{(X_j^2 - 1)(\widetilde{X}_j^2 - 1) \mid \mathbf{y}, \widetilde{\mathbf{y}}\right\}\|\mathbf{y} - \widetilde{\mathbf{y}}\|\right] \\ &= -E\left\{E\left(X_j^2 - 1 \mid \mathbf{y}\right)E(\widetilde{X}_j^2 - 1 \mid \widetilde{\mathbf{y}})\|\mathbf{y} - \widetilde{\mathbf{y}}\|\right\} \\ &= E\Big[\left\{\text{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\text{T}})\mathbf{\Sigma}_{\mathcal{D}}^{-1}\left(\mathbf{x}_{\mathcal{D}}\mathbf{x}_{\mathcal{D}}^{\text{T}} - \mathbf{\Sigma}_{\mathcal{D}}\right)\mathbf{\Sigma}_{\mathcal{D}}^{-1}\text{cov}(\mathbf{x}_{\mathcal{D}}, X_j)\right\} \\ &\quad \left\{\text{cov}(X_j, \mathbf{x}_{\mathcal{D}}^{\text{T}})\mathbf{\Sigma}_{\mathcal{D}}^{-1}\left(\widetilde{\mathbf{x}}_{\mathcal{D}}\widetilde{\mathbf{x}}_{\mathcal{D}}^{\text{T}} - \mathbf{\Sigma}_{\mathcal{D}}\right)\mathbf{\Sigma}_{\mathcal{D}}^{-1}\text{cov}(\mathbf{x}_{\mathcal{D}}, X_j)\right\}\|\mathbf{y} - \widetilde{\mathbf{y}}\|\Big]. \qquad (5.19) \end{aligned}$$

Therefore, it follows that

$$\max_{j\in\mathcal{I}} \lambda_j \leq \lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\boldsymbol{\Lambda}_{\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\right) \max_{j\in\mathcal{I}}\left\{\|\mathrm{cov}(X_j,\mathbf{x}_{\mathcal{D}})\|^4\right\} \lambda_{\max}(\boldsymbol{\Sigma}_{\mathcal{D}}^{-2}).$$

Use the fact that $\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{B}\mathbf{C}) \leq \lambda_{\max}(\mathbf{B})\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{C})$ for any matrix $\mathbf{B} \geq 0$. We obtain $\lambda_{\max}\left(\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\boldsymbol{\Lambda}_{\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D}}^{-1}\right) \leq \lambda_{\max}\left(\boldsymbol{\Lambda}\right)/\lambda_{\min}^2\left(\boldsymbol{\Sigma}_{\mathcal{D}}\right)$. In addition,

$$\max_{j\in\mathcal{I}}\left\{\|\mathrm{cov}(X_j,\mathbf{x}_{\mathcal{D}})\|^4\right\} \leq \lambda_{\max}^2\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{I}},\mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{D}},\mathbf{x}_{\mathcal{I}})\right\}.$$

These two results imply that

$$\max_{j\in\mathcal{I}} \lambda_j \leq \lambda_{\max}\left(\boldsymbol{\Lambda}_{\mathcal{D}}\right)\lambda_{\max}^2\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{I}},\mathbf{x}_{\mathcal{D}}^{\mathrm{T}})\mathrm{cov}(\mathbf{x}_{\mathcal{D}},\mathbf{x}_{\mathcal{I}})\right\}/\lambda_{\min}^4\left(\boldsymbol{\Sigma}_{\mathcal{D}}\right),$$

which completes the proof of (5.12). □

5.4.2 Proof of Theorem 5.2

We reproduce the Hoeffding's inequality (Hoeffding, 1963) for the sake of readability.

Lemma 5.1. *Let $X_1,\ldots,X_n$ be independent random variables. Assume that the X_i's are almost surely bounded; that is,* $\Pr\left(a_i \leq X_i \leq b_i\right) = 1$ *for* $1 \leq i \leq n$. *Then*

$$\Pr\left\{\sum_{i=1}^{n}\left(X_i - E\left(X_i\right)\right) \geq \tau\right\} \leq \exp\left(-\frac{2\tau^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right).$$

For notational clarity, we let $I_n(X_j,\mathbf{y}) = \mathbf{1}\left\{(X_j,\mathbf{y}) \in \Omega_n\left(X_j,\mathbf{y}\right)\right\}$ *and*

$$\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right) = X_0X_{ij}\|\mathbf{y}_i-\mathbf{y}_0\|I_n(X_{ij},\mathbf{y}_i) - E\left\{X_0X_j\|\mathbf{y}-\mathbf{y}_0\|I_n(X_j,\mathbf{y})\right\}.$$

