ASSOCIATION TESTS WITH INCOMPLETE COVARIATES AND HIGH-DIMENSIONAL AUXILIARY VARIABLES

FENG JIAHUI

PhD

The Hong Kong Polytechnic University

2022

The Hong Kong Polytechnic University

Department of Applied Mathematics

Association Tests with Incomplete Covariates and High-Dimensional Auxiliary Variables

FENG Jiahui

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

August 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

FENG Jiahui (Name of student)

Abstract

In many clinical and epidemiological studies, investigators are interested in testing the presence of association between an outcome variable and covariates of interest. Such analyses are often complicated by missing data. When variables of interest are missing for some subjects, it is desirable to use observed auxiliary variables, which are sometimes high-dimensional, to impute or predict the missing values to improve statistical efficiency. Although many methods have been developed for prediction using high-dimensional variables, it is challenging to perform valid inference based on the predicted values. In this dissertation, we propose novel association testing methods involving missing data with the goal of detecting relevant predictors for outcomes of interest.

We first focus on parametric models and develop an association test for an outcome variable and a partially missing covariate, where the missing values can be predicted using a set of high-dimensional auxiliary variables. The proposed analysis consists of a model selection step and a testing step. Specifically, in the first step, we select a subset of auxiliary variables and fit a regression model of the covariate of interest against the selected features. In the second step, we perform the score test for the covariate in the outcome model under the full likelihood, which includes both the outcome model and the missing covariate model. We then extend the proposed method to a class of semiparametric transformation models for potentially right-censored survival outcomes. We propose a supremum test, where we consider multiple choices of transformation functions, perform

individual score test under each outcome model, and take the supremum of the individual test statistics as the proposed test statistic. We show that the proposed testing procedure improves the test performance when the outcome model is unknown.

The validity and advantages of the proposed methods are demonstrated both theoretically and numerically. We establish the asymptotic properties of the proposed test statistics under regularity conditions and show the validity of the tests under data-driven model selection procedures. We evaluate the proposed methods through extensive simulation studies, and show their superior performances over some existing methods. Real data analyses are carried out on major cancer genomic studies.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Wong Kin Yau, for his invaluable guidance and lasting encouragement throughout the whole research process. During every single discussion with him, I benefit from his constructive suggestions and enlightening ideas. Dr. Wong shows me how to be a scholar with knowledge, curiosity and dedication towards research. He is and would continue to be my role model in academic. I am truly honored to study under and work with Dr. Wong.

Second, I would like to acknowledge Prof. Zhao Xingqiu for her administrative support throughout my postgraduate studies. I also want to thank my MA advisor Dr. Jiang Binyan, who has helped me tremendously with my PhD application and offered me a job as a research assistant before I started my doctoral study, which I will always be grateful. Acknowledgments also go to Dr. Liu Chun-ling, Catherine, who cares a lot about my academic study and career path.

Moreover, I would like to thank my committee members, Prof. Li Jialiang, Prof. Tong Tiejun and Dr. Zhang Guofeng for their helpful comments on the dissertation and the organization of my oral examination.

Last but certainly not least, my deepest gratitude goes to my family. My late grandmother, who took care of me, supported me and loved me. I am sorry I cannot be there in your last days. Also, I am extremely grateful to my parents and my little brother for their supports and understanding along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Statistical Analysis of Genomic Data

Statistical analysis with genomic data has been an area of considerable interest in the past decades, with a major goal of understanding complex human diseases and revealing the underlying biological mechanisms. This era of genomic study has seen some revolutionary advances in high-throughput technologies, such as microarray, whole-genome sequencing, single-cell sequencing, and RNA sequencing. We are now equipped with various types of genomic data, including DNA methylation, RNA expression, copy number alteration and protein expression, on a large number of subjects. This burst of genomic data poses challenges as well as opportunities on clinicians and researchers to gain a deeper understanding about disease mechanisms at a molecular level, and subsequently develop better prevention, diagnosis and treatment strategies. It is essential to develop valid and effective statistical approaches to utilize the information from different genomic measurements to investigate the relationships between genomic variables and interested outcomes, such as cancer phenotypes.

Here we give a brief introduction about some commonly studied types of genomic data.

DNA alteration DNA, or deoxyribonucleic acid, is the molecule that carries genetic information in all living things. The alterations in DNA sequences can lead to functional consequences. Genetic diseases, such as human cancers, are commonly driven by DNA alterations. DNA alterations can take several forms, such as single nucleotide polymorphisms, copy number variation and chemical modification.

RNA expression RNA, or ribonucleic acid, is composed of nucleotides. RNA plays an essential role in multiple physiological processes such as coding, decoding and regulation of gene expressions. The process of producing RNA from DNA is called transcription. The messenger RNA (mRNA) can further convey genetic information from RNA to protein; this process is termed translation. Another type of RNA, microRNA (miRNA), is a class of non-coding RNA molecules that does not involve in protein synthesis but regulates gene expression at the post-transcriptional level. miRNAs present diverse expression patterns and regulate various biologic processes.

Protein Protein is the basic cellular component in an organism and is involved in almost every biological process. Following transcription and translation, the functions of a protein are modulated by a set of post-translational modifications, including phosphorylation and methylation. Changes in protein expression level have been shown to be highly correlated with tumor progression.

One primary interest in genomic studies is to identify genomic features that are associated with outcomes of interest. This is of particular relevance to the field of cancer study, where one of the most important and fundamental objectives is to gain more insight into the molecular and genetic basis of cancer. Human diseases are influenced by a large number of factors including inherited variation, gene mutation and environmental exposures. The environmental exposures, along with sex and age are regularly studied and found to be associated with such diseases, but limited amount of risk can be explained by these fac-

tors. Various studies have demonstrated that genomic factors are of moderate importance in risk of diseases (Rosenwald et al., 2002; Rosenwald et al., 2003; Metzeler et al., 2008; Kim et al., 2013). For instance, Rosenwald et al. (2002) conducted a study of the diffuse large B-cell lymphoma (DLBCL), which is one of the most common types of lymphoma worldwide. The patients were grouped into three subgroups on the basis of gene expressions and presented significant differences on the overall survival after chemotherapy. Metzeler et al. (2008) studied the prognostic properties of gene expressions identified by the supervised principle component analysis (Bair et al., 2004) in cytogenetically normal acute myeloid leukemia, and developed a gene signature to predict the overall survival. One limitation of the above studies is that the conducted analyses only include a single type of genomic measurements, under most scenarios, the gene expressions.

In recent years, increasing efforts have been made to integrative analysis to detect important genomic features influencing human diseases, especially with the advent of multi-platform genomic data. Integrative analysis refers to the analysis that combines multiple types of data under a unified framework. For example, Bussey et al. (2006) performed an analysis on a panel of 60 human cancer cell lines to study the relationships among DNA copy number, mRNA expression and drug sensitivity. Shen et al. (2009) developed an integrative clustering method to incorporate different data types. Kristensen et al. (2012) studied the breast cancer heterogeneity by integrating different layers of molecular data into the analysis. Wong et al. (2019a) proposed a statistical method based on boosting to effectively integrate multiple layers of genomic information for predicting survival time. It has been argued that using multiple types of genomic data to explore associations between genomic features and outcomes of interest can be more powerful than using a single type of genomic measurements for several reasons. First, aggregation of information obtained across data types can enrich association signals and thus improve statistical efficiency. Individual feature analysis is usually inefficient because the activities

of one type of genomic features can only explain a part of the biological process underlying particular phenotypes. Further, integrative analysis can capture the interactive effects of multiple genomic features. Human complex diseases may depend on not only individual type of genomic features but also interactions among different types of genomic features. By including multiple data types into analysis, we are able to capture the indirect effects of genomic features on a phenotype through other genomic features. For example, Pollack et al. (2002) found a positive correlation between the variation in gene copy number and variation in gene expression in breast cancer cells. Finally, individual feature analysis can be vulnerable to unobserved information. There are situations that one type of data is subject to missing, the missing values can be inferred from other types of data by the underlying associations among them.

Although integrative analysis has demonstrated its great potential in revealing the complex molecular architecture of human diseases, the joint analysis of multiple data types must tackle some statistical challenges. One problem in genomic studies is missing data, arising due to costs or other constraints. This poses challenges on conventional statistical methods that do not accommodate missing data. Another critical issue is that the dimension of genomic variables is large when multiple data types are considered in the analysis. Many existing methods are not applicable to high-dimensional data. New statistical methods that can account for the problem of missing data and high dimensionality need to be studied. In this dissertation, we seek to develop novel statistical methods to identify the associations between partially observed genomic variables and outcomes of interest, where another set of high-dimensional genomic data is available for the prediction of the missing values.

1.2 Survival Analysis

In cancer genomic studies, the outcome of interest is often an event time, such as time to death since initial diagnosis, time to tumor progression, and response time to a medical treatment. One common feature presented in time-to-event data is censoring, which arises when the event is only known to have occurred in a certain period of time. Typical types of censoring include right censoring, interval censoring, and left censoring. In survival analysis, it is of interest to ascertain the relationship between the time-to-event outcome and some variables, including clinical characteristics and genomic features, and formulate the effects of the covariates on the survival time.

The proportional hazards model (Cox, 1972) is widely used in survival analysis. Let T denote the survival time. Given a subject with covariates $\boldsymbol{X}$, the hazard function of T, or instantaneous rate of occurrence of the event, is specified by

$$\lambda(t \mid \boldsymbol{X}) = \lambda(t) \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}),$$

where $\lambda(\cdot)$ is an unknown baseline hazard function and $\boldsymbol{\alpha}$ is a vector of regression parameters. It is easy to see that the hazard ratio between two subjects is constant over time. Cox (1972, 1975) proposed to estimate $\boldsymbol{\alpha}$ by maximizing the partial likelihood, and the advantage of the partial likelihood method is that the estimation of the nonparametric baseline hazard function is avoided. Breslow (1972) suggested an estimation approach for the nonparametric baseline hazard function using the joint likelihood function involving parameters $\boldsymbol{\alpha}$ and λ. The asymptotic properties of the maximum partial likelihood estimator and the Breslow estimator of the cumulative baseline hazard function are established in Andersen and Gill (1982) via the counting-process martingale theory.

Another commonly used model in survival analysis is the proportional odds model (Pettitt, 1982; Bennett, 1983a, 1983b; Dabrowska & Doksum, 1988). Under this model,

the odds ratio of the survival probabilities between two subjects is constant over time, while the hazard ratio converges to one as t goes to infinity rather than staying constant. The proportional odds model assumes that

$$-\text{logit}\{S(t \mid \boldsymbol{X})\} = g(t) + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X},$$

where $\text{logit}(x) = \log\{x/(1-x)\}$, $g(t)$ is an arbitrary increasing function, and $S(t \mid \boldsymbol{X})$ is the survival function given covariates $\boldsymbol{X}$. Bennett (1983b) proposed to estimate $\boldsymbol{\alpha}$ by maximizing the likelihood function with $g(t) = \varphi \log t$, where φ is some nonnegative parameter. Bennett's estimator of $\boldsymbol{\alpha}$ is the maximum profile likelihood estimator with the nuisance parameter φ estimated out. Maximum likelihood estimation (MLE) for the proportional odds model was studied by Murphy et al. (1997).

Both the proportional hazards model and the proportional odds model are special cases of the transformation model. Under the transformation model, the cumulative hazard function for T conditional on $\boldsymbol{X}$ takes the form

$$\Lambda(t \mid \boldsymbol{X}) = G\{\Lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X})\}, \tag{1.1}$$

where $\Lambda(\cdot)$ is an unknown increasing function in $[0, \tau]$ with $\Lambda(0) = 0$, $G(\cdot)$ is a prespecified transformation function that is strictly increasing with $G(0) = 0$, and τ is the end-of-study time. For example, we can consider the class of Box-Cox transformations

$$G(x) = \begin{cases} \{(1+x)^{\rho} - 1\}/\rho & \text{for } \rho > 0, \\ \log(1+x) & \text{for } \rho = 0, \end{cases}$$

where ρ is a prespecified transformation parameter. In this family, $\rho = 1$ corresponds to the proportional hazards model, and $\rho = 0$ corresponds to the proportional odds model.

Alternatively, we can consider the class of logarithmic transformations

$$G(x) = \begin{cases} r^{-1}\log(1+rx) & \text{for } r > 0, \\ x & \text{for } r = 0, \end{cases}$$

where r is a prespecified transformation parameter. Clearly, the choices of $r = 0$ and $r = 1$ yield the proportional hazards model and the proportional odds model, respectively. Under (1.1), the model of T can be expressed as a linear transformation model, with

$$\log \Lambda(T) = -\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \epsilon,$$

where ϵ is an error term with $P(\epsilon < t) = 1 - \exp[-G\{\exp(t)\}]$. Particularly, the choices of the extreme value distribution with $P(\epsilon < t) = 1 - \exp\{-\exp(t)\}$ and the standard logistic distribution with $P(\epsilon < t) = \exp(t)/\{1 + \exp(t)\}$ yield the proportional hazards model and the proportional odds model, respectively.

For semiparametric transformation models, Cheng et al. (1995) proposed a generalized estimating equation to estimate $\boldsymbol{\alpha}$ with right-censored survival data. They used the inverse weight of the Kaplan-Meier estimator for the survival function of the censoring variable to adjust the censoring under the assumption that the censoring distribution is independent of covariates. Chen et al. (2002) relaxed such an assumption and proposed an estimator of $\boldsymbol{\alpha}$ using a general estimating equation in terms of counting process notations, which can be reduced to the partial likelihood score equation under the proportional hazards model. Zeng and Lin (2007) and Zeng et al. (2016) studied nonparametric maximum likelihood estimation (NPMLE) methods with right-censored data and interval-censored data, respectively, and showed that the estimators are consistent and asymptotically efficient. Efficient expectation-maximization (EM) algorithms were developed for the computation of the proposed NPMLE.

1.3 Incomplete Data Analysis

One complication in statistical analysis is the presence of missing data. For example, in sample surveys, it is common that some participants do not answer all questions due to refusal or other reasons. Missing data are also commonly encountered in longitudinal studies such as clinical trials, where some subjects may drop out before the end of study. The problem of missing data is especially prevalent in large scale genomic studies, where multiple types of genomic data are collected on a large number of subjects, often over different locations and time periods. For example, in The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/), over 11,000 subjects with 33 cancer types were measured for multiple types of genomic data, including DNA methylations, mutations, RNA expressions, and protein expressions, but protein expressions were not measured for a substantial number of subjects. As another example, in the Trans-Omics for Precision Medicine (TOPMed) program (https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed), whole-genome sequencing data are available on hundreds of thousands of subjects, but other types of genomic data, such as RNA sequencing, methylation, and metabolites, are available for only tens of thousands of subjects or fewer. The missing data problem arises when one attempts to analyze these data under a unified framework.