Lemma 5.2. *If* $E\left(X_j^2\|\mathbf{y}\|^2\right) < \infty$ *and* $E\left(X_j^2\right) < \infty$, *then*

$$\begin{aligned} & \Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right| \geq \varepsilon_n\right\} \\ \leq \;& 2q^{1/2}\left(32K_n^{\star}K_n^2\varepsilon_n^{-1}\right)^{q+1}\exp\left(-\frac{n\varepsilon_n^2}{512q^{1/2}K_n^{\star}K_n^2}\right). \end{aligned} \tag{5.20}$$

Proof of Lemma 5.2. For a fixed point $(X_0, \mathbf{y}_0) \in \Omega_n(X_0, \mathbf{y}_0)$, it can be easily verified that

$$E\left\{\xi_{X_0,\mathbf{y}_0}^2\left(X_j,\mathbf{y}\right)\right\} \leq X_0^2 E\left\{X_j^2\|\mathbf{y}-\mathbf{y}_0\|^2\right\} \leq 2X_0^2 E\left\{X_j^2\left(\|\mathbf{y}\|^2+\|\mathbf{y}_0\|^2\right)\right\}.$$

Consequently, we apply the Chebyshev's inequality to obtain that

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right| > \frac{1}{2}\varepsilon_n\right\} \leq \frac{1}{2}, \tag{5.21}$$

when n is sufficiently large, where $\varepsilon_n > 0$ depends possibly on n.

Let $\left(X'_{ij},\mathbf{y}'_i\right)$'s be independent copies of $(X_{ij},\mathbf{y}_i)$'s for $1 \leq i \leq n$, and σ_i's be the Rademacher sequence (i.e., an independent and identically distributed sequence taking value ± 1 with probability $1/2$). By symmetry,

$$\sigma_i\left\{\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)-\xi_{X_0,\mathbf{y}_0}\left(X'_{ij},\mathbf{y}'_i\right)\right\}$$

has the same distribution as $\left\{\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)-\xi_{X_0,\mathbf{y}_0}\left(X'_{ij},\mathbf{y}'_i\right)\right\}$. In addition, the symmetrization Lemma in Pollard (1984) implies that

$$\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right| \geq \varepsilon_n\right\}$$

$$
\begin{aligned}
&\leq 2\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\left\{\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)-\xi_{X_0,\mathbf{y}_0}\left(X'_{ij},\mathbf{y}'_i\right)\right\}\right|\geq\frac{1}{2}\varepsilon_n\right\}\\
&= 2\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\sigma_i\left\{\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)-\xi_{X_0,\mathbf{y}_0}\left(X'_{ij},\mathbf{y}'_i\right)\right\}\right|\geq\frac{1}{2}\varepsilon_n\right\}\\
&\leq 4\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\sigma_i\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right|\geq\frac{1}{4}\varepsilon_n\right\}.
\end{aligned}
$$

Let $\mathcal{F}_n=\{\xi_{X_0,\mathbf{y}_0}(X_{ij},\mathbf{y}_i):(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)\}$ be a class of functions indexed by $(X_0,\mathbf{y}_0)$. Given the observations $\mathcal{X}\otimes\mathcal{Y}=\{(X_{ij},\mathbf{y}_i),1\leq i\leq n\}$, we delicately choose functions $\xi_1^0,\ldots,\xi_m^0$ from $\mathcal{F}_n$ such that

$$
\min_{1\leq s\leq m}\frac{1}{n}\sum_{i=1}^{n}\left|\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)-\xi_j^0\left(X_{ij},\mathbf{y}_i\right)\right|\leq\frac{1}{8}\varepsilon_n \tag{5.22}
$$

for each $\xi_{X_0,\mathbf{y}_0}$ in $\mathcal{F}_n$. We will discuss how to decide m later on.

Let P_n be the empirical measure that puts equal mass n^{-1} at each of the n observations $(X_{ij},\mathbf{y}_i)$, and $N(\varepsilon_n,P_n,\mathcal{F}_n)$ be the minimum m for all sets that satisfies (5.22). Denote by $\xi^*_{X_0,\mathbf{y}_0}$ for the ξ_s^0 at which the minimum is achieve, we then have

$$
\begin{aligned}
&\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\sigma_i\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right|\geq\frac{1}{4}\varepsilon_n\Big|\mathcal{X}\otimes\mathcal{Y}\right\}\\
\leq\ &\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\sigma_i\xi^*_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right|\geq\frac{1}{8}\varepsilon_n\Big|\mathcal{X}\otimes\mathcal{Y}\right\}\\
\leq\ &N(\varepsilon_n,P_n,\mathcal{F}_n)\max_{1\leq s\leq N(\varepsilon_n,P_n,\mathcal{F}_n)}\Pr\left\{\left|n^{-1}\sum_{i=1}^{n}\sigma_i\xi_s^0\left(X_{ij},\mathbf{y}_i\right)\right|\geq\frac{1}{8}\varepsilon_n\Big|\mathcal{X}\otimes\mathcal{Y}\right\}.
\end{aligned}
$$