Rubin (1976) presented a general taxonomy of missing data mechanism based on the propensity of missing data. To formalize this problem, let $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^{\mathrm{T}}$ denote the study variable with dimension K. Let $\boldsymbol{R} = (R_1, \ldots, R_K)^{\mathrm{T}}$ denote the indicator vector where R_j $(j = 1, \ldots, K)$ takes value one or zero depending on whether the corresponding element in $\boldsymbol{Z}$ is observed or not, defined by

$$R_j = \begin{cases} 1 & \text{if } Z_j \text{ is observed,} \\ 0 & \text{if } Z_j \text{ is missing.} \end{cases}$$

Then the data can be partitioned as $\boldsymbol{Z} = (\boldsymbol{Z}_{\text{obs}}, \boldsymbol{Z}_{\text{mis}})$, where $\boldsymbol{Z}_{\text{obs}} = \{Z_j : R_j = 1\}$ is the observed part of $\boldsymbol{Z}$, and $\boldsymbol{Z}_{\text{mis}} = \{Z_j : R_j = 0\}$ is the missing part of $\boldsymbol{Z}$. Let $p(\boldsymbol{R} \mid \boldsymbol{Z})$ denote the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{Z}$. If $\boldsymbol{R}$ and $\boldsymbol{Z}$ are independent, then $p(\boldsymbol{R} \mid \boldsymbol{Z}) = p(\boldsymbol{R})$, and the data is missing completely at random (MCAR). In this case, there are no systematic differences between observed data and missing data. One example is the situation that patients drop out clinical trials because of personal reasons that have nothing to do with the issues under study. The second type of missing mechanism is missing at random (MAR), where $\boldsymbol{R}$ does not depend on the missing values of $\boldsymbol{Z}$ given the observed values of $\boldsymbol{Z}$, which we write as $p(\boldsymbol{R} \mid \boldsymbol{Z}_{\text{mis}}, \boldsymbol{Z}_{\text{obs}}) = p(\boldsymbol{R} \mid \boldsymbol{Z}_{\text{obs}})$. Two-phase sampling study is a classical MAR situation, where the outcome and inexpensive covariates are observed for all subjects in phase 1, and then a sub-group of subjects is selected for measurements on expensive covariates in phase 2 based on the results of phase 1. Lastly, data are said to be missing not at random (MNAR) if the missingness depends on the missing values. For example, consider a survey about income. MNAR would occur if participants refuse to respond because of their levels of income. Among the three types of missing mechanism, MNAR is the weakest assumption while MCAR is the strongest.

Various methods have been developed in the past decades for statistical analysis with missing data (Ibrahim et al., 2005; Little & Rubin, 2019). A naive approach to handle missing data is to perform a complete-case analysis, where subjects with missing data are discarded. Such an approach is obviously inefficient because information of partially observed subjects would be discarded. In addition, the complete-case analysis is generally invalid under situations other than MCAR, since the complete cases need not to be a representative sample from the population, and consequently substantial bias may be introduced to the estimation. An alternative approach is single imputation, where the missing values are imputed by plausible values based on the observed data, and conventional methods are then applied to the imputed dataset. However, although estimation

based on imputed data may be more efficient than a complete-case analysis, conventional inferential procedures based on (singly) imputed data are generally invalid. Rubin (1987) mentioned that the standard error estimator from the single imputation method is systematically underestimated because the uncertainty from the imputation process is not incorporated into the analysis. More sophisticated statistical methods to handle missing data can be broadly classified into three categories, namely the likelihood-based approach, the multiple imputation (MI) approach, and the inverse-probability weighting (IPW) approach.

1.3.1 Likelihood-Based Approach

The method of maximizing the observed data likelihood is often used in regression analysis with missing data. The observed data likelihood is obtained by finding the marginal distribution of the observed data resulting from the integration of the joint distribution of full data over the missing variables. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ denote n independent realizations of $\boldsymbol{Z}$ with probability density function $f(\boldsymbol{Z}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters of interest. Let $\boldsymbol{R}_i$ $(i = 1, \ldots, n)$ denote the indicator vector corresponding to $\boldsymbol{Z}_i$. The likelihood based on the observed data $\{(\boldsymbol{Z}_{\text{obs},i}, \boldsymbol{R}_i), i = 1, \ldots, n\}$ is given by

$$L_{\text{obs}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int f(\boldsymbol{Z}_i; \boldsymbol{\theta}) p(\boldsymbol{R}_i \mid \boldsymbol{Z}_i) \, \mathrm{d}\nu(\boldsymbol{Z}_{\text{mis},i}),$$

where ν is some dominating measure for $\boldsymbol{Z}_{\text{mis},i}$. Under scenarios that (a) the probability function $p(\boldsymbol{R}_i \mid \boldsymbol{Z}_i)$ does not involve parameter $\boldsymbol{\theta}$, and (b) the missing mechanism is MAR or MCAR, the missingness is considered to be ignorable (Rubin, 1976) because $p(\boldsymbol{R}_i \mid \boldsymbol{Z}_i) = p(\boldsymbol{R}_i \mid \boldsymbol{Z}_{\text{obs},i})$. In this case, the missing data probability function can be factored out from the likelihood function. Specifically, we can write the observed data

likelihood as

$$
\begin{aligned}
L_{\text{obs}}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \int f(\boldsymbol{Z}_i; \boldsymbol{\theta}) p(\boldsymbol{R}_i \mid \boldsymbol{Z}_{\text{obs},i}) \, \mathrm{d}\nu(\boldsymbol{Z}_{\text{mis},i}) \\
&= \prod_{i=1}^{n} \int f(\boldsymbol{Z}_i; \boldsymbol{\theta}) \, \mathrm{d}\nu(\boldsymbol{Z}_{\text{mis},i}) \times \prod_{i=1}^{n} p(\boldsymbol{R}_i \mid \boldsymbol{Z}_{\text{obs},i}).
\end{aligned}
$$

Therefore, likelihood inference can be conducted by considering only the likelihood for $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ and not the missing mechanism.

Since the observed data likelihood is obtained by integrating the full data likelihood over the missing values of $\boldsymbol{Z}$, it is usually difficult to maximize the likelihood function with standard methods. The EM algorithm (Dempster et al., 1977) is an iterative numerical technique often used to compute the MLE from the observed data likelihood function. It consists of two steps: (1) the E-step computes the expectation of log-likelihood of $\boldsymbol{\theta}$ given the observed data and some initial value of $\boldsymbol{\theta}$; (2) the M-step maximizes the expectation from the previous step to find an estimator of $\boldsymbol{\theta}$. The two steps are iterated until convergence.

The likelihood-based methods have been extensively studied with different parametric models in the presence of missing data. Fuchs (1982) derived the MLE for the parameters of log-linear models with missing data. Little and Schluchter (1985) applied it to mixed continuous and categorical variables with missing values by combining the multivariate normal model for continuous variables and multinomial model for categorical data into the analysis. Ibrahim (1990) developed an EM algorithm for generalized linear models with missing discrete covariates, and later extended this method to incomplete categorical covariates in survival analysis (Lipsitz & Ibrahim, 1996), and continuous or mixed categorical and continuous covariates with missingness (Ibrahim et al., 1999). The parametric model approach, albeit convenient and easy to implement, is sensitive to model misspecification. Many efforts have been made to semiparametric likelihood methods

in which the missing covariate model or the missing data mechanism is assumed to be nonparametric. For example, Lawless et al. (1999) studied the semiparametric methods with a nonparametric missing covariate model where the missingness depends on a stratification outcome variable, and the asymptotic properties of the semiparametric MLE of parameters of interest is established in Breslow et al. (2003). Chatterjee et al. (2003) proposed a pseudoscore estimator for two-phase designs. Zhang and Rockette (2005, 2006) considered a semiparametric likelihood approach with a generalized linear outcome model and an unspecified covariate model when some variables MAR. Zhao et al. (2009) studied the problem where covariates and/or responses are missing by design, and proposed a semiparametric likelihood estimation method with nonparametric assumptions about the conditional distribution of missing covariate given some always observed variables. Kim and Yu (2011) proposed semiparametric estimators with nonignorable nonresponse data by considering a semiparametric logistic regression model for the response probability.

1.3.2 Multiple Imputation Approach

The MI approach was first proposed by Rubin (1978) for complex surveys with missingness. The basic idea of MI is to create multiple complete datasets by filling in the missing entries with imputed values based on the observed data. Similar to the single imputation method, MI has the practical advantage of allowing the standard complete-date methods to be used on the imputed datasets. Moreover, the MI method improves on the single imputation method by producing more accurate variance estimator.

The procedure of MI involves three steps: imputation, estimation, and pooling of results. In the step of imputation, multiple copies of missing values, say M, are generated from a posterior predictive distribution of the missing values conditional on the observed data. In each dataset, the missing values are replaced by the imputed values. In the estimation step, standard complete-data methods can be applied to analyze each com-

pleted dataset, which leads to M estimated parameters $\widehat{\boldsymbol{\theta}}_m$ and corresponding estimated variance $\boldsymbol{W}_m$ for $m = 1, \ldots, M$. The results from each analysis will differ because of the variability introduced by the imputation process. In the last step, all the results are combined to obtain a single set of results. Specifically, the pooled estimator of $\boldsymbol{\theta}$ is obtained by taking the average of the M estimators from previous step:

$$\widehat{\boldsymbol{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\boldsymbol{\theta}}_m.$$

The variance estimator of $\widehat{\boldsymbol{\theta}}$ is derived using Rubin's rules (Rubin, 1987):

$$\widehat{\boldsymbol{V}} = \boldsymbol{W} + (1 + \frac{1}{M})\boldsymbol{B},$$

where $\boldsymbol{W} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{W}_m$ is the within imputation variance, and $\boldsymbol{B} = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{\boldsymbol{\theta}}_m - \widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_m - \widehat{\boldsymbol{\theta}})^{\mathrm{T}}$ is the between imputation variance. The between imputation variance caused by the differences in the imputed values across M datasets reflects the uncertainty of the missing data.

1.3.3 Inverse-Probability Weighting Approach

An alternative class of methods to deal with missing data is by weighting the complete cases with the inverse of the probability of being non-missing (Zhao & Lipsitz, 1992; Robins et al., 1994). The likelihood-based methods under some scenarios can ignore the missing mechanism when performing inference on the parameter of interest. On the contrary, the IPW approach directly involves the missing mechanism.

Let $\mu(\boldsymbol{Z}; \boldsymbol{\theta})$ be some estimating function for $\boldsymbol{\theta}$ based on the full data, for example, the score function in a parametric model. The estimating equation that would be solved

to estimate the regression parameter is given by

$$\sum_{i=1}^{n} \mu(\boldsymbol{Z}_i; \boldsymbol{\theta}) = 0.$$

We say that the estimating function is valid if $\mathrm{E}\{\mu(\boldsymbol{Z}; \boldsymbol{\theta}_0)\} = 0$ with $\boldsymbol{\theta}_0$ being the true value of $\boldsymbol{\theta}$. Define $r_i = 1$ if $\boldsymbol{Z}_i$ is fully observed, and $r_i = 0$ otherwise. The complete-case analysis corresponds to solving the estimating equation with fully observed cases $\sum_{i=1}^{n} r_i \mu(\boldsymbol{Z}_i; \boldsymbol{\theta}) = 0$, which is invalid unless the missing data is MCAR. The IPW approach corrects the bias of a complete-case analysis by reweighting each complete case by the probability of being observed $p_i(\boldsymbol{Z}) = P(r_i = 1 \mid \boldsymbol{Z}_i)$. The estimating equation is in the form of

$$\sum_{i=1}^{n} \frac{r_i}{p_i(\boldsymbol{Z})} \mu(\boldsymbol{Z}_i; \boldsymbol{\theta}) = 0.$$

For example, if the missingness is related with sex, and male subjects are less likely to respond, then the inverse probability term would up-weight the male cases and down-weight the female cases to reflect the entire population. In practice, the missing mechanism is usually unknown and needs to be estimated. The estimate of $p_i(\boldsymbol{Z})$ can be obtained, for example, by a logistic regression of r_i on some always observed covariates in $\boldsymbol{Z}_i$. However, the estimator of $\boldsymbol{\theta}$ under IPW approach can be biased if the model of missing mechanism is misspecified, and the estimator is generally inefficient because only the information contained in the complete cases is used. To improve efficiency, Robins et al. (1994, 1995) proposed the augmented inverse probability approach by including the additional information contained in the incomplete cases into the analysis. The augmented estimating equation takes the form of

$$\sum_{i=1}^{n} \left\{ \frac{r_i}{p_i(\boldsymbol{Z})} \mu(\boldsymbol{Z}_i; \boldsymbol{\theta}) + \Big(1 - \frac{r_i}{p_i(\boldsymbol{Z})}\Big) \eta(\boldsymbol{Z}_{\mathrm{obs},i}; \boldsymbol{\theta}) \right\} = 0,$$

where $\eta(\boldsymbol{Z}_{\text{obs}};\boldsymbol{\theta})$ is some function of the observed data that takes the place of μ for incomplete cases. Rotnitzky and Robins (1995) showed that the optimal estimator of $\boldsymbol{\theta}$ is obtained when $\eta(\boldsymbol{Z}_{\text{obs}};\boldsymbol{\theta}) = \mathrm{E}\{\mu(\boldsymbol{Z};\boldsymbol{\theta}) \mid \boldsymbol{Z}_{\text{obs}}\}$. Another advantage of the augmented estimating equation is that the resulting estimator is doubly robust in the sense that it is consistent when either the missing mechanism model or the distribution of the missing covariates, but not necessarily both, is correctly specified (Bang & Robins, 2005; Kang & Schafer, 2007).

1.3.4 Association Testing for Incomplete Data

Despite extensive studies on regression analysis with missing data, association testing with incomplete data has received less attention. Although association testing typically can be done under a regression framework, the testing problem should not be treated as a trivial special case. There are issues of interest in hypothesis testing that are not pertinent under a general regression analysis framework. First, in estimation, strong assumptions concerning the distribution of the incomplete variables are usually made to ensure desirable theoretical properties, such as consistency of the estimators. By contrast, in hypothesis testing, we are primarily concerned with the theoretical properties of estimators under the null hypothesis (or contiguous alternatives), and correct specification of the full model is not required for a test to be valid. As a result, much more relaxed model assumptions could be considered for association tests than for regression analyses in general. Second, estimation is generally performed under specific models. On the other hand, in hypothesis testing, we may only be interested in the existence of association between a covariate and an outcome variable, not necessarily under particular models.

For illustration, let Y denote the response variable, S denote the covariate of interest that might be missing, $\boldsymbol{X}$ denote a vector of other covariates. Let R be the indicator of

whether S is observed, i.e., $R = 1$ if S is observed, and $R = 0$ otherwise. Assume that $Y \mid (\boldsymbol{X}, S) \sim f(Y; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S, \boldsymbol{\zeta})$, where $\boldsymbol{\alpha}$ and β are the regression parameters, and $\boldsymbol{\zeta}$ is a set of nuisance parameters, and $S \mid \boldsymbol{X} \sim g(S \mid \boldsymbol{X}; \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a parameter vector. For a sample of size n, the observed data consist of $\{(Y_i, \boldsymbol{X}_i, R_i S_i, R_i) : i = 1, \ldots, n\}$. We are interested in testing the null hypothesis $H_0 : \beta = 0$.

To perform association tests with missing genotype data, Hu et al. (2015) considered a linear regression model and a logistic regression model for quantitative Y and binary Y, respectively, on S and $\boldsymbol{X}$, and proposed a score test based on imputed genotype data. Specifically, the score statistic under H_0 takes the form

$$U = \sum_{i=1}^{n} \{Y_i - h(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{X}_i)\}\widetilde{S}_i,$$

where $h(x) = x$ for the linear model of Y, and $h(x) = e^x/(1 + e^x)$ for the logistic model of Y, $\widehat{\boldsymbol{\alpha}}$ is the estimator of $\boldsymbol{\alpha}$ under the null hypothesis, and $\widetilde{S}_i$ is the imputed value of S_i if $R_i = 0$ and the observed value of S_i, otherwise. They also proposed a robust variance estimator for U that properly accounts for the differential quality between observed and imputed genotypes. Under outcome-dependent sampling designs, Derkach et al. (2015) and Lawless (2018) proposed to model the variable with missing values and studied the score test based on the full likelihood. The full likelihood function for the observed data is given by

$$L(\boldsymbol{\alpha}, \beta, \boldsymbol{\zeta}, \boldsymbol{\xi}) = \prod_{R_i=1} f(Y_i \mid \boldsymbol{X}_i, S_i; \boldsymbol{\alpha}, \beta, \boldsymbol{\zeta}) g(S_i \mid \boldsymbol{X}_i; \boldsymbol{\xi}) \times \prod_{R_i=0} f_1(Y_i \mid \boldsymbol{X}_i; \boldsymbol{\alpha}, \beta, \boldsymbol{\zeta}, \boldsymbol{\xi}),$$

where $f_1(Y \mid \boldsymbol{X}; \boldsymbol{\alpha}, \beta, \boldsymbol{\zeta}, \boldsymbol{\xi})$ is the probability density function of Y given $\boldsymbol{X}$. The score statistic for β is

$$U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\xi}}) = \sum_{i=1}^{n} A(Y_i, \boldsymbol{X}_i; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\zeta}}) \big\{ R_i S_i + (1 - R_i) \widehat{\mathrm{E}}(S_i \mid \boldsymbol{X}_i) \big\},$$

where $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\xi}})$ are estimators of $(\boldsymbol{\alpha}, \boldsymbol{\zeta}, \boldsymbol{\xi})$ under H_0, $\widehat{\mathrm{E}}(S \mid \boldsymbol{X}) = \int s g(s \mid \boldsymbol{X}; \widehat{\boldsymbol{\xi}}) \, \mathrm{d}\nu(s)$, and $A(Y, \boldsymbol{X}; \boldsymbol{\alpha}, \boldsymbol{\zeta}) = \partial \log f(Y; t + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}, \boldsymbol{\zeta}) / \partial t|_{t=0}$. Derkach et al. (2015) treated $g(\cdot)$ as nonparametric and required $\boldsymbol{X}$ to be discrete. Lawless (2018) assumed parametric models for the missing variable S on $\boldsymbol{X}$. Under extreme phenotype sampling designs in genetic association studies, Bjørnland et al. (2018) considered a similar model-based score test and a complete-case score test based on the conditional likelihood given the sampling mechanism. Wong et al. (2019b) proposed to model the incomplete variable semiparametrically and developed a score test that is robust against misspecification of the missing variable model. The proposed score statistic of Wong et al. (2019b) is similar to that of Derkach et al. (2015) and Lawless (2018) with the posterior mean of S given $\boldsymbol{X}$ replaced by the imputed value of S.

In many applications, there often exist auxiliary variables that are associated with the missing covariate or correlated with the missingness. Auxiliary variables refer to variables that are available for all subjects but not included in the main analysis. By incorporating auxiliary variables into the analysis framework, the analysis performance may be improved by reducing bias and increasing efficiency. Collins et al. (2001) discussed two auxiliary variable selection strategies with ML and MI procedures. One is inclusive strategy by including a large amount of auxiliary variables into the missing data analysis. The other one is restrictive strategy, which uses few or no auxiliary variables. They conclude that the inclusive strategy is recommended since it reduces the chances of omitting variables correlated with the missing covariate or the cause of missingness and have substantial gains in terms of both bias and efficiency. Collins et al. (2001) also mentioned that the MNAR mechanism can be converted to MAR mechanism by adding auxiliary variables into the analysis model, the reason is that the auxiliary variables enter into the model as observed variables, and can help to eliminate the relationship between outcome variable and the missing covariate part. Graham (2003) developed two structural equation

models to incorporate auxiliary variables without affecting the substantive interpretation of the parameters, and adopted MLE for estimation. Hardt et al. (2012) studied the performances of MI with different choices of auxiliary variables and found that too many auxiliary variables can potentially lead to biased estimators. They suggested that the number of auxiliary variables included in the imputation model should be no more than one third of the number of cases with complete data.

1.4 High-Dimensional Data Analysis

With the fast developments in high-throughput technologies, we often encounter data that are high-dimensional in biomedical studies, where the number of variables p is larger than the sample size n. For example, in the DLBCL study (Rosenwald et al., 2002), hundreds of thousands of genetic variants such as gene expressions were collected from 240 patients. Under such regimes, classical statistical methods often fail to provide valid estimation or prediction. Consider a linear regression model

$$Y = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X} + \epsilon, \tag{1.2}$$

where Y is the response variable, $\boldsymbol{X}$ is a p-vector of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-vector of regression parameters, and ϵ is a random error term with mean zero. For simplicity, we assume that $\boldsymbol{X}$ and Y are centered. Let $(Y_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$ denote n independent realizations of $(Y, \boldsymbol{X})$. In low-dimensional settings, an estimator of $\boldsymbol{\beta}$ can be obtained by

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)^2.$$

The above problem has an explicit solution provided that $\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}}$ is invertible. Under the scenario of $n < p$, the least-squares estimator of $\boldsymbol{\beta}$ is no longer well-defined. To

deal with the large number of variables, it is common to use penalization methods to simultaneously select variables and estimate coefficients.

1.4.1 Estimation Methods with High-Dimensional Data

The objective of variable selection is to identify a subset of variables that are relevant to the response variable. Conventional variable selection methods include forward selection and backward selection. In forward selection, the initial model has only the intercept term and variables are then iteratively added until some statistical criterion is satisfied. The backward selection method starts with the full model and removes variables with least importance until meeting some criterion. However, the backward selection method is not applicable for high-dimensional data, and both methods suffer from low prediction accuracy (Harrell et al., 1996). An alternative approach is the best subset selection by choosing a subset of variables that minimizes some criterion from all possible combinations of the variables. It is computationally infeasible for high-dimensional data because there are 2^p candidate models to consider.

In the past decades, penalization methods have been extensively studied in the literature because of their advantages in computation. To encourage sparsity, a penalty term is employed on the regression parameters so that a large number of regression coefficients could be shrunk to zero. The least absolute shrinkage and selection operator (lasso) is introduced in the seminal work of Tibshirani (1996), and the estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is a nonnegative regularization/tuning parameter. The second term on the right-hand side of the above equation is termed ℓ_1 penalty. The first term on the right-hand side above measures the goodness-of-fit and tends to be smaller when more variables are

included, while the ℓ_1 penalty term measures the model complexity and increases as more variables enter into the model. The lasso performs variable selection in the way that some estimated coefficients can be shrunk to exactly zero with large enough λ. The larger value of λ leads to more shrinkage. Lasso is closely related to the bridge regression (Frank & Friedman, 1993; Fu, 1998) with bridge penalty function $\lambda \sum_{j=1}^{p} |\beta_j|^q$ for $0 < q \leq \infty$. The choice of $q = 1$ corresponds to the lasso, and $q = 2$ to the ridge regression. Although lasso obtains high prediction accuracy by applying shrinkage to the coefficients due to the bias-variance trade-off, there is no guarantee in the consistency of variable selection except under some simple settings (Meinshausen & Bühlmann, 2006). In order to achieve consistent variable selection in lasso, a nontrivial condition is usually required (Zhao & Yu, 2006).

To overcome this problem, Zou (2006) developed the adaptive lasso in which the ℓ_1 penalty is replaced by a weighted version. Specifically, the adaptive lasso estimator is given by

$$\widehat{\boldsymbol{\beta}}_{\text{adalasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where $\boldsymbol{w} = 1/|\widehat{\boldsymbol{\beta}}|^{\gamma}$ is the vector of weights that adjusts the size of shrinkage to each coefficient, where $\gamma \geq 0$, and $\widehat{\boldsymbol{\beta}}$ is a consistent initial estimator of $\boldsymbol{\beta}$ such as the least-squares estimator. Moreover, the adaptive lasso possesses the oracle property with fixed p in the sense that the penalized estimators are asymptotically equivalent to the estimators obtained by the true model without penalization, whereas the lasso does not. Huang et al. (2008) further explored the oracle property of adaptive lasso when p is larger than n by considering a different weight vector $\boldsymbol{\omega}$ with the jth element $\omega_j = |\sum_{i=1}^{n} X_{ij}^2 / \sum_{i=1}^{n} X_{ij} Y_i|^{\gamma}$, where X_{ij} is the jth element of $\boldsymbol{X}_i$.

The lasso and adaptive lasso are essentially convex optimization problems, and some considered non-convex penalty functions. One prominent example is the smoothly clipped absolute deviation (SCAD) proposed by Fan and Li (2001). The SCAD estimator is in

the form of

$$\widehat{\boldsymbol{\beta}}_{\text{SCAD}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)^2 + p_\lambda(\boldsymbol{\beta}),$$

with the derivative of the penalty defined by

$$p'_\lambda(u) = \lambda \Big\{ I(u \le \lambda) + \frac{(a\lambda - u)_+}{(a-1)\lambda} I(u > \lambda) \Big\},$$

for some $a > 2$ and nonnegative λ. Fan and Li (2001) showed that the SCAD estimator enjoys the oracle property. A similar penalty called minimax concave penalty (MCP) is proposed by Zhang (2010) with $p'_\lambda(u) = 1 - u/(a\lambda)$ for $a > 0$.

When the variables are predefined into groups, it is useful to consider the group lasso (Yuan & Lin, 2006). Suppose that there are L groups. The group lasso estimator is defined by

$$\widehat{\boldsymbol{\beta}}_{\text{grouplasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i)^2 + \lambda \sum_{l=1}^{L} \sqrt{\rho_l} \|\boldsymbol{\beta}_l\|_2,$$

where $\sqrt{\rho_l}$ accounts for the varying group size, and $\boldsymbol{\beta}_l$ denotes the coefficients of $\boldsymbol{X}_i$'s in the l-th group for $l = 1, \ldots, L$. The group lasso penalty performs like the lasso penalty on the group level, and the coefficients within a group could all shrunk to be zeros with appropriate λ. When there is only one variable within each group, the group lasso reduces to lasso.

Another variable selection method for high-dimensional data is screening based on some association measurements between individual predictors and response. Unlike penalization methods, screening measures the effect of each variable individually and rank the variables by the measurements. The sure independence screening (SIS) procedure proposed by Fan and Lv (2008) is based on the marginal correlation of $\boldsymbol{X}$ and Y. A subset of relevant features is determined by the rank of marginal correlations, and defined as

$$M = \{1 \le j \le p : |\omega_j| \text{ is among the first } d \text{ largest of all}\},$$

where ω_j is the empirical correlation coefficient of Y and the jth component of $\boldsymbol{X}$, and d is some integer smaller than p. The SIS procedure reduces the dimension of predictors from a large scale to a moderate size. The probability of all the important variables are included in the final submodel tends to one, which is known as the sure screening property. Other examples of screening methods include rank correlation screening (Li et al., 2012a) and distance correlation screening (Li et al., 2012b).

1.4.2 High-Dimensional Inference

For high-dimensional data, estimation and selection methods are widely available. The problem of inference, however, remains relatively less explored and challenging. A major difficulty of such a problem is to find the limiting distribution of estimators. Recent work on the area of high-dimensional inference can be roughly divided into two classes: methods that make inference of parameters in the full, high-dimensional regression model, and methods that make inference under a (randomly) selected model.

The methods that make inference of parameters in the full model are often based on debiased estimators (Bühlmann, 2013; Zhang & Zhang, 2014; van de Geer et al., 2014; Javanmard & Montanari, 2014a, 2014b) or decorrelated score functions (Ning & Liu, 2017; Li et al., 2021). Consider the lasso estimator $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$. The inference of $\boldsymbol{\beta}$ constructed around $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ would be problematic since the lasso estimator is a biased estimator of $\boldsymbol{\beta}$. Zhang and Zhang (2014) proposed to correct the bias of the lasso estimators by a low-dimensional projection method in order to form valid confidence intervals for individual regression coefficients. van de Geer et al. (2014) extended the method of Zhang and Zhang (2014) to generalized linear models with convex loss function and established asymptotic efficiency for the debiased estimators. Ning and Liu (2017) provided a general framework for high-dimensional inference based on the decorrelated score functions, the proposed method is applicable to generalized linear models and additive hazards model. The aforementioned

methods are all performed under the sparsity assumption that only a few predictors have contributions to the response.

The second class of methods studies the post-selection inference problem. It is well-known that in general, conventional inferential procedures, such as the F test and the t test, on a selected model are invalid, because the parameters to be estimated or tested arise from a data-driven model selection procedure and are "random". Specifically, let $\mathcal{K}^* \subset \{1, \ldots, p\}$ denote the (random) set of indices of the selected variables, $\mathcal{K}$ denote the observed value of $\mathcal{K}^*$, and $\boldsymbol{X}_{\mathcal{K}}$ denote the elements of $\boldsymbol{X}$ indexed by $\mathcal{K}$. Instead of the full model (1.2), we consider the following model

$$Y = \boldsymbol{\beta}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{K}} + \widetilde{\epsilon}, \tag{1.3}$$

where $\boldsymbol{\beta}_{\mathcal{K}}$ is the vector of the regression coefficients corresponding to predictors in $\boldsymbol{X}_{\mathcal{K}}$, and $\widetilde{\epsilon}$ is an error term. The true value of the regression parameter in the sub-model (1.3) is $\boldsymbol{\beta}_{\mathcal{K}0} = \arg\min_{\boldsymbol{\beta}_{\mathcal{K}}} \mathrm{E}\|Y - \boldsymbol{\beta}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{K}}\|^2$, which is generally not the same as the corresponding components of the true value of $\boldsymbol{\beta}$ in the full model (1.2). This is because the regression coefficients in the full model (1.2) measure the effect of a given variable on Y conditional on the other $p-1$ variables, whereas the coefficients in the model (1.3) capture the effect of a variable in $\boldsymbol{X}_{\mathcal{K}}$ on Y, adjusted for the other variables in the selected model. Another concern is that the target $\boldsymbol{\beta}_{\mathcal{K}}$ changes with the model selection procedure and thus is random. The randomness lies in the choice of which parameters to estimate instead of the parameters themselves. The subset of parameters to be estimated and draw inference on can be expressed as $\{\beta_{\mathcal{K},j} : \mathcal{K} \subset \{1, \ldots, p\}, j \in \mathcal{K}\}$. We wish to form inferences for the parameters $\beta_{\mathcal{K}^*,j}$ in the model $\mathcal{K}^*$ selected. For instance, suppose there exists a confidence interval $C_{\mathcal{K}^*,j}$ such that

$$P(\beta_{\mathcal{K}^*,j} \in C_{\mathcal{K}^*,j}) \geq 1 - \alpha,$$

where α is the significance level. However, the event inside the probability is not well-defined because for $j \notin \mathcal{K}$, the parameter $\beta_{\mathcal{K},j}$ is undefined.

To deal with the above issues, one approach is to perform conditional inference for the model parameters given the model selection event. In specific, we may consider a confidence interval $C_{\mathcal{K},j}$ such that for any $j \in \mathcal{K}$, we have

$$P(\beta_{\mathcal{K},j} \in C_{\mathcal{K},j} \mid \mathcal{K}^* = \mathcal{K}) \geq 1 - \alpha.$$

Lee et al. (2016) considered a model selected by lasso and derived valid confidence intervals of the lasso coefficients conditional on the selection event. Tibshirani et al. (2016) carried out valid inferences after the forward stepwise regression, least angle regression, and the lasso selection events. Tian et al. (2018) considered a similar problem with unknown noise level in the regression model. This approach is dependent on distributional assumptions and is applicable only when the model is selected based on a prespecified formal selection procedure, such as forward selection and lasso.

An alternative approach is to develop uniformly valid inferential procedures that can be used after arbitrary model selection. Similarly, we require that the confidence interval satisfies

$$P(\beta_{\mathcal{K}^*,j} \notin C_{\mathcal{K}^*,j} \text{ for any } j \in \mathcal{K}^*) \leq \alpha.$$

This is different from the conditional inference methods since the confidence interval is universally valid after any model selection procedures. Berk et al. (2013) constructed post-selection inference (PoSI) for linear regression coefficients in the submodel (1.3), the simultaneous inference is valid under all possible model selection events instead of a particular selection procedure such as lasso. This approach is later extended to the problem of prediction (Bachoc et al., 2019) and to misspecified non-linear settings (Bachoc et al., 2020). These PoSI-based methods can be computationally NP-hard. Kuchibhotla et al.

(2020) provided computationally efficient inferences for coefficients that are valid under arbitrary model selection and allow misspecification of the regression model. Such procedures are based on uniform tail probability inequalities and thus are often conservative.