Now we need to determine the order of $N(\varepsilon_n,P_n,\mathcal{F}_n)$. For each set satisfying

(5.22), each $\xi_s^0(X_{ij}, \mathbf{y}_i)$ has a pair $(X_{s0}, \mathbf{y}_{s0})$ such that

$$\xi_s^0(X_{ij}, \mathbf{y}_i) = \xi_{X_{s0}, \mathbf{y}_{s0}}(X_{ij}, \mathbf{y}_i).$$

Then for each $(X_0, \mathbf{y}_0) \in \Omega_n(X_0, \mathbf{y}_0)$, we can have that

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n} |\xi_{X_0,\mathbf{y}_0}(X_{ij}, \mathbf{y}_i) - \xi_{X_{s0},\mathbf{y}_{s0}}(X_{ij}, \mathbf{y}_i)| \\
\leq\ & n^{-1}\sum_{i=1}^{n} |X_0 - X_{s0}|\,|X_{ij}|\,\|\mathbf{y}_i - \mathbf{y}_0\| I_n(X_{ij}, \mathbf{y}_i) + n^{-1}\sum_{i=1}^{n} |X_{s0}X_{ij}|\,\|\mathbf{y}_0 - \mathbf{y}_{s0}\| I_n(X_{ij}, \mathbf{y}_i) \\
+\ & E\{|X_0 - X_{s0}|\,|X_{ij}|\,\|\mathbf{y}_i - \mathbf{y}_0\| I_n(X_{ij}, \mathbf{y}_i)\} + E\{|X_{s0}X_{ij}|\,\|\mathbf{y}_0 - \mathbf{y}_{s0}\| I_n(X_{ij}, \mathbf{y}_i)\} \\
\leq\ & 2\left\{K_n q^{1/2}(2K_n^{\star})\,|X_0 - X_{s0}| + K_n^2\|\mathbf{y}_0 - \mathbf{y}_{s0}\|\right\}.
\end{aligned}$$

Next we want to bound the right-hand side of the above formula by $\varepsilon_n/8$. For each $(X_0, \mathbf{y}_0) \in \Omega_n(X_0, \mathbf{y}_0)$, we need a pair $(X_{s0}, \mathbf{y}_{s0}) \in \Omega_n(X_{s0}, \mathbf{y}_{s0})$, such that $|X_0 - X_{s0}| \leq K_n^{-1} q^{-1/2}(2K_n^{\star})^{-1}\varepsilon_n/16$ and $\|\mathbf{y}_0 - \mathbf{y}_{s0}\| \leq K_n^{-2}\varepsilon_n/16$. Therefore, the covering number $N(\varepsilon_n, P_n, \mathcal{F}_n)$ needed to satisfied (5.22) is bounded by

$$\{64q^{1/2}K_n^{\star}K_n^2\varepsilon_n^{-1}\}\left(32K_n^{\star}K_n^2\varepsilon_n^{-1}\right)^q = O\{\left(K_n^2K_n^{\star}\varepsilon_n^{-1}\right)^{q+1}\}.$$

That is,

$$N(\varepsilon_n, P_n, \mathcal{F}_n) \leq O\{\left(K_n^2K_n^{\star}\varepsilon_n^{-1}\right)^{q+1}\}.$$

Given the observations $\mathcal{X} \otimes \mathcal{Y}$, the function $\sigma_i\xi_s^0(X_{ij}, \mathbf{y}_i)$ is bounded. With the Hoeffding's inequality, we have

$$\Pr\left\{\left|n^{-1}\sum_{i=1}^{n}\sigma_i\xi_s^0(X_{ij}, \mathbf{y}_i)\right| \geq \frac{1}{8}\varepsilon_n \middle| \mathcal{X} \otimes \mathcal{Y}\right\} \leq 2\exp\left(-\frac{2n^2(\varepsilon_n/8)^2}{\sum_{i=1}^{n} 4\xi_s^0(X_{ij}, \mathbf{y}_i)}\right) \wedge 1.$$