1.5 Outline of Dissertation

In this dissertation, we focus on the association between partially observed genomic features and outcomes of interest in genomic studies. We develop new statistical testing methods that integrate multiple data types to detect relevant risk factors, and investigate and understand their relationships with the disease outcomes. The remaining of the dissertation is organised as follows.

In Chapter 2, we consider the association test problem between a phenotype and an incomplete covariate, where the incomplete covariate may be associated with potentially high-dimensional auxiliary variables. We consider a MAR scenario, where the missing mechanism may depend on the outcome of interest and observed covariates, and a complete-case analysis or a single imputation approach is generally invalid. We develop a two-step test procedure that integrates high-dimensional auxiliary variables and identifies important genomic features associated with the phenotype outcome. We model the missing covariate against a subset of auxiliary variables, and construct the score test for the covariate effect on the outcome variable. We show that the proposed test procedure, though derived by assuming a prespecified covariate model, is valid even when the selection event of the auxiliary variables is random.

In Chapter 3, we consider right-censored survival outcomes and extend the approach proposed in Chapter 2 to semiparametric outcome models. We specify the link between the survival outcome and the covariates using a transformation model, which includes the proportional hazards model and the proportional odds model as special cases. We propose a new testing method by considering multiple choices of the transformation functions to

improve test efficiency when the outcome model is unknown. We show that the type I error of the proposed test is preserved, even when the outcome model and/or the missing covariate model is misspecified. We establish the asymptotic normality of the proposed test statistic under some regularity conditions.

In Chapter 4, we make a brief summary of the dissertation and discuss some future research directions.

Chapter 2

Score Tests with an Incomplete Covariate in Parametric Regression Models

2.1 Model and the Post-Selection Score Test

Consider an outcome of interest Y, a covariate of interest S, a vector of other covariates $\boldsymbol{X}$, and a potentially high-dimensional vector of auxiliary variables $\boldsymbol{A}$. For example, in genomic studies, Y may represent a disease phenotype, S may represent a genomic variable of interest, $\boldsymbol{X}$ may represent clinical or demographic variables, and $\boldsymbol{A}$ may represent other types of genomic or environmental variables collected in the study. The vector of covariates $\boldsymbol{X}$ includes a constant component of 1. Assume that

$$Y \mid (\boldsymbol{X}, S) \sim F_Y(; \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S), \tag{2.1}$$

where $\boldsymbol{\alpha}$ and β are regression parameters, and F_Y is a distribution function such that for some known function $\mu(\cdot)$, $\mathrm{E}[\{Y - \mu(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)\}(\boldsymbol{X}^{\mathrm{T}}, S)^{\mathrm{T}}] = \mathbf{0}$ at the true values

of $\boldsymbol{\alpha}$ and β. This formulation includes as special cases the linear regression model, with $\mu(x) = x$, and the logistic regression model, with $\mu(x) = e^x/(1+e^x)$. The parameter β captures the effect of S on Y given $\boldsymbol{X}$. In cancer genomic studies, we typically set $\boldsymbol{X}$ to be clinical or demographic variables and do not include mediator variables in the effect of S on Y (such as downstream variables of S) in $\boldsymbol{X}$. In this case, β represents the total effect of S after accounting for clinical/demographic covariates. We do not assume an explicit model for S but allow an arbitrary association structure with $(\boldsymbol{X}, \boldsymbol{A})$. Because the major purpose of fitting the model of S is to predict missing S values, we can set $\boldsymbol{A}$ to be (potential) predictive variables of S.

Suppose that S may be missing, and let R be the indicator of whether S is observed. Specifically, $R = 1$ if S is observed, and $R = 0$ otherwise. We assume that R is conditionally independent of $(S, \boldsymbol{A})$ given $(Y, \boldsymbol{X})$. This missing mechanism is common in two-phase studies, where the outcome Y and basic covariates $\boldsymbol{X}$ are measured for all subjects in the first phase, and subjects with certain outcome or covariate values are selected to be measured for an expensive covariate S in the second phase. We do not allow R to depend directly on $\boldsymbol{A}$, because the auxiliary variables, though completely observed, may not be selected into the model of S. If R depends on a component of $\boldsymbol{A}$ that is associated with S and is not selected, then the missing mechanism becomes not at random. For a sample of size n, the observed data consist of $\{(Y_i, \boldsymbol{X}_i, \boldsymbol{A}_i, R_i S_i, R_i) : i = 1, \ldots, n\}$. The assumed relationships among these variables are illustrated in Figure 2.1.

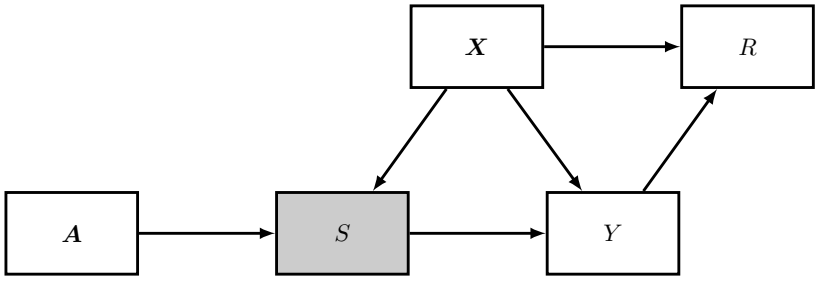


Figure 2.1: Relationships among the completely and incompletely observed variables.

We wish to test the null hypothesis $H_0 : \beta = 0$. Due to the presence of missing data, we propose to fit a working model of S on $(\boldsymbol{X}, \boldsymbol{A})$ and adopt the score test based on the full model (including both the models of Y and S). Fitting a working model of S against $(\boldsymbol{X}, \boldsymbol{A})$ allows us to utilize information about the missing S values contained in the auxiliary variables and is generally more efficient than ignoring the auxiliary variables. The score test is considered, instead of the Wald test and the likelihood-ratio test, because it only involves estimation under the null hypothesis, whereas the other two tests involve estimation under the alternative hypothesis. Note that estimation under the alternative hypothesis is more challenging because the likelihood generally involves an integration without a closed-form expression.

Because $\boldsymbol{A}$ is potentially high-dimensional, maximum likelihood estimation for the model of S may be infeasible. Also, the model of S is only of secondary interest, so full specification of the model may not be necessary. Therefore, we propose a two-step approach, where in the first step, we select a low-dimensional subset of $\boldsymbol{A}$ into the model of S, and in the second step, we perform a score test based on the model of Y and a working model of S.

In the first step, we perform variable selection on $\boldsymbol{A}$. Let $\mathcal{K}^*$ be a general model selection operator, such that for an m-vector of outcome variables $\mathcal{Y}$ and an $(m \times p)$-matrix of covariates $\mathcal{Z}$, $\mathcal{K}^*(\mathcal{Y}, \mathcal{Z}) : \mathbb{R}^m \times \mathbb{R}^{m \times p} \to \mathcal{C}_p$, where $\mathcal{C}_p$ is the collection of subsets of $\{1, \ldots, p\}$. For example, for marginal screening (Fan & Lv, 2008; Fan & Song, 2010) with a quantitative outcome variable and standardized $\mathcal{Z}$, $\mathcal{K}^*$ can be defined as $\mathcal{K}^* : (\mathcal{Y}, \mathcal{Z}) \mapsto \{j : |\mathcal{Y}^{\mathrm{T}} \mathcal{Z}_j| > \lambda\}$, where λ is a tuning parameter, and $\mathcal{Z}_j$ is the jth column of $\mathcal{Z}$. Likewise, for lasso (Tibshirani, 1996),

$$\mathcal{K}^* : (\mathcal{Y}, \mathcal{Z}) \mapsto \left\{ j : \widehat{\boldsymbol{\gamma}}_j \neq 0, \text{ where } (\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_p)^{\mathrm{T}} = \arg\min_{\boldsymbol{\gamma}} \left(\|\mathcal{Y} - \mathcal{Z}\boldsymbol{\gamma}\|^2 + \lambda \|\boldsymbol{\gamma}\|_1 \right) \right\}.$$

We use this operator to select a model for S based on the residual $S - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}} \boldsymbol{X}$ and $\boldsymbol{A}$, where

$\widehat{\boldsymbol{\gamma}}_X \equiv (\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}})^{-1} \sum_{i=1}^n R_i \boldsymbol{X}_i S_i$ is the least-squares estimator of S on $\boldsymbol{X}$ using the subjects with $R = 1$. The selected components of $\boldsymbol{A}$ are $\mathcal{K}^*(\mathcal{S} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}_X, \boldsymbol{A})$, where $\mathcal{S}$ is a vector that consists of $\{S_i : R_i = 1\}$, and $\boldsymbol{X}$ and $\boldsymbol{A}$ are matrices that consist of rows of $\{\boldsymbol{X}_i : R_i = 1\}$ and $\{\boldsymbol{A}_i : R_i = 1\}$, respectively. For simplicity of presentation, we write $\mathcal{K}^* = \mathcal{K}^*(\mathcal{S} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}_X, \boldsymbol{A})$ and let $\mathcal{K}$ be the observed value of $\mathcal{K}^*$.

Let $\boldsymbol{W}_{\mathcal{K}}$ denote the vector that consists of $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ indexed by $\mathcal{K}$. In the second step, we fit model (2.1) and the working model $S = \boldsymbol{\gamma}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}} + \delta$ under the null hypothesis H_0, where δ is a mean-zero error term, and $\boldsymbol{\gamma}_{\mathcal{K}}$ is a vector of regression parameters. In particular, we estimate $\boldsymbol{\gamma}_{\mathcal{K}}$ by $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}} \equiv (\sum_{i=1}^n R_i \boldsymbol{W}_{\mathcal{K},i} \boldsymbol{W}_{\mathcal{K},i}^{\mathrm{T}})^{-1} \sum_{i=1}^n R_i \boldsymbol{W}_{\mathcal{K},i} S_i$, the least-squares estimator using the subjects with observed S values. Let $\widehat{\boldsymbol{\alpha}}$ be the Z-estimator of $\boldsymbol{\alpha}$ under H_0, such that $\sum_{i=1}^n \{Y_i - \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{X}_i)\} \boldsymbol{X}_i = 0$. The (scaled) score statistic for β is

$$U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = \frac{1}{n^{1/2}} \sum_{i=1}^n \{Y_i - \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{X}_i)\} \{R_i S_i + (1 - R_i) \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}\}.$$

Note that this coincides with the imputation-based score statistic, that is, the score statistic when the missing values of S are imputed by the estimated mean $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}}$.

To obtain an asymptotic size α test, we need to derive the asymptotic distribution of $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})$ under H_0. This is highly challenging, because the model selection event $\{\mathcal{K}^* = \mathcal{K}\}$ is random, and the usual arguments based on the Taylor's series expansion of the score statistic do not apply. Nevertheless, as we establish in Section 2.2, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})$, properly scaled by a variance term that can be consistently estimated by an empirical sum-of-squares estimator, is asymptotically normal. In particular, the variance term resembles that derived based on the usual Taylor's series expansion on the score statistic. Let $\boldsymbol{\alpha}_0$ be the true value of $\boldsymbol{\alpha}$, and for a given selected model $\mathcal{K}$, define $\boldsymbol{\gamma}_{0\mathcal{K}} \equiv \arg\min_{\boldsymbol{\gamma}} \mathrm{E}\{R(S - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}})^2\}$ as the true value of $\boldsymbol{\gamma}_{\mathcal{K}}$. Let $\boldsymbol{I}_{\alpha\alpha} = \mathrm{E}\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \boldsymbol{X} \boldsymbol{X}^{\mathrm{T}}\}$, $\boldsymbol{I}_{\beta\alpha} = -\mathrm{E}[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \boldsymbol{X} \{RS + (1 - R) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}}\}]$, $\boldsymbol{I}_{\gamma\gamma} = \mathrm{E}(R \boldsymbol{W}_{\mathcal{K}} \boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})$,

and $\boldsymbol{I}_{\beta\gamma} = \mathrm{E}[\{Y - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\}(1-R)\boldsymbol{W}_{\mathcal{K}}]$, where μ' denotes the first derivative of μ. If the model $\mathcal{K}$ is prespecified, then the Taylor's series expansion of $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ at $(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_{0\mathcal{K}})$ yields

$$U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[\{Y_i - \mu(\alpha_0^{\mathrm{T}}\boldsymbol{X}_i)\}\big\{R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\} + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\Big] + o_p(1) \tag{2.2}$$

under regularity conditions. Based on this expansion, we can estimate the asymptotic variance of $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ by $\widehat{\sigma}^2(\mathcal{K}) = n^{-1}\sum_{i=1}^{n}\{\widehat{\sigma}_i(\mathcal{K}) - \overline{\sigma}(\mathcal{K})\}^2$, where

$$\widehat{\sigma}_i(\mathcal{K}) = \{Y_i - \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{X}_i)\}\big\{R_i S_i + (1-R_i)\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\} + \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}R_i(S_i - \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}),$$

$\overline{\sigma}(\mathcal{K}) = n^{-1}\sum_{i=1}^{n}\widehat{\sigma}_i(\mathcal{K})$, and $\widehat{\boldsymbol{I}}_{\alpha\alpha}$, $\widehat{\boldsymbol{I}}_{\beta\alpha}$, $\widehat{\boldsymbol{I}}_{\beta\gamma}$, and $\widehat{\boldsymbol{I}}_{\gamma\gamma}$, are the empirical counterparts of $\boldsymbol{I}_{\alpha\alpha}$, $\boldsymbol{I}_{\beta\alpha}$, $\boldsymbol{I}_{\beta\gamma}$, and $\boldsymbol{I}_{\gamma\gamma}$, respectively, with the expectations replaced by empirical means and true parameters replaced by estimators. We can show that even though this variance term is derived based on fixed $\mathcal{K}$, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})/\widehat{\sigma}(\mathcal{K}^*)$ converges to the standard normal distribution under H_0. Therefore, for an asymptotic size α test, we reject H_0 if $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})^2/\widehat{\sigma}^2(\mathcal{K}^*) > \chi^2_{1,\alpha}$.

The proposed test does not require correct specification of the models of Y and S. For the outcome model, we only require that $\mathrm{E}[\{Y - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\}(\boldsymbol{X}^{\mathrm{T}}, S)^{\mathrm{T}}] = \boldsymbol{0}$ under the null hypothesis, because an empirical variance estimator is used instead of a model-based estimator. For the covariate model, as detailed in Section 2.2, we require the association structure between S and $\boldsymbol{X}$ to be correctly specified but allow arbitrary association between S and $\boldsymbol{A}$; correct specification of the association between S and $\boldsymbol{X}$ is generally needed (Derkach et al., 2015; Lawless, 2018). The association structure between S and

$\boldsymbol{A}$ affects the power of the test but not its validity under the null hypothesis.

2.2 Asymptotic Properties of the Post-Selection Score Test

For any $\mathcal{K}$, let $\boldsymbol{\gamma}_{0X}$ and $\boldsymbol{\gamma}_{0A,\mathcal{K}}$ be the subvectors of $\boldsymbol{\gamma}_{0\mathcal{K}}$ that correspond to $\boldsymbol{X}$ and the selected components of $\boldsymbol{A}$, respectively. Define

$$\begin{aligned}
\sigma_1^2(\mathcal{K}) &= \mathrm{Var}\big[\epsilon\big\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}\big\}\big] \\
\sigma_2^2(\mathcal{K}) &= \mathrm{Var}\Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\big\{\mathrm{E}(\epsilon \mid R, \boldsymbol{X})\boldsymbol{X} - \mathrm{E}(\epsilon\boldsymbol{X} \mid R)\big\} \\
&\qquad + \big\{\mathrm{E}(\epsilon \mid R, \boldsymbol{X}) - \mathrm{E}(\epsilon \mid R)\big\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K}} \\
&\qquad + \big\{\mathrm{E}(\epsilon \mid R, \boldsymbol{X}) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K}}\big\}R(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\Big] \\
\sigma_3^2(\mathcal{K}) &= \mathrm{Var}\Big\{(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\mathrm{E}(\epsilon\boldsymbol{X} \mid R) + \mathrm{E}(\epsilon \mid R)\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K}}\Big\},
\end{aligned}$$

where $\epsilon = Y - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})$, and let $\sigma^2(\mathcal{K}) = \sum_{k=1}^{3}\sigma_k^2(\mathcal{K})$. Let $\|\cdot\|_{\psi_\xi}$ be an Orlicz norm, such that $\|X\|_{\psi_\xi} = \inf\{\eta > 0 : \mathrm{E}(e^{|X|^\xi/\eta^\xi}) \le 2\}$, and $\|\cdot\|$ be the Euclidean norm. We assume the following conditions. Some conditions involve a generic positive constant M.

(C1) For some $\xi \in (0, 2]$, $\|Y\|_{\psi_\xi} + \|S\|_{\psi_\xi} + \max_j \|A_j\|_{\psi_\xi} < M$. The covariate $\boldsymbol{X}$ is bounded, so that $P(\|\boldsymbol{X}\| < M) = 1$. Also, the estimator $\widehat{\boldsymbol{\alpha}}$ is strongly consistent under $\beta = 0$, $\mu(\cdot)$ is twice continuously differentiable, and $\lambda_{\min}[\mathrm{E}\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}] > M^{-1}$, where $\lambda_{\min}(C)$ denotes the minimum eigenvalue of the matrix C.

(C2) There exists a sequence of collections of models Ω_n, such that $P(\mathcal{K}^* \in \Omega_n) \to 1$, $\sup_{\mathcal{K}\in\Omega_n}|\mathcal{K}| = O(n^\tau)$, and $\log|\Omega_n| = O(n^\kappa)$, where τ and κ are constants that satisfy $\tau < 4\xi/(5\xi+12)$, $5\tau/4 + 3\kappa/\xi < 1$, and $\tau + 4\kappa/\xi < 1$, and $|\mathcal{C}|$ denotes the cardinality of the set $\mathcal{C}$. Also, $\inf_{\mathcal{K}\in\Omega_n}\lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})\} > M^{-1}$, $\sup_{\mathcal{K}\in\Omega_n}\mathrm{E}\{(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^4\} <$

M, and $\inf_{\mathcal{K}\in\Omega_n} \sigma^2(\mathcal{K}) > M^{-1}$.

(C3) The probability $P(R = 1 \mid Y, \boldsymbol{X}) > M^{-1}$ almost surely.

(C4) Under $\beta = 0$, the residual $(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$ and $\boldsymbol{X}$ are independent, and $\boldsymbol{A}$ is independent of $(Y, \boldsymbol{X})$.

(C5) The models selected based on the estimated residuals $(S_i - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ and the actual residuals $(S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ are such that

$$P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) \neq \mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A})\Big\} = o(1)$$

and

$$\sup_{\mathcal{K}\in\Omega_n} \frac{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) = \mathcal{K}\Big\}}{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A}) = \mathcal{K}\Big\}} < M.$$

(C6) For a random sample of size m, let $\widetilde{\mathcal{S}} = (S_1, \ldots, S_m)^{\mathrm{T}}$, $\widetilde{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)^{\mathrm{T}}$, and $\widetilde{\mathcal{A}} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)^{\mathrm{T}}$. The random variable

$$\sup_{\mathcal{K}\in\Omega_m} \left| \frac{P\{\mathcal{K}^*(\widetilde{\mathcal{S}} - \widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X}, \widetilde{\mathcal{A}}) = \mathcal{K} \mid \widetilde{\mathcal{A}}\}}{P\{\mathcal{K}^*(\widetilde{\mathcal{S}} - \widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X}, \widetilde{\mathcal{A}}) = \mathcal{K}\}} - 1 \right|$$

converges to 0 in mean as $m \to \infty$.

Remark 2.1. Condition (C1) imposes constraints on the tail probabilities of the observed variables. With $\xi = 1$ or $\xi = 2$, we assume each component of $(Y, S, \boldsymbol{A})$ to be sub-exponential or sub-Gaussian, respectively. To maintain a flexible model for Y, we assume that $\boldsymbol{X}$ is bounded. Desired theoretical results could be obtained by only requiring $\max_j \|X_j\|_{\psi_\xi} < M$, but additional conditions on μ would be required. Condition (C2) allows the set of "possibly-selected models" Ω_n to grow exponentially with n and the size of the selected model to increase at a polynomial rate of n. For example, for $\xi =$

2, we allow $\sup_{\mathcal{K}\in\Omega_n}|\mathcal{K}| = O(n^{1/4})$ and $|\Omega_n| = O\{\exp(n^{1/4})\}$. Note that if the model selection procedure yields consistent selection, then Ω_n can be chosen to be a singleton set, consisting only of the true model. In our setting, we allow the model selection event to be genuinely random even when n increases to infinity. Condition (C3) ensures that a nonvanishing portion of subjects have observed S.

Remark 2.2. Condition (C4) requires that S exhibits a linear association structure with $\boldsymbol{X}$ and that $(Y, \boldsymbol{X})$ are independent of the auxiliary variables. This guarantees that $(Y, \boldsymbol{X})$ are independent of the model selection event, which is based on the residuals in the model of S and the auxiliary variables. In cancer genomic studies where $\boldsymbol{X}$ represents demographic variables and $\boldsymbol{A}$ represents genomic variables (such as gene expressions in the tumor), $\boldsymbol{X}$ and $\boldsymbol{A}$ are plausibly independent. In general, because $\boldsymbol{X}$ is low-dimensional, the independence between $\boldsymbol{A}$ and $\boldsymbol{X}$ can be (approximately) achieved by projecting components of $\boldsymbol{A}$ to the orthogonal complement of the span of $\boldsymbol{X}$ or functions of $\boldsymbol{X}$. The independence between $\boldsymbol{A}$ and Y can be relaxed to allow some auxiliary variables that are not associated with S to depend on Y; the technical formulation of the relaxed condition is deferred to Appendix 1. For marginal screening, the relaxed condition allows the auxiliary variables not in any models in Ω_n to depend on Y (and $\boldsymbol{X}$). Requiring the (potentially) selected auxiliary variables to be independent of Y is quite reasonable under the null hypothesis, because these variables are generally associated with S. If they are also associated with Y, then except at some specific parameter values, S and Y would be marginally associated, and the null hypothesis does not hold.

Remark 2.3. Conditions (C5) and (C6) impose mild conditions on the model selection operator. Condition (C5) requires that the model selected based on the estimated residuals and that selected based on the actual residuals are asymptotically equal. This is easily satisfied, because the least-squares estimator $\widehat{\boldsymbol{\gamma}}_X$ is consistent. Condition (C6) requires that the marginal probability of selecting a model is asymptotically equal to the condi-

tional probability of the same event given the auxiliary variables. This is true of common model selection operators, which select a model based on the association between the outcome and the covariates, and the covariates alone do not contain information about the model selection event. We discuss the verification of these conditions under a marginal screening procedure in Section 2.6.2.

We impose conditions on the number of possibly-selected models instead of the total number of auxiliary variables, because the former is directly relevant to the asymptotic distribution of the score statistic. Nevertheless, for a given maximal selected model size $q_n \equiv \sup_{\mathcal{K}\in\Omega_n} |\mathcal{K}|$, we have

$$r_n \equiv |\Omega_n| \leq \sum_{s=1}^{q_n} \binom{p_n}{s} \leq \left(\frac{ep_n}{q_n}\right)^{q_n},$$

where p_n is the total number of auxiliary variables. The condition on r_n is satisfied if $\log p_n = O(n^{\kappa-\tau})$, with κ and τ satisfying the inequalities in condition (C2). In fact, if most auxiliary variables are only weakly associated with S, then r_n could be much smaller than the above upper bound.

We have the following results.

Theorem 2.1. *Under conditions (C1)–(C6) and H_0, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})/\sigma(\mathcal{K}^*)$ converges weakly to the standard normal distribution.*

Theorem 2.2. *Under conditions (C1)–(C6) and H_0,*

$$\mathrm{E}\Big\{ \sup_{\mathcal{K}\in\Omega_n} |\widehat{\sigma}^2(\mathcal{K}) - \sigma^2(\mathcal{K})| \Big\} = o(1).$$

Remark 2.4. Theorem 2.1 states that the scaled score statistic, which is derived from a randomly selected model, converges in distribution to a standard normal distribution marginally. A key step in the proof is to show that the score statistic can be (asymp-

totically) written as a sum of independent variables that are mean zero conditional on the model selection event and possibly other components of the observed data. Then, we can employ the Lindeberg approach to the proof of the central limit theorem to establish the desired result. Theorem 2.2 states that the scaling term of the score statistic in Theorem 2.1 can be uniformly consistently estimated by the proposed sum-of-squares estimator over the set of possibly-selected models Ω_n.

Combining the above results, we have the following corollary.

Corollary 2.1. *Under conditions (C1)–(C6) and H_0, $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})/\widehat{\sigma}(\mathcal{K}^*)$ converges weakly to the standard normal distribution.*

We outline the proof of Theorem 2.1 here and relegate the complete proofs of Theorems 2.1 and 2.2 to Section 2.6.5. By a version of the portmanteau theorem (Pollard, 2002, p. 177), it suffices to prove that for any function g with bounded derivatives up to the third order,

$$\mathrm{E}\Big[g\Big\{\frac{U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})}{\sigma(\mathcal{K}^*)}\Big\}\Big] \to \mathrm{E}\{g(Z)\}, \tag{2.3}$$

where Z is a standard normal variable. The first step of the proof is to expand $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ as

$$\begin{aligned}
&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\big\{\epsilon_i - \mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\big\}\big\{R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\} \\
&+ \frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\boldsymbol{X}_i - \mathrm{E}(\epsilon\boldsymbol{X} \mid R_i)\big\} \\
&+ \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) - \mathrm{E}(\epsilon \mid R_i)\big\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i} + \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}\big\}R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\Big] \\
&+ \frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big\{(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\mathrm{E}(\epsilon\boldsymbol{X} \mid R_i) + \mathrm{E}(\epsilon \mid R_i)\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\Big\} + o_p(1) \\
\equiv &\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{1i}(\mathcal{K}) + \frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{2i}(\mathcal{K}) + \frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{3i}(\mathcal{K}) + o_p(1),
\end{aligned}$$

where the $o_p(1)$ terms converge in mean to zero uniformly over $\mathcal{K} \in \Omega_n$. As a result, the left-hand side of (2.3) can be written as

$$\int_{\mathcal{K}\in\Omega_n} \mathrm{E}\Big[g\Big\{n^{-1/2}\sum_{i=1}^{n}\frac{U_{1i}(\mathcal{K})+U_{2i}(\mathcal{K})+U_{3i}(\mathcal{K})}{\sigma(\mathcal{K})}\Big\}\Big|\mathcal{K}^*=\mathcal{K}\Big]\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K})+o(1), \qquad (2.4)$$

where $\mathcal{P}_{\mathcal{K}^*}$ is the probability measure of $\mathcal{K}^*$.

The main argument of the proof is to show that $n^{-1/2}\sum_{i=1}^{n} U_{ki}(\mathcal{K})$ for $k = 1, 2, 3$ in (2.4) can in turn be replaced by normal variables. Note that conditional on $\mathcal{O}_1 \equiv (R_i, S_i, \boldsymbol{W}_{\mathcal{K},i})_{i=1,\ldots,n}$, $U_{11}(\mathcal{K}), \ldots, U_{1n}(\mathcal{K})$ are mean zero and independent. For $i = 1, \ldots, n$, let

$$\widetilde{U}_{1i}(\mathcal{K}) = \mathrm{Var}(\epsilon \mid R_i, \boldsymbol{X}_i)^{1/2}\big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\}Z_{1i},$$

where $Z_{11}, \ldots, Z_{1n}$ are i.i.d. standard normal random variables that are independent of the observed data. Because the first and second moments of U_{1i} and $\widetilde{U}_{1i}$ given $\mathcal{O}_1$ match and $\{\mathcal{K}^* = \mathcal{K}\}$ is implied by $\mathcal{O}_1$, the moments given $\{\mathcal{K}^* = \mathcal{K}\}$ also match. We then use Lindeberg's telescoping argument for the central limit theorem (Chung, 2001, p. 211) to show that $n^{-1/2}\sum_{i=1}^{n} U_{1i}(\mathcal{K})$ in (2.4) can be replaced by $n^{-1/2}\sum_{i=1}^{n}\widetilde{U}_{1i}(\mathcal{K})$. We further show that the term can be replaced by a normal variable with mean zero and variance $\sigma_1^2(\mathcal{K})$.

Next, we show that under condition (C5), the event $\{\mathcal{K}^* = \mathcal{K}\}$ in the conditional expectation in (2.4) can be replaced by $\{\mathcal{K}_0^* = \mathcal{K}\}$, where $\mathcal{K}_0^* \equiv \mathcal{K}_0^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A})$ is the selected model based on the actual residual $(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$. Then, we note that $\{\mathcal{K}_0^* = \mathcal{K}\}$ is implied by $\mathcal{O}_2 \equiv (R_i, \boldsymbol{A}_i, S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)_{i=1,\ldots,n}$, and conditional on $\mathcal{O}_2$, $U_{21}(\mathcal{K}), \ldots, U_{2n}(\mathcal{K})$ are mean zero and independent; under this conditional probability space, the random element in $U_{2i}(\mathcal{K})$ is $\boldsymbol{X}_i$. We can similarly show that $n^{-1/2}\sum_{i=1}^{n} U_{2i}(\mathcal{K})$ in (2.4) can be replaced by a normal variable with mean zero and variance $\sigma_2^2(\mathcal{K})$.

Finally, we show that after $n^{-1/2}\sum_{i=1}^{n} U_{1i}(\mathcal{K})$ and $n^{-1/2}\sum_{i=1}^{n} U_{2i}(\mathcal{K})$ are replaced by normal variables, the conditional expectation in (2.4) can be replaced by a marginal expectation under condition (C6). It is easy to see that $U_{31}(\mathcal{K}), \ldots, U_{3n}(\mathcal{K})$ are mean zero and independent, and thus $n^{-1/2}\sum_{i=1}^{n} U_{3i}(\mathcal{K})$ can be replaced by a normal variable with mean zero and variance $\sigma_3^2(\mathcal{K})$. Combining the above results, we conclude that the variable in the function g in (2.4) can be replaced by a standard normal variable, and the desired result follows.

In the conventional argument for the asymptotic distribution of the score statistic, we expand $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ as in (2.2), and the asymptotic normality of the score statistic (given $\boldsymbol{X}$ and $\boldsymbol{A}$) follows from the central limit theorem. However, conditional on the model selection event $\{\mathcal{K}^* = \mathcal{K}\}$, $(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i})_{i=1,\ldots,n}$ are dependent, and the central limit theorem does not apply. In our proof, instead of relying on the independence of $(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i})$'s, we establish the asymptotic normality based on the (conditional) independence and mean-zero property of functions of $\boldsymbol{X}_i$'s given the model selection event.

2.3 Simulation Studies

Let $\boldsymbol{X} = (X_1, \ldots, X_5)^{\mathrm{T}}$, where (X_1, X_2, X_3) are mean-zero multivariate normal variables with $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$ $(j, k = 1, 2, 3)$, $X_4 \sim$ Bernoulli(0.1), $X_5 \sim$ Bernoulli(0.2), and X_4 and X_5 are independent of each other and (X_1, X_2, X_3). Let $\boldsymbol{A}$ be a p-vector of independent standard normal variables. We set $S = \boldsymbol{\gamma}_X^{\mathrm{T}} \boldsymbol{X} + \boldsymbol{\gamma}_A^{\mathrm{T}} \boldsymbol{A} + \boldsymbol{\gamma}_{A,2}^{\mathrm{T}} \boldsymbol{A}^2 + \delta$, where $\boldsymbol{A}^2$ is a p-vector of the squared components of $\boldsymbol{A}$, δ is standard normal, $\boldsymbol{\gamma}_X = (0.1, \ldots, 0.1)^{\mathrm{T}}$, and $\boldsymbol{\gamma}_{A,2}$ is 0.1 at the first 5 components and 0 elsewhere. We considered two different values of $\boldsymbol{\gamma}_A$. In Setting 1, we set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components and 0 at the remaining components, whereas in Setting 2, we set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components, 0.02 at the subsequent 80 components, and 0 at the remaining components. In Setting 1, the model is sparse, and a small number of auxiliary variables have strong

effects on S. In Setting 2, the model contains a mixture of strong and weak signals from the auxiliary variables.