Combining the above results, we can easily have

$$\begin{aligned}
&\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right|\geq\varepsilon_n\right\}\\
\leq\;& O\left\{\left(K_n^2K_n^\star/\varepsilon_n\right)^{q+1}\right\}E\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}2\exp\left(-\frac{2n^2(\varepsilon_n/8)^2}{\sum_{i=1}^{n}4\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)}\right)\wedge 1\right\}.
\end{aligned}$$

Because $\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\leq 4q^{1/2}K_n^\star K_n^2$, Lemma 5.2 follows immediately. □

Let

$$\eta\left(X_{ij},\mathbf{y}_i\right)=E\left[X_{ij}\widetilde{X}_{ij}\|\mathbf{y}_i-\widetilde{\mathbf{y}}_i\|\mathbf{1}\{|X_{ij}|\vee\|\widetilde{X}_{ij}\|\leq K_n,\|\mathbf{y}_i\|_\infty\vee\|\widetilde{\mathbf{y}}_i\|_\infty\leq K_n^\star\}\middle|X_{ij},\mathbf{y}_i\right],$$

where $a\vee b=\max\{a,b\}$. Clearly, $|\eta\left(X_{ij},\mathbf{y}_i\right)|\leq|X_{ij}E(\widetilde{X}_{ij}\|\mathbf{y}_i-\widetilde{\mathbf{y}}_i\|\mid X_{ij},\mathbf{y}_i)|\leq 2\left\{qE(X_j^2)\right\}^{1/2}(K_nK_n^\star)$. Thus a direct application of the Hoeffding's inequality (Lemma 5.1) entails the following result.

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\eta\left(X_{ij},\mathbf{y}_i\right)-E\left\{\eta\left(X_j,\mathbf{y}\right)\right\}\right|\geq\varepsilon_n\right\}\leq 2\exp\left(-\frac{n\varepsilon_n^2}{8qE(X_j^2)\{K_nK_n^\star\}^2}\right).\tag{5.23}$$

These lemmas pave the road for proving Theorem 5.2, which is detailed below.

Proof of Theorem 5.2. For the sake of notational clarity, we write $\varepsilon_n=cn^{-\kappa}$ and X_{ij} in place of X_{ij}^2-1 in the sequel. With the Bonferroni inequality, the probability that $\left|\widehat{\lambda}_j-\lambda_j\right|\geq\varepsilon_n$ is clearly bounded by

$$\begin{aligned}
&\Pr\left(\left|\sum_{i,s=1}^{n}\frac{X_{ij}S=X_{sk}\|\mathbf{y}_i-\mathbf{y}_s\|\mathbf{1}\{(X_{ij},\mathbf{y}_i)\in\Omega_n\left(X_{ij},\mathbf{y}_i\right)\text{ and }(S=X_{sk},\mathbf{y}_s)\in\Omega_n(S=X_{sk},\mathbf{y}_s)\}}{n(n-1)}\right.\right.\\
&\qquad\left.\left.-E\left[X_j\widetilde{X}_j\|\mathbf{y}-\widetilde{\mathbf{y}}\|\mathbf{1}\{(X_j,\mathbf{y})\in\Omega_n\left(X_j,\mathbf{y}\right)\text{ and }(\widetilde{X}_j,\widetilde{\mathbf{y}})\in\Omega_n(\widetilde{X}_j,\widetilde{\mathbf{y}})\}\right]\right|\geq\frac{\varepsilon_n}{2}\right)\\
&+\Pr\left(\left|\sum_{i,s=1}^{n}\frac{X_{ij}S=X_{sk}\|\mathbf{y}_i-\mathbf{y}_s\|\mathbf{1}\{(X_{ij},\mathbf{y}_i)\in\Omega_n^c\left(X_{ij},\mathbf{y}_i\right)\text{ or }(S=X_{sk},\mathbf{y}_s)\in\Omega_n^c(S=X_{sk},\mathbf{y}_s)\}}{n(n-1)}\right.\right.
\end{aligned}$$

$$- E\left[X_j\widetilde{X}_j\|\mathbf{y}-\widetilde{\mathbf{y}}\|\mathbf{1}\{(X_j,\mathbf{y})\in\Omega_n^c\left(X_j,\mathbf{y}\right)\text{ or }(\widetilde{X}_j,\widetilde{\mathbf{y}})\in\Omega_n^c(\widetilde{X}_j,\widetilde{\mathbf{y}})\}\right]\Bigg|\geq\frac{\varepsilon_n}{2}\Bigg). \quad (5.24)$$