We considered a quantitative and a binary outcome variable Y. For the quantitative outcome, we set $Y = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}+\beta S+\epsilon$, where ϵ is standard normal, and $\boldsymbol{\alpha} = (1, -1, 1, -1, 1)^{\mathrm{T}}$. For the binary outcome, we set $\mathrm{logit}\{P(Y = 1 \mid \boldsymbol{X}, S)\} = -2.2 + \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S$, where $\boldsymbol{\alpha}$ is the same as that under the linear model; the proportion of subjects with $Y = 1$ is about 15–20%. We considered two missing-data mechanisms. The first mechanism is missing completely at random (MCAR), where the missing-data status is independent of other variables. The second mechanism is missing at random (MAR), where for the quantitative outcome, an equal number of subjects at the two extreme tails of the distribution of Y were selected to have observations on S, whereas for the binary outcome, all subjects with $Y = 1$ were selected, and a fraction of subjects with $Y = 0$ were selected to attain the desired missing proportion. We considered sample sizes of $n = 500$ and 1000 and numbers of auxiliary variables of $p = 200$, 500, 1000, 1500, and 2000. For the alternative hypothesis, we set $\beta = 2n^{-1/2}$ and $6n^{-1/2}$ for the quantitative and binary outcome variables, respectively. For each setting, we simulated 100,000 and 10,000 replicates for $\beta = 0$ and $\beta \neq 0$, respectively.

We compare the performance of five tests: (1) the standard score test using complete data only; (2) the standard score test with missing data imputed under a working linear model of S on $\boldsymbol{X}$ and components of $\boldsymbol{A}$ selected using marginal screening, where a component of $\boldsymbol{A}$ is selected if its absolute empirical correlation with $S - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}$ among the subjects with complete data is larger than a certain threshold; (3) the score test based on the full likelihood with a working linear model of S against $\boldsymbol{X}$ alone; (4) the proposed test, where the working model of S is selected in the same way as (2); and (5) the score test based on the full likelihood with a linear model of S against $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ that are associated with S. We refer to methods (1)–(5) as the complete-case anal-

ysis, the simple imputation method, the covariate-only method, the proposed method, and the true model method, respectively. In the simple imputation, proposed, and true model methods, only first-order terms of $\boldsymbol{A}$ are in the working models, so none of the models is "correct". Nevertheless, according to our theory, the proposed method is still valid under such misspecification. For the simple imputation and proposed methods, the threshold for screening is selected using BIC. For the covariate-only and true model methods, the variance of the score statistic is estimated using the proposed empirical sum-of-squares estimator instead of the usual estimator based on the second derivative of the log-likelihood. This is for ease of comparison with the proposed method, and the two variance estimators are asymptotically equivalent. The true model method is a gold standard but is not practical, because it requires knowledge of the relevant predictors of S.

The results under a missing proportion of 60% are plotted in Figures 2.2 and 2.3, and the results under a missing proportion of 30% are plotted in Figures 2.4 and 2.5 in Section 2.6.6; the power of methods that inflate the type I error is not presented. The significance level is set to be 0.05. Under missing at random and the linear outcome model, both the complete-case analysis and simple imputation method inflate the type I error, because they underestimate the variance of the score statistic. The covariate-only method and the true model method preserve the type I error; they do not involve model selection, and their validity follows from a conventional argument. The proposed method, despite involving model selection variability, preserves the type I error; in fact, under Setting 2, any given model is selected at most 0.006% and 3.804% of the times over all simulation replicates with sample size 500 and 1000, respectively. The pattern of results under missing at random and the binary outcome model are similar, but the complete-case analysis preserves the type I error due to the validity of inference based on the prospective likelihood under a case-control study and the logistic regression model

(Prentice & Pyke, 1979). Under missing completely at random, all methods preserve the type I error.

Under the alternative hypothesis, the simple imputation method under missing at random has relatively high power due to underestimation of the variance of the score statistic; this is similar for the complete-case analysis under missing at random and the logistic outcome model. Under settings where the complete-case analysis preserves the type I error, the complete-case analysis and the covariate-only method have similar power, because both methods do not incorporate information of the auxiliary variables. As expected, the proposed method utilizes information about the missing data contained in the auxiliary variables and tends to yield higher power than the covariate-only method. The power gain from the incorporation of auxiliary variables can be small or even negative when the number of auxiliary variables p is much larger than the (effective) sample size $\sum_{i=1}^{n} R_i$. In this case, the variable selection procedure cannot effectively identify the relevant auxiliary variables. This results in the inclusion of many noise variables into the working model of S, which in turn results in a worse fit than the covariate-only model that has no noise variables.

The true model method tends to have high power, because it uses the true model of S. Nevertheless, it is less powerful than the proposed method in some scenarios under Setting 2. This is because the true model contains many auxiliary variables with weak signals, and the extra information contained in the variables does not compensate the variability involved in the estimation of their effects. This illustrates that even when the true model is known, it may be desirable to perform variable selection and retain only the variables with strong signals.

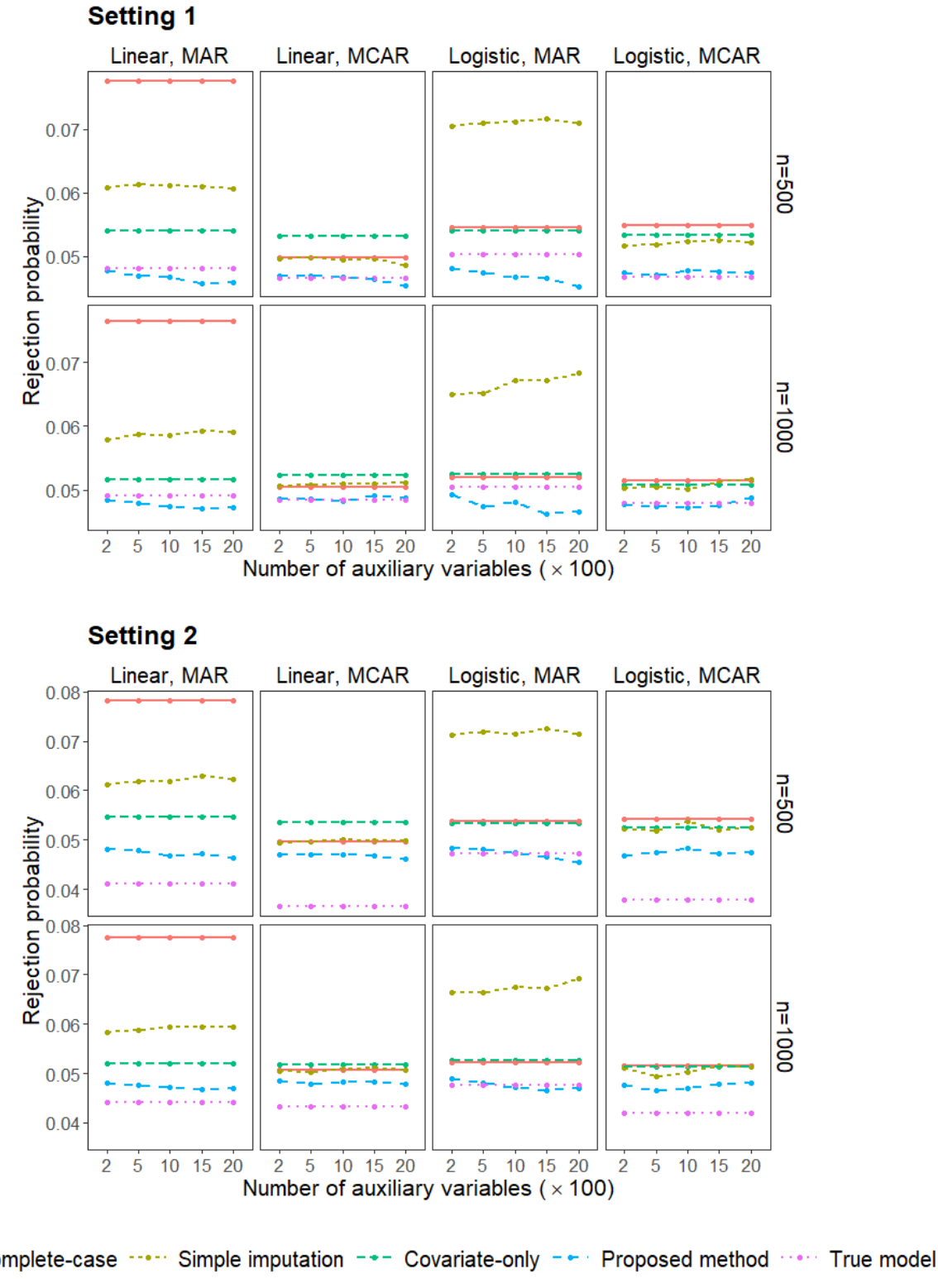


Figure 2.2: Rejection probabilities under a missing proportion of 60% and the null hypothesis.

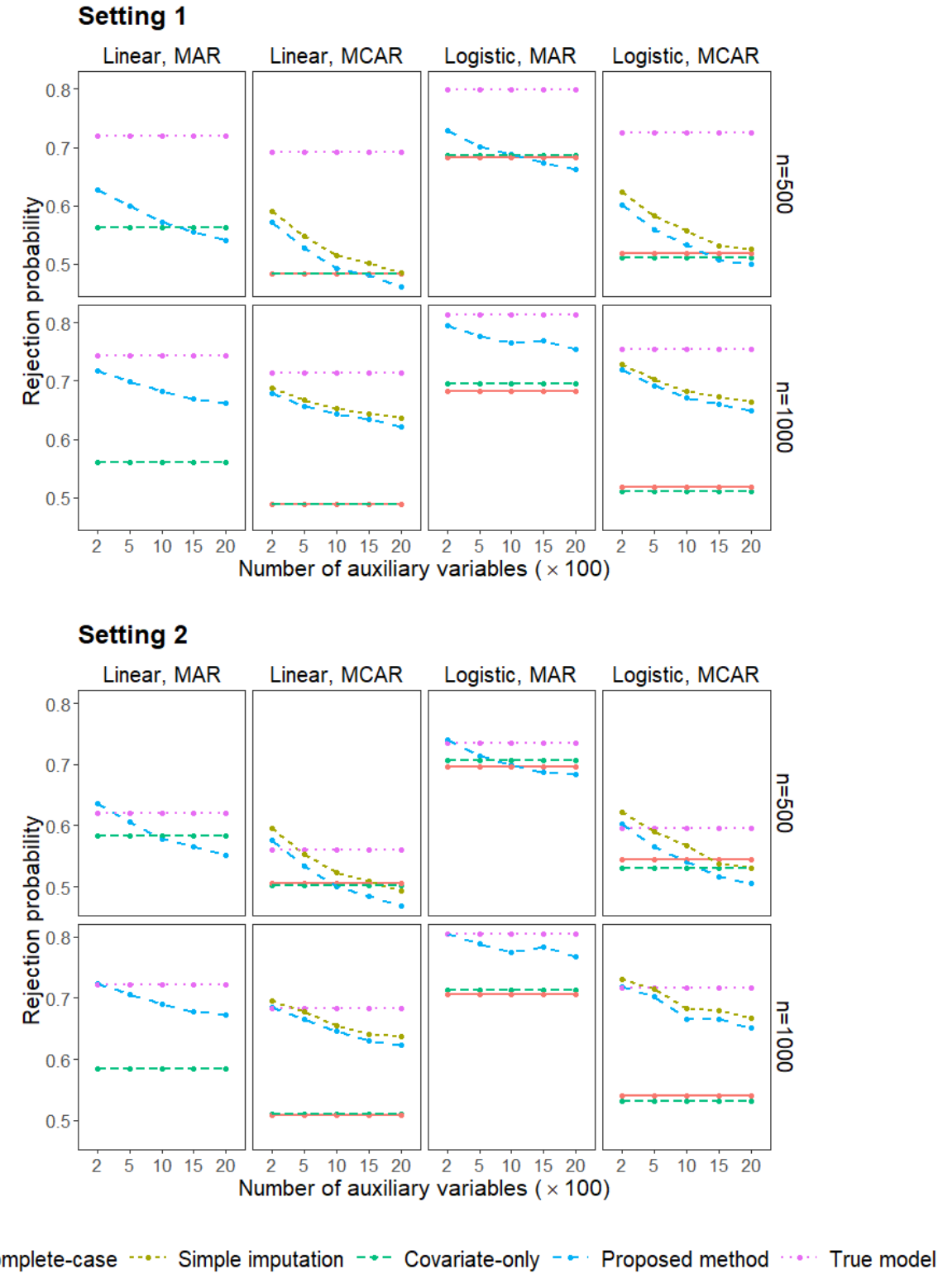


Figure 2.3: Rejection probabilities under a missing proportion of 60% and the alternative hypothesis.

2.4 A Real Study

We analyzed a dataset of patients with colorectal adenocarcinoma from TCGA (The Cancer Genome Atlas Network, 2012), available at http://gdac.broadinstitute.org/. In the study, demographic and clinical data, including age at diagnosis, sex, and tumor stage, as well as genomic data, including the expressions of RNA and protein, were measured. After removing subjects with missing clinical data, the sample size is 600. The expressions of 18,068 genes, measured by RNA sequencing, are available for most subjects. The expressions of 204 proteins or phospho-proteins are available for only 78.2% of the subjects.

We focused on the association between individual protein expressions and tumor stage. We set the outcome variable to be tumor stage, dichotomized into stage I/II and stage III/IV, with respective proportions of 0.56 and 0.44. In a single analysis, we set the covariate of interest S to be the expression of a protein or phospho-protein. We set sex and age at diagnosis to be the covariates in $\boldsymbol{X}$ and set gene expressions as auxiliary variables. In the resulting model, β represents the association between a protein and tumor stage for subjects with given age and sex. Note that the auxiliary variables are plausibly independent of $\boldsymbol{X}$, as required by condition (C4). The gene expression data are incomplete, and we impute the missing values using k-nearest neighbor imputation with $k = 10$. We set the auxiliary variables $\boldsymbol{A}$ to be the top 200 principal components of the gene expressions; they appear to be more predictive than the individual gene expressions. We performed the proposed test with the working model of S selected by the correlation-based marginal screening procedure in the simulation studies, and the screening threshold was selected by BIC. For comparison, we performed the score test using complete data only and the covariate-only method described in the simulation studies.

A total of 46 proteins were identified to be significantly associated with tumor stage at $\alpha = 0.05$ under at least one of the three tests. Among the significant proteins, 76% have smaller p-values under the proposed method than the complete-case analysis, and 78%

have smaller p-values under the proposed method than the covariate-only method. Many of the proteins that are more significant under the proposed method have been identified to be related to the progression of colorectal adenocarcinoma; the significant proteins and some relevant references are given in Table 2.1 in Section 2.6.6. This suggests that the proposed method is more powerful than the other two methods.

To investigate whether the power gain stems from the auxiliary variables, we inspect the relationship between the significance level and the variation explained by the gene expressions in the protein models. For a given protein, we let Z_1 and Z_2 be the indicators of whether the proposed method yields a smaller p-value than the complete-case analysis and the covariate-only method, respectively. Let R^2 be the coefficient of partial determination of the gene expressions, that is, the percentage of variation explained by the gene expressions given that sex and age are in the model. Among the significant proteins, the sample correlation between Z_1 and R^2 and that between Z_2 and R^2 are 0.32 and 0.22, respectively. In addition, we classify each protein into one of two groups based on whether it is more significant under the proposed method than the complete-case analysis. Then, we test the difference in mean of R^2 between the two groups using the two-sample Wilcoxon test, and the p-value is 0.0381. A similar analysis comparing the proposed method and the covariate-only method yields a p-value of 0.1271. The results suggest that proteins with better fit of the imputation model tend to have higher power gain, especially when compared with the complete-case analysis.

2.5 Discussion

In this chapter, we consider the association test between an outcome variable and an incomplete covariate, where the missing covariate values could be imputed using high-dimensional auxiliary variables. We propose a simple two-step procedure that does not involve accounting for the variability of model selection in the first step and prove that such

a procedure is asymptotically valid. This is in contrast with the conventional statistical intuition that standard inferential procedures on selected models are invalid and proper adjustments are needed (Fithian et al., 2014; Lee et al., 2016). In the current setting, the model that involves variable selection is only of secondary interest. Although the fit of this model would affect the power of the test, the variability of model selection does not affect the asymptotic distribution of the score statistic.

A linear working model is assumed for the incomplete covariate S, but the validity of the score test does not depend on the correctness of this model. In fact, as demonstrated in the simulation studies, a simple working model may yield higher power than the true model when the latter is complex and involves many unknown parameters. Nevertheless, we require S to exhibit a linear association with the low-dimensional covariates $\boldsymbol{X}$ in the outcome model. To relax this assumption, one may instead assume a nonparametric association between S and $\boldsymbol{X}$ (Derkach et al., 2015).

We focus on the asymptotic property of the score test under the null hypothesis. Evaluation of the asymptotic power of the test under contiguous alternatives is highly challenging, because the power depends on specifics of the model selection operator. The evaluation is even more complicated when R depends on Y, in which case the missing mechanism for the data on $(S, \boldsymbol{X}, \boldsymbol{A})$ is not at random. To provide some insights to the power gain from the auxiliary variables, we evaluate the power under prespecified, fixed-dimensional sets of auxiliary variables in Section 2.6.3. Under missing completely at random, inclusion of more auxiliary variables always increases the (asymptotic) power. Although the power generally does not have a simple form under missing at random, numerical evaluations suggest that the power tends to increase with the number of auxiliary variables. One should note, however, that these results are asymptotic and may not apply when the number of auxiliary variables is large compared to the sample size.

One possible extension is by considering more general outcome models. In this work,

we consider a parametric model of the phenotype outcome. However, in cancer genomic studies such as TCGA, some outcomes of interest are (possibly censored) time to events, such as time to cancer progression or death. It is of interest to consider semiparametric survival models for univariate or recurrent event times.

2.6 Technical Details and Additional Results

2.6.1 Relaxation of Condition (C4)

Let $\mathcal{M}_n \equiv \{j : j \in \mathcal{K} \text{ for some } \mathcal{K} \in \Omega_n\}$ be the collection of all "possibly selected" auxiliary variables and $\mathcal{M}_n^{\mathrm{C}}$ be its complement. Let $\mathcal{S}$ and $\mathcal{X}$ be the vector or matrix of the values of S_i and $\boldsymbol{X}_i$ for subjects with $R_i = 1$ as defined in Section 2.1, and $\mathcal{A}_{\mathcal{M}_n}$ be the matrix that consists of rows of $\{\boldsymbol{A}_{\mathcal{M}_n,i} : R_i = 1\}$. For any given $(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n})$, let $\mathcal{K}(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n})$ be the collection of models that could be selected under the given data values, that is,

$$\mathcal{K}(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n}) = \Big\{\mathcal{K} : \mathcal{K}^*\{\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, (\mathcal{A}_{\mathcal{M}_n}, \widetilde{\mathcal{A}}_{\mathcal{M}_n^{\mathrm{C}}})\} = \mathcal{K} \text{ for some } \widetilde{\mathcal{A}}_{\mathcal{M}_n^{\mathrm{C}}} \in \mathbb{R}^{(\sum_i R_i)\times(p_n - |\mathcal{M}_n|)}\Big\}.$$

For any $\mathcal{K} \in \Omega_n$, define

$$\overline{\mathcal{K}} = \Big\{\mathcal{M} : \mathcal{M} \in \mathcal{K}(\widetilde{\mathcal{S}}, \widetilde{\mathcal{X}}, \widetilde{\mathcal{A}}_{\mathcal{M}_n}) \text{ for some } (\widetilde{\mathcal{S}}, \widetilde{\mathcal{X}}, \widetilde{\mathcal{A}}_{\mathcal{M}_n}) \text{ such that } \mathcal{K} \in \mathcal{K}^*(\widetilde{\mathcal{S}}, \widetilde{\mathcal{X}}, \widetilde{\mathcal{A}}_{\mathcal{M}_n})\Big\}.$$

We can understand $\overline{\mathcal{K}}$ as the collection of models that are "close" to $\mathcal{K}$: there exist auxiliary variable values $\widetilde{\mathcal{A}}_{\mathcal{M}_n}$ that are compatible with the selection of $\mathcal{K}$ as well as the selection of other elements of $\overline{\mathcal{K}}$. For marginal screening, because the selection of components of $\boldsymbol{A}_{\mathcal{M}_n}$ depends only on $(\mathcal{S}, \mathcal{X}, \mathcal{A}_{\mathcal{M}_n})$ but not $\mathcal{A}_{\mathcal{M}_n^{\mathrm{C}}}$, $\overline{\mathcal{K}}$ consists of models that include variables in $\mathcal{K}$ along with a subset of variables in $\boldsymbol{A}_{\mathcal{M}_n^{\mathrm{C}}}$.

We can replace condition (C4) in Theorems 2.1 and 2.2 and Corollary 2.1 by

(C4') Under $\beta = 0$, the residual $(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$ and the covariate $\boldsymbol{X}$ are independent, and $\boldsymbol{A}_{\mathcal{M}_n}$ is independent of $(Y, \boldsymbol{X})$. Also, $\sum_{\mathcal{K}\in\Omega_n} P(\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}) \to 0$.

For marginal screening, elements of $\{\overline{\mathcal{K}} : \mathcal{K} \in \Omega_n\}$ are mutually exclusive. The second part of condition (C4') is automatically satisfied, because

$$\sum_{\mathcal{K}\in\Omega_n} P(\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}) = P\Big(\bigcup_{\mathcal{K}\in\Omega_n} \{\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}\}\Big) \leq P(\mathcal{K}^* \notin \Omega_n) \to 0.$$

2.6.2 Model Selection Events Under Marginal Screening

We discuss the model selection events and how conditions (C5) and (C6) can be established under a marginal screening procedure. Let S be an outcome of interest, $\boldsymbol{X}$ be a set of low-dimensional covariates, and $\boldsymbol{A} \equiv (A_1, \ldots, A_{p_n})^{\mathrm{T}}$ be a set of high-dimensional covariates. The observed data consist of $(S_i, \boldsymbol{X}_i, \boldsymbol{A}_i)_{i=1,\ldots,n}$; suppose that there are no missing data. For simplicity of presentation, suppose that $\boldsymbol{A}$ is mean zero and uncorrelated with $\boldsymbol{X}$; otherwise we can replace each component of $\boldsymbol{A}$ by its projection onto the orthogonal complement of the span of the observed $\boldsymbol{X}$ in the sequel. Consider a marginal screening procedure, such that the jth component of $\boldsymbol{A}$ is selected if and only if $\big|n^{-1/2}\sum_{i=1}^n (S_i - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}_i)A_{ij}\big| > \lambda_n$, where λ_n is a selection threshold, and $\widehat{\boldsymbol{\gamma}}_X$ is the least-squares estimator of $\boldsymbol{\gamma}_X$. Let $\boldsymbol{\gamma}_{0X} = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\mathrm{E}(\boldsymbol{X}S)$ and $\rho_{nj} = \mathrm{Cov}(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}, A_j)$. For $j = 1, \ldots, p_n$, we write

$$\begin{aligned}
&\frac{1}{n^{1/2}}\sum_{i=1}^n (S_i - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}_i)A_{ij} \\
&= n^{1/2}\rho_{nj} + \frac{1}{n^{1/2}}\sum_{i=1}^n \big\{(S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)A_{ij} - \rho_{nj}\big\} - \frac{1}{n^{1/2}}(\widehat{\boldsymbol{\gamma}}_X - \boldsymbol{\gamma}_{0X})^{\mathrm{T}}\sum_{i=1}^n \boldsymbol{X}_i A_{ij}.
\end{aligned}$$

Under condition (C1), we can show that for any diverging sequence k_n,

$$P\Big\{\Big|\frac{1}{n^{1/2}}(\widehat{\boldsymbol{\gamma}}_X-\boldsymbol{\gamma}_{0X})^{\mathrm{T}}\sum_{i=1}^{n}\boldsymbol{X}_iA_{ij}\Big|>k_n\Big\}\lesssim\exp\big\{-Mk_n^{\min(1,\xi)/2}\big\}$$

and

$$P\Big[\Big|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\big\{(S_i-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)A_{ij}-\rho_{nj}\big\}\Big|>k_n\Big]\lesssim\exp\big(-Mk_n^{\xi/2}\big)$$

for $j=1,\ldots,p_n$ and some large enough constant M, where $A\lesssim B$ means that $A\le CB$ for some positive constant C. If we choose $\lambda_n\gg(\log p_n)^{2/\xi}$, then

$$P\bigg[\sup_{j=1,\ldots,p_n}\Big|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big\{(S_i-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)A_{ij}-\rho_{nj}-(\widehat{\boldsymbol{\gamma}}_X-\boldsymbol{\gamma}_{0X})^{\mathrm{T}}\boldsymbol{X}_iA_{ij}\Big\}\Big|>\lambda_n\bigg]\to0.$$

If there exists a model $\mathcal{K}_n$ such that $\min_{j\in\mathcal{K}_n}|\rho_{nj}|\gg\lambda_nn^{-1/2}\gg\max_{j\notin\mathcal{K}_n}|\rho_{nj}|$, then $\mathcal{K}_n$ is selected with probability going to 1. In this case, there is no model selection variability, and conditions (C5) and (C6) are clearly satisfied with $\Omega_n=\{\mathcal{K}_n\}$.

Alternatively, suppose that $n^{1/2}\rho_{nj}=\lambda_n-c_j$ for $j=1,\ldots,q_n$ and $n^{1/2}|\rho_{nj}|\ll\lambda_n$ for $j=q_n+1,\ldots,p_n$, where $c_1,\ldots,c_{q_n}$ are uniformly bounded constants. In this case, the (marginal) signal strength and the selection threshold are of the same order, and the selection event of A_j may be nondegenerate. Let $Z_{nj}=n^{-1/2}\sum_{i=1}^{n}\big\{(S_i-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)A_{ij}-\rho_{nj}\big\}$ and $b_{nj}=-n^{-1/2}(\widehat{\boldsymbol{\gamma}}_X-\boldsymbol{\gamma}_{0X})^{\mathrm{T}}\sum_{i=1}^{n}\boldsymbol{X}_iA_{ij}$. The selection probability of $\mathcal{K}_n\subset\{1,\ldots,q_n\}$ is

$$P\Big(Z_{nj}+b_{nj}>c_j\text{ for }j\in\mathcal{K}_n,\ Z_{nj}+b_{nj}<c_j\text{ for }j\in\{1,\ldots,q_n\}\setminus\mathcal{K}_n\Big)+O(p_ne^{-M\lambda_n^{\xi/2}}).$$

Condition (C5) requires that the models selected based on the actual residuals and the estimated residuals are equal asymptotically. Let Ω_n be the collection of all subsets of $\{1,\ldots,q_n\}$, $\mathcal{K}^*=\{j:|n^{-1/2}\sum_{i=1}^{n}(S_i-\widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}_i)A_{ij}|>\lambda_n\}$, and $\mathcal{K}_0^*=\{j:|n^{-1/2}\sum_{i=1}^{n}(S_i-$

$\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)A_{ij}| > \lambda_n\}$. We have

$$
\begin{aligned}
&P(\mathcal{K}^* \neq \mathcal{K}_0^*) \\
\leq\, &P(\mathcal{K}^* \neq \mathcal{K}_0^*, \mathcal{K}^* \in \Omega_n, \mathcal{K}_0^* \in \Omega_n) + P(\mathcal{K}^* \notin \Omega_n, \mathcal{K}_0^* \notin \Omega_n) \\
\leq\, &P\Big(\cup_{j=1}^{q_n} [\{j \in \mathcal{K}^*, j \notin \mathcal{K}_0^*\} \cup \{j \in \mathcal{K}_0^*, j \notin \mathcal{K}^*\}]\Big) + O\{p_n \exp(-M\lambda_n^{\xi/2})\} \\
\leq\, &q_n \max_{j=1,\ldots,q_n} \big\{P(Z_{nj} > c_j, Z_{nj} + b_{nj} < c_j) + P(Z_{nj} < c_j, Z_{nj} + b_{nj} > c_j)\big\} + o(1). \quad (2.5)
\end{aligned}
$$

For any diverging sequence k_n and $j = 1, \ldots, q_n$, we have

$$
\begin{aligned}
&P(Z_{nj} > c_j, Z_{nj} + b_{nj} < c_j) \\
\leq\, &P(Z_{nj} > c_j, Z_{nj} - n^{-1/2}k_n < c_j) + P(n^{1/2}|b_{nj}| > k_n) \\
=\, &P(Z_{nj} < c_j + n^{-1/2}k_n) - P(Z_{nj} < c_j) + P(n^{1/2}|b_{nj}| > k_n) \\
=\, &\frac{\partial}{\partial \epsilon} P(Z_{nj} < c_j + \epsilon)\Big|_{\epsilon=\widetilde{\epsilon}} n^{-1/2}k_n + O[\exp\{-Mk_n^{\min(1,\xi)/2}\}], \quad (2.6)
\end{aligned}
$$

where $\widetilde{\epsilon}$ is some value within $(0, n^{-1/2}k_n)$. Because Z_{nj} converges to the normal distribution, the derivative term on the right-hand side above is bounded. Under the given conditions on q_n, we can find k_n such that $n^{-1/2}k_nq_n + q_n \exp\{-Mk_n^{\min(1,\xi)/2}\} = o(1)$. Under this choice of k_n and combining (2.5) and (2.6), we have $P(\mathcal{K}^* \neq \mathcal{K}_0^*) \to 0$.

To derive the second part of condition (C5), let $\Psi_n = \{\sup_{j=1,\ldots,q_n} n^{1/2}|b_{nj}| < k_n\}$ for some diverging sequence k_n. For $\mathcal{K} \in \Omega_n$, let $\mathcal{K}^- = \{1, \ldots, q_n\} \setminus \mathcal{K}$. We have

$$
\begin{aligned}
P(\mathcal{K}^* = \mathcal{K}) \leq\, &P(Z_{nj} + b_{nj} > c_j \text{ for } j \in \mathcal{K}, Z_{nj} + b_{nj} < c_j \text{ for } j \in \mathcal{K}^-, \Psi_n) \\
&+ P(\mathcal{K}^* \notin \Omega_n) + P(\Psi_n^{\mathrm{C}}) \\
=\, &P(\mathcal{K}_0^* = \mathcal{K}) + O\{p_n \exp(-M\lambda_n^{\xi/2})\} + O[q_n \exp\{-Mk_n^{\min(1,\xi)/2}\}] \\
&+ n^{-1/2}k_n \sum_{k=1}^{q_n} \frac{\partial}{\partial \epsilon_k} P(Z_{nj} > c_j + \epsilon_j \text{ for } j \in \mathcal{K}, Z_{nj} < c_j + \epsilon_j \text{ for } j \in \mathcal{K}^-)\Big|_{\epsilon=\widetilde{\epsilon}},
\end{aligned}
$$

where $\widetilde{\epsilon} = (\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_{q_n})$ such that $|\widetilde{\epsilon}_k| < n^{-1/2}k_n$. Note that each term in the summation on the right-hand side above is of the same order as $P(\mathcal{K}_0^* = \mathcal{K})$. Again, by an appropriate choice of k_n, we have $\sup_{\mathcal{K}\in\Omega_n} P(\mathcal{K}^* = \mathcal{K})/P(\mathcal{K}_0^* = \mathcal{K}) \to 1$.