Condition (B.2) implies immediately that $\Pr\left\{(X_j,\mathbf{y})\in\Omega_n^c\left(X_j,\mathbf{y}\right)\right\}=o(n^{-\kappa})$, thus the second quantity in the right hand side of (5.24) is bounded by

$$n\Pr\left\{(X_j,\mathbf{y})\in\Omega_n^c\left(X_j,\mathbf{y}\right)\right\}.$$

To prove (5.13), it remains to study the convergence rate of the first quantity in the right hand side of (5.24). Again with the Bonferroni inequality, Lemma 5.2 and (5.23), the first quantity is less than or equal to

$$\begin{aligned}
& n\Pr\left\{\sup_{(X_0,\mathbf{y}_0)\in\Omega_n(X_0,\mathbf{y}_0)}\left|n^{-1}\sum_{i=1}^{n}\xi_{X_0,\mathbf{y}_0}\left(X_{ij},\mathbf{y}_i\right)\right|\geq\frac{\varepsilon_n}{4}\right\} \\
+ \quad & \Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\eta\left(X_{ij},\mathbf{y}_i\right)-E\left\{\eta\left(X_j,\mathbf{y}\right)\right\}\right|\geq\frac{\varepsilon_n}{4}\right\} \\
\leq \quad & 2\left\{q^{1/2}\left(128K_n^{\star}K_n^2\varepsilon_n^{-1}\right)^{q+1}n\exp\left(-\frac{n\varepsilon_n^2}{8192q^{1/2}K_n^{\star}K_n^2}\right)+\exp\left(-\frac{n\varepsilon_n^2}{128qE(X_j^2)\{K_nK_n^{\star}\}^2}\right)\right\}.
\end{aligned}$$

Because we choose $K_n^{\star}$ to be sufficiently large, we can use only $K_n^{\star}>1$, the first quantity in the curly parenthesis is clearly bounded by the second quantity. Thus the proof of (5.13) is completed if $n^{1-2\kappa}/\left(K_n^2K_n^{\star 2}\right)\to\infty$.

Next we show the ranking consistency property. Let $\delta=\min_{j\in\mathcal{D}}\lambda_j-\max_{j\in\mathcal{I}}\lambda_j$. Theorem 5.1 ensures that $\delta>0$. After simple algebraic calculation,

$$\Pr\left(\min_{j\in\mathcal{D}}\widehat{\lambda}_j\leq\max_{j\in\mathcal{I}}\widehat{\lambda}_j\right)=\Pr\left(\min_{j\in\mathcal{D}}\widehat{\lambda}_j-\min_{j\in\mathcal{D}}\lambda_j+\delta\leq\max_{j\in\mathcal{I}}\widehat{\lambda}_j-\max_{j\in\mathcal{I}}\lambda_j\right)$$

$$
\begin{aligned}
&= \Pr\left(\left|\min_{j\in\mathcal{D}}\widehat{\lambda}_j - \min_{j\in\mathcal{D}}\lambda_j\right| \geq \delta/2 \text{ or } \left|\max_{j\in\mathcal{I}}\widehat{\lambda}_j - \max_{j\in\mathcal{I}}\lambda_j\right| \geq \delta/2\right) \\
&\leq \Pr\left(\left|\min_{j\in\mathcal{D}}\widehat{\lambda}_j - \min_{j\in\mathcal{D}}\lambda_j\right| \geq \delta/2\right) + \Pr\left(\left|\max_{j\in\mathcal{I}}\widehat{\lambda}_j - \max_{j\in\mathcal{I}}\lambda_j\right| \geq \delta/2\right) \\
&\leq \Pr\left(\sup_{j\in\mathcal{D}}|\widehat{\lambda}_j - \lambda_j| \geq \delta/2\right) + \Pr\left(\sup_{j\in\mathcal{I}}|\widehat{\lambda}_j - \lambda_j| \geq \delta/2\right).
\end{aligned}
$$

Both terms converge to 0 in probability by (5.13), which entails (5.14) immediately.

The proof of (5.15) follows literally the proof for the second part of Theorem 4 in Fan and Song (2009). Let $\mathcal{E}_n = \left\{\max_{j\in\mathcal{D}}\left|\widehat{\lambda}_j - \lambda_j\right| \leq cn^{-\kappa}\right\}$. On the event $\mathcal{E}_n$ and under condition (B.3) $\lambda_j \geq 2cn^{-\kappa}$, we have $\widehat{\lambda}_j \geq cn^{-\kappa}$, for all $j \in \mathcal{D}$. Therefore,

$$
\begin{aligned}
\Pr(\mathcal{D} \subseteq \widehat{\mathcal{D}}) &\geq \Pr(\mathcal{E}_n) = 1 - \Pr(\mathcal{E}_n^c) = 1 - \Pr\left\{\min_{j\in\mathcal{D}}\left|\widehat{\lambda}_j - \lambda_j\right| \geq cn^{-\kappa}\right\} \\
&= 1 - s_n \Pr\left\{\left|\widehat{\lambda}_j - \lambda_j\right| \geq cn^{-\kappa}\right\} \geq 1 - O\left\{s_n \exp\left(-n^{1-2\kappa}/\xi_c\right)\right\},
\end{aligned}
$$

for some constant ξ_c, where s_n is the cardinality of $\mathcal{D}$. This proves (5.15). □

Chapter 6

Discussion and Future Research

6.1 Model Misspecification

We theoretically compare covariate-adjusted and -unadjusted approaches for generalized linear models. Simulation studies confirm these theoretical results and suggest the results are applicable to cases with moderate sample size. In general, a covariate-adjusted model allows correct interpretations of the treatment effect, while an unadjusted model can lead to misleading results. On the other hand, a covariate-adjusted model is associated with reduced precision. When the covariate effects are small, then an unadjusted approach might still be preferable in some cases. Thus, there is a trade-off between accuracy and precision. Researchers should choose suitable approaches to achieve the goal, keeping in mind their advantages and disadvantages.

When the treatment interacts with covariates, β^* is usually biased when the covariate-adjusted model is true. The interaction leads to two different scenarios: 1) treatment A is always superior to treatment B, while the magnitude of difference between A and B is different in different subgroups; 2) treatment A is superior to

B in on some subgroups and is inferior to treatment B in other subgroups. In scenario 2, although both β^* and β both represent the "average" treatment effect, they may not the best way to define the treatment effect. In this case, treatment effects may be better quantified in subgroups. Consequently, the results in Section 3.2 may not be very relevant in scenario 2.

We only compare adjusted and unadjusted models. At this point, we are not able to provide general results to compare all possible covariate-adjusted models. Nevertheless, we believe all comparisons involve trade-off between accuracy and precision. In our opinion, a model selection procedure is a good way to balance accuracy and precision, although its practical implementation is itself an interesting and promising research area with many unsolved issues. The results form a basis to evaluate a covariate-adjusted approach in terms of bias and precision. However, we acknowledge that further research is needed.

We compare covariate-adjusted and -unadjusted generalized linear models with the same dispersion parameter. Extension to cases with different dispersion parameters are of interest for future research.

6.2 Feature Ranking and Screening

We proposed two unified sure independence ranking and screening procedures for ultrahigh dimensional data. We established their ranking consistency property and sure screening properties. Our numerical comparison indicates that for some settings, the proposed procedures outperform the sure independence screening procedure (Fan and Lv, 2008) and nonparametric independence screening procedure (Fan, Feng and Song, 2011). The proposed procedures are appealing in situations in which there are a huge amount of candidate variables but little information

about the underlying true model.

We want to emphasize that, these two new approaches are complementary, rather than alternative, methodologies to the existing sure independence ranking and screening literature. For example, our simulations indicate that, when the important predictors are not symmetrically relevant to the response variable, the existing independence ranking and screening procedures are usually more efficient. This phenomenon also motivates us to develop an efficient marginal utility which can identify all important predictors exhaustively under the framework (4.2) and (5.8). This is a challenging and yet important problem, which deserves our further study.

After feature screening, the next step is to apply refined variable selection techniques to fully identify important variables, such as the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001), Dantzig selector (Candes and Tao, 2007), the penalized linear unbiased selector (Zhang, 2010, PLUS) and their variations such as the elastic net (Zou and Hastie, 2005), the adaptive LASSO (Zou, 2006) and nonnegative garrote (Yuan and Lin, 2007), etc. In this dissertation, we only emphasize the features of variable screening without further discussing the impact of variable selection. There are a lot of open topics for future research to consider the two steps together; for example, how to incorporate the two steps well enough so that we can reduce the false selection rate.

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the likelihood principle. Proceedings of the Second International Symposium of Information Theory, ed. B. N. Petrov and F. Csáki. Budapest: Akadémiai Kiado.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control* **19**, 716-723.

[3] Behrendt, C. E. and Gehan, E. A. (2009). Treatment-subgroup interaction: An example from a published, phase II clinical trial. *Contemporary Clinical Trials*, **30**, 279-281.

[4] Bickel, P. J. and Levina E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227.

[5] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics* **35**, 2313-2404.

[6] Candes, E. Wakin, M. and Boyd, S. (2007). Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications* **14**, 877-905.

[7] Choi, S. and Wette, R. (1969). Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias. *Technometrics* **11**, 683-690.

[8] Cox, D. R. (1958). Planning of Experiments. New York: Wiley.

[9] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). Analysis of longitudinal data (2nd ed.). Oxford, U. K.: Oxford University.

[10] Drake, C. and McQuarrie, A. (1995). A note on the bias due to omitted confounders. *Biometrika* **82**, 633-638.

[11] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annual of Statistics* **32**, 407-499.

[12] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association* **106**, 544–557

[13] Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**, 710-723.

[14] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B.* **70**, 849–911.

[15] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research* **10**, 1829-1853.

[16] Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38**, 3567–3604

[17] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309-368.

[18] Fisher, R. A. (1925). Theory of Statistical Estimation. *Proceedings of the Cambridge Philosophical Society* **22**, 700-725.

[19] Ford, I. and Norrie, J. (2002). The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Statistics in medicine* **21**, 2899-2908.

[20] Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

[21] Friedman, J. H. (2008). Fast sparse regression and classification. In Proceedings of the 23rd International Workshop on Statistical Modelling, 27-57.

[22] Gail, M. and Simon, R. (1985). Testing for qualitative interaction between treatment effects and patient subsets. *Biometrics* **41**, 361-372.

[23] Gail, M., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with non-linear regressions and omitted covariates. *Biometrika* **71**, 431-44.

[24] Grouin, J. M., Coste, M. and Lewis, J. (2005). Subgroup Analyses in Randomized Clinical Trials: Statistical and Regulatory Issues. *Journal of Biopharmaceutical Statistics* **15**, 869-882.

[25] Grouin, J. M., Day, S. and Lewis, J. (2004). Adjustment for baseline covariates: an introductory note. *Statistics in medicine* **23**, 697-699.

[26] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models. *J. R. Statist. Soc. B* **55**, 757-796.

[27] Hauck, W. W., Anderson, S. and Marcus, S. M. (1998). Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials? *Controlled Clinical Trials* **19**, 249 - 256.

[28] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**, 13–30.

[29] Hoerl, A. E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

[30] Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Annals of Statistics* **33**, 1617-1642.

[31] Kim, Y., Choi, H. and Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association* **103**, 1665–1673.

[32] Knol, M. J., van der Tweel, I., Grobbee, D. E., Numans, M. E. and Geerlings, M. I. (2007). Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *Int J Epidemiol* **36**, 1111-1118.

[33] Koch, G. G., Tangen, C. M., Jung, J. W. and Amara, I. A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* **17**, 1863-1892.

[34] Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79-86.

[35] Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

[36] LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *University of California Publications in Statistics* **1**, 277-330.

[37] Lee, M. L. T. and Whitmore, G. A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statist. Sci.* **21**, 501-513.

[38] Lin, X., and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96**, 1045-1056.

[39] Lin, D. Y., and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association* **96**, 103-126.

[40] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annual of Statistics* **37**, 3498-3528.

[41] Martinussen, T. and Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89**, 283-298.

[42] McClelland, G. H. and Judd, C. M. (1993). Statistical Difficulties of Detecting Interactions and Moderator Effects. *Psychological Bulletin* **114**, 376-390.

[43] McCullagh, P., Nelder, J. (1989). Generalized linear models, second edition. Chapman and Hall.

[44] Nelder, J., Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* **135**, 370-384.

[45] Neuhuas, J. M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *Journal of American Statistical Association* **93**, 1124-1129.

[46] Neuhaus, J. M. and Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* **80**, 807-815.

[47] Neuhaus, J. M., Kalbfleisch, J. D., Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25-35.

[48] Pennell, M. L., Whitmore, G. A. and Lee, M. T. (2010). Bayesian random-effects threshold regression with application to survival data with nonproportional hazards. *Biostatistics* **1**, 111-126.

[49] Peto, R. (1982). Statistical aspects of cancer trials. *In Treatment of Cancer*, K. E. Halnan (ed.), 867-871. London: Chapman and Hall.

[50] Pocock, S. J. (2004). Clinical trials: a practical approach. John Wiley.

[51] Pocock, S. J., Assmann, S. E., Enos, L. E. and Kasten, L. E. (2002). Subgroup analysis, covariate adjusted and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in medicine* **21** 2917 - 2930.

[52] Pollard, D. (1984) *Convergence of Stochastic Processes*, Springer, New York.

[53] Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69**, 331-342.

[54] Richardson, D. B. and Kaufman, J. S. (2009). Estimation of the Relative Excess Risk Due to Interaction and Associated Confidence Bounds. *American Journal of Epidemiology* **169**, 756-760.

[55] Robinson, L. D., Dorroh, J. R., Lien, D. and Tiku, M. L. (1998). The effects of covariate adjusted in generalized linear models. *Communications in Statistics - Theory and Methods* **27**,1653-1675.

[56] Robinson, L. D. and Jewell, N. P. (1991). Some Surprising Results About Covariate Adjusted in Logistic Regression Models. *International Statistical Review* **58**, 227-240.

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

[58] Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp1, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheeld, V. C., and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* **103**, 14429-14434.

[59] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **19**, 461-464.

[60] Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons Inc.

[61] Senn, S. (2004). Conditional and Marginal Models: Another View. *Statistical Science* **19**, 228-230.

[62] Senn, S. (2007). *Statistical issues in Drug development*, John Wiley and Sons.

[63] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proc. Natn. Acad. Sci. USA* **100**, 9440-9445.

[64] Stürmer, T. and Brenner, H. (2002). Flexible Matching Strategies to Increase Power and Efficiency to Detect and Estimate Gene-Environment Interactions in Case-Control Studies. *American Journal of Epidemiology* **155**, 593-602.

[65] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

[66] Tibshirani, R. (1996). Regression shrinkage and selection via LASSO. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

[67] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82-86.

[68] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **60**, 595-603.

[69] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104**, 1512–1524.

[70] Weng, H. Y., Hsueh, Y. H. and Messam, L. L. and Hertz-Picciotto, I. (2009). Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. *Am J Epidemiol* **169**, 1182-1190.

[71] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

[72] Yuan, M. and Lin, Y. (2007). On the nonnegative garrote estimator. *Journal of the Royal Statistical Society, Series B.* **69**, 143–161.

[73] Zeger, S. L., Liang, L. Y., Albert, P. A. (1998). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-60.

[74] Zou, G. Y. (2008). On the estimation of additive interaction by use of the four-by-two table and beyond. *Am J Epidemiol* **168**, 212-224.

[75] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

[76] Zhu, L. P, Li, L., Li, R. and Zhu, L. X. (2010). Model-free feature screening for ultrahigh dimensional data, *Submitted*.

[77] Zhu, L. P, Zhu, L. X. and Feng, Z. H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association* **105**, 1455-1466.

[78] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

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

[80] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics* **36**, 1509–1566.

Vita

Junyi Lin
Department of Statistics, Penn State University
326 Thomas Building, University Park, PA 16802
PHONE: (814) 321-1029 EMAIL: jul216@psu.edu

Education

Ph.D. in Statistics, The Pennsylvania State University, USA, 2011 (Expected)
M.S. in Statistics, The Pennsylvania State University, USA, 2011
M.S. in Mathematics, Tsinghua University, China, 2007
B.S. in Mathematics, Tsinghua University, China, 2005

Professional Experience

Research Assistant	01/2010 to present
Advisor: Prof. Runze Li, Department of Statistics, PSU	
Quantitative Analyst (Intern)	05/2010-08/2010
Manager: Agus Sudjianto, Bank of America, Charlotte, NC	
Research Assistant (Intern)	05/2009-08/2009
Advisor: Dr. Lei Nie, FDA, Silver Spring, MD	

Publications

Li, R., **Lin, J**. and Zhu, L. (2011). A Unified Sure Independence Ranking and Screening Procedure for Ultrahigh Dimensional Data. Journal of the Royal Statistical Society, Series B, submitted.

Shiyko, M., **Lin, J**., Li, R., Burkhalter, J. and Ostroff, J. (2011). Joint Modeling of Program-Satisfaction Trajectories and Time-To-Non-Adherence with a Handheld Device in a Smoking Cessation Study. *Multivariate Behavioral Research*, submitted.

Lin, J., Sudjianto, A. and Singhal, H. (Under revision). Non-proportional Hazard Approaches: First Hitting Time Models and Frailty Models Revisited.

Lin, J., Nie, L. and Li, R. (2011). Variance-bias trade-off in generalized linear regression models. *The Canadian Journal of Statistics,* submitted.

Lin, J. and Wang, X. (2008). New Brownian Bridge in Quasi-Monte Carlo Methods for Computational Finance. *Journal of Complexity* **24**, 109-133.

Conference Presentations

Lin, J., Nie, L., & Li, R. (2010). *Trade off between accuracy and precision in generalized linear regression models*. ENAR, New Orleans.

Lin, J., Sudjianto, A., & Singhal, H. (2011). *Non-Proportional Hazard Approaches: First Hitting Time Models and Frailty Models Revisited*. JSM, Miami.