Condition (C6) requires that $P(\mathcal{K}^* = \mathcal{K} \mid \boldsymbol{A})$ and $P(\mathcal{K}^* = \mathcal{K})$ are equal asymptotically over $\mathcal{K} \in \Omega_n$. Let $\sigma_j^2(A_j) = \mathrm{Var}(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X} \mid A_j)A_j^2$. Note that

$$P(\mathcal{K}^* = \mathcal{K} \mid \boldsymbol{A}) = P\Big(\widetilde{Z}_{nj} > \widetilde{c}_j \text{ for } j \in \mathcal{K}, \widetilde{Z}_{nj} < \widetilde{c}_j \text{ for } j \in \mathcal{K}^- \mid \boldsymbol{A}\Big) + o(1),$$

where $\widetilde{Z}_{nj} = \{n^{-1}\sum_i \sigma_j^2(A_{ij})\}^{-1/2}Z_{nj}$ and $\widetilde{c}_j = \{n^{-1}\sum_i \sigma_j^2(A_{ij})\}^{-1/2}c_j$. Based on a similar expansion as the above, the first term on the right-hand side above is

$$\begin{aligned}
&P\Big(\widetilde{Z}_{nj} > \widetilde{c}_{0j} \text{ for } j \in \mathcal{K}, \widetilde{Z}_{nj} < \widetilde{c}_{0j} \text{ for } j \in \mathcal{K}^- \mid \boldsymbol{A}\Big) \\
&+ n^{-1/2}k_n \sum_{k=1}^{q_n} \frac{\partial}{\partial \epsilon_k} P\Big(\widetilde{Z}_{nj} > \widetilde{c}_{0j} + \epsilon_j \text{ for } j \in \mathcal{K}, \widetilde{Z}_{nj} < \widetilde{c}_{0j} + \epsilon_j \text{ for } j \in \mathcal{K}^- \mid \boldsymbol{A}\Big)|_{\epsilon=\widetilde{\epsilon}} \\
&+ O\bigg[P\Big\{\sup_{j=1,\ldots,q_n} n^{1/2}|\widetilde{c}_j - \widetilde{c}_{0j}| > k_n\Big\}\bigg],
\end{aligned}$$

where $\widetilde{c}_{0j} = [E\{\sigma_j^2(A_j)\}]^{-1/2}c_j$. If $\sigma_j^2(A_j) \lesssim A_j^d$ for some $d \geq 2$ and $\sigma_j^2(A_j)$ is bounded away from zero, then $P\{n^{1/2}|\widetilde{c}_j - \widetilde{c}_{0j}| > k_n\} \lesssim \exp(-Mk_n^{\xi/d})$. Condition (C6) can then be established by noting that the first term above converges to a limit that does not depend on $\boldsymbol{A}$, and the second and third terms converge to 0 faster than the first term under an appropriate choice of k_n.

2.6.3 Evaluation of Power

We evaluate the power of the score test under a prespecified set of auxiliary variables. According to (2.2), the score statistic can be expanded as $n^{-1/2}\sum_{i=1}^n \boldsymbol{\Psi}(\ell_{\beta,i}, \boldsymbol{\ell}_{\alpha,i}^{\mathrm{T}}, \boldsymbol{\ell}_{\gamma,i}^{\mathrm{T}})^{\mathrm{T}} + o_p(1)$, where $\boldsymbol{\Psi} = (1, \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}, \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1})$, $\ell_{\beta,i} = \{Y_i - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\}$, $\boldsymbol{\ell}_{\alpha,i} = \{Y_i - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i$, and $\boldsymbol{\ell}_{\gamma,i} = R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\boldsymbol{W}_{\mathcal{K},i}$. Let $(\ell_\beta, \boldsymbol{\ell}_\alpha, \boldsymbol{\ell}_\gamma)$ be

$(\ell_{\beta,i}, \boldsymbol{\ell}_{\alpha,i}, \boldsymbol{\ell}_{\gamma,i})$ for a generic subject. Under a contiguous alternative of $\beta = \beta_n = n^{-1/2}b$ for some constant b, the score test statistic converges to a noncentral chi-square distribution with the noncentrality parameter

$$\mathcal{C} = \lim_{n\to\infty} \frac{[n^{1/2}\boldsymbol{\Psi}\{\mathrm{E}(\ell_\beta), \mathrm{E}(\boldsymbol{\ell}_\alpha)^{\mathrm{T}}, \mathrm{E}(\boldsymbol{\ell}_\gamma)^{\mathrm{T}}\}^{\mathrm{T}}]^2}{\boldsymbol{\Psi}\mathrm{E}\{(\ell_\beta, \boldsymbol{\ell}_\alpha^{\mathrm{T}}, \boldsymbol{\ell}_\gamma^{\mathrm{T}})^{\otimes 2}\}\boldsymbol{\Psi}^{\mathrm{T}}}, \tag{2.7}$$

where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ for any vector $\boldsymbol{a}$. We evaluate $\mathcal{C}$ under different missing mechanisms and outcome models.

First, consider the linear model $Y = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S + \epsilon$, where $\mathrm{E}(\epsilon \mid \boldsymbol{X}, S, \boldsymbol{A}) = 0$ and $\mathrm{Var}(\epsilon \mid \boldsymbol{X}, S, \boldsymbol{A}) = \sigma^2$. Consider an extreme-tail sampling scheme, such that $R = I(Y \in \Omega)$ with $\Omega = (-\infty, C_2) \cup (C_1, \infty)$ for some constants $C_2 < C_1$. Under the contiguous alternative,

$$n^{1/2}\mathrm{E}(\ell_\beta) = n^{1/2}\mathrm{E}\{\epsilon R(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\} + b\mathrm{E}[S\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\}].$$

The first expectation on the right-hand side above is

$$\begin{aligned}
&\mathrm{E}\{\epsilon I(Y \in \Omega)(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\} \\
&= \mathrm{E}\big[(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\mathrm{E}\{\epsilon I(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X} + \beta_n S + \epsilon \in \Omega) \mid \boldsymbol{W}_{\mathcal{K}}, S\}\big] \\
&= \beta_n \mathrm{E}\Big[(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2 \frac{\partial}{\partial t}\mathrm{E}\{\epsilon I(t + \boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X} + \epsilon \in \Omega) \mid \boldsymbol{W}_{\mathcal{K}}\}\big|_{t=0}\Big] + o(n^{-1/2}) \\
&= \beta_n \mathrm{E}\{(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2 h_1(\boldsymbol{X})\} + o(n^{-1/2}),
\end{aligned}$$

where $h_1(\boldsymbol{X}) = (C_1 - \boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})f_\epsilon(C_1 - \boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}) - (C_2 - \boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})f_\epsilon(C_2 - \boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})$, and f_ϵ is the density of ϵ. Thus,

$$n^{1/2}\mathrm{E}(\ell_\beta) \to b\mathrm{E}\{(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2 h_1(\boldsymbol{X})\} + b\mathrm{E}[S\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\}].$$

Similarly,

$$n^{1/2}\mathrm{E}(\boldsymbol{\ell}_\gamma) \to b\mathrm{E}\Big\{h_2(\boldsymbol{X})(S-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2\boldsymbol{W}_{\mathcal{K}}\Big\},$$

where $h_2(\boldsymbol{X}) = f_\epsilon(C_1-\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}) - f_\epsilon(C_2-\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})$. Simple algebraic manipulations yield $n^{1/2}\mathrm{E}(\boldsymbol{\ell}_\alpha) = b\mathrm{E}(S\boldsymbol{X})$. With $h_3(\boldsymbol{X}) = \mathrm{E}\{\epsilon I(\epsilon \le C_2-\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X} \text{ or } \epsilon \ge C_1-\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\}$ and $h_4(\boldsymbol{X}) = P(\epsilon \le C_2-\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}) + P(\epsilon \ge C_1-\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})$, $\boldsymbol{I}_{\beta\gamma} = -\mathrm{E}\{h_3(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\}$ and $\boldsymbol{I}_{\gamma\gamma} = \mathrm{E}\{h_4(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}}\}$. Combining the above results, we can derive that (the limit of) the numerator of $\mathcal{C}$ is

$$\begin{aligned}
&b^2\Big(\mathrm{E}\big\{(S-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2h_1(\boldsymbol{X})\big\} + \mathrm{E}\big[(S-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})\{\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}} + h_4(\boldsymbol{X})(S-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\}\big] \\
&- \mathrm{E}\big\{h_3(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\big\}^{\mathrm{T}}\mathrm{E}\big\{h_4(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}}\big\}^{-1}\mathrm{E}\big\{h_2(\boldsymbol{X})(S-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2\boldsymbol{W}_{\mathcal{K}}\big\}\Big)^2,
\end{aligned}$$

where $\boldsymbol{\gamma}_{0X} = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\mathrm{E}(\boldsymbol{X}S)$. Note that under the contiguous alternative, the limit of the denominator of $\mathcal{C}$ is equal to that under the null. The denominator is

$$\begin{aligned}
&\mathrm{Var}(\ell_\beta + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{\ell}_\alpha + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{\ell}_\gamma) \\
&= \mathrm{E}\Big[\epsilon\mathcal{P}_X^{\perp}\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\} - \mathrm{E}\big\{h_3(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\big\}^{\mathrm{T}}\mathrm{E}\big\{h_4(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}}\big\}^{-1}R(S-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\boldsymbol{W}_{\mathcal{K}}\Big]^2,
\end{aligned}$$

where $\mathcal{P}_X^{\perp}$ denotes the projection onto the orthogonal space of $\boldsymbol{X}$, i.e., $\mathcal{P}_X^{\perp}(T) = T - \boldsymbol{X}^{\mathrm{T}}\mathrm{E}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\mathrm{E}(\boldsymbol{X}T)$ for any random variable T.

When S is missing completely at random, $\boldsymbol{I}_{\beta\gamma} = \boldsymbol{0}$. In this case, the numerator of $\mathcal{C}$ is

$$\begin{aligned}
&b^2\Big(\mathrm{E}\big[S\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\}\big] - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\mathrm{E}\big[\boldsymbol{X}\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\}\big]\Big)^2 \\
&= b^2\Big[p_R\mathrm{E}\big\{\mathcal{P}_X^{\perp}(S)\big\}^2 + (1-p_R)\mathrm{E}\big\{\mathcal{P}_X^{\perp}(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\big\}^2\Big]^2,
\end{aligned}$$

where $p_R = P(R = 1)$. The denominator of $\mathcal{C}$ is

$$\mathrm{Var}(\ell_\beta + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{\ell}_\alpha) = \sigma^2\mathrm{E}\big[\mathcal{P}_X^\perp\big\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}\big]^2,$$

We have

$$\mathcal{C} = \frac{b^2\big[p_R\mathrm{E}\big\{\mathcal{P}_X^\perp(S)\big\}^2 + (1-p_R)\mathrm{E}\big\{\mathcal{P}_X^\perp(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\big\}^2\big]^2}{\sigma^2\mathrm{E}\big[\mathcal{P}_X^\perp\big\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}\big]^2}.$$

Following the arguments in Section S.4 of Wong et al. (2019b), we can show that a test with a larger set of auxiliary variables has a larger noncentrality parameter and thus is more powerful.

Next, consider the logistic regression model $\mathrm{logit}\{P(Y = 1 \mid \boldsymbol{X}, S)\} = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S$. In this case, $\mu(x) = e^x/(1+e^x)$ and $\mu'(x) = e^x/(1+e^x)^2$. For a case-control study, we can set $R = 1 - (1-Y)\omega$, where ω is a Bernoulli variable that is independent of the data with $P(\omega = 1) = p_\omega$. In this case, the numerator of $\mathcal{C}$ is

$$\begin{aligned}
&b^2\Big\{\mathrm{E}\big\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})S^2\big\} - p_\omega\mathrm{E}\big\{\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2\big\} \\
&- \mathrm{E}\big(\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}[p_R(\boldsymbol{X})S + \{1-p_R(\boldsymbol{X})\}\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}]\big)^{\mathrm{T}}\mathrm{E}\big\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\big\}^{-1}\mathrm{E}\big\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})S\boldsymbol{X}\big\} \\
&- p_\omega^2\mathrm{E}\big\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\big\}^{\mathrm{T}}\mathrm{E}\big\{p_R(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}}\big\}^{-1}\mathrm{E}\big\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})^2\boldsymbol{W}_{\mathcal{K}}\big\}\Big\}^2,
\end{aligned}$$

where $p_R(\boldsymbol{X}) = P(R = 1 \mid \boldsymbol{X}) = 1 - p_\omega/(1 + e^{\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}})$. The denominator of $\mathcal{C}$ is

$$\begin{aligned}
&\mathrm{Var}(\ell_\beta + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{\ell}_\alpha + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{\ell}_\gamma) \\
=&\,\mathrm{E}\Big[\{Y - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\}\widetilde{\mathcal{P}}_X^\perp\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\} \\
&- p_\omega\mathrm{E}\big\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\big\}^{\mathrm{T}}\mathrm{E}\{p_R(\boldsymbol{X})\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}}\}^{-1}R(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\boldsymbol{W}_{\mathcal{K}}\Big]^2,
\end{aligned}$$

where $\widetilde{\mathcal{P}}_X^\perp(T) = T - \boldsymbol{X}^{\mathrm{T}}\mathrm{E}\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\}^{-1}\mathrm{E}\{\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\boldsymbol{X}T\}$ for any random variable

T.

When S is missing completely at random, $\boldsymbol{I}_{\beta\gamma} = \mathbf{0}$. The denominator of $\mathcal{C}$ simplifies to

$$\mathrm{Var}(\ell_\beta + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1} \boldsymbol{\ell}_\alpha) = \mathrm{E}\big[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \widetilde{\mathcal{P}}_X^{\perp}\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}}\}\big]^2.$$

The numerator of $\mathcal{C}$ is

$$b^2\Big(p_R \mathrm{E}\big[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X})\{\widetilde{\mathcal{P}}_X^{\perp}(S)\}^2\big] + (1-p_R)\mathrm{E}\big[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X})\{\widetilde{\mathcal{P}}_X^{\perp}(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}})\}^2\big]\Big)^2.$$

Therefore, the noncentrality parameter under MCAR is

$$\mathcal{C} = \frac{b^2\big(p_R \mathrm{E}\big[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X})\{\widetilde{\mathcal{P}}_X^{\perp}(S)\}^2\big] + (1-p_R)\mathrm{E}\big[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X})\{\widetilde{\mathcal{P}}_X^{\perp}(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}})\}^2\big]\big)^2}{\mathrm{E}\big[\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \widetilde{\mathcal{P}}_X^{\perp}\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}}\}\big]^2}.$$

As in the linear model, the power increases as the set of auxiliary variables expands.

Based on the limiting distribution of the score test statistic established above, we can evaluate the power of the test under different sets of auxiliary variables. We consider Setting 2 in the simulation studies in Section 2.3. We plot the asymptotic power under the linear and logistic regression models, with auxiliary variables $A_1, \ldots, A_q$ for $q = 1, \ldots, 200$ in Figure 2.6. As expected, under MCAR, the power increases as more auxiliary variables are included into the model. The same pattern holds for the two MAR mechanisms considered.

2.6.4 Additional Theoretical Results

Before proving the theorems, we present the following lemmas. All lemmas are stated under the null hypothesis H_0.

Lemma 2.1. *Under conditions (C1)–(C3), the inequalities*

$$\sup_{\mathcal{K}\in\Omega_n}\left\|\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}-\boldsymbol{\gamma}_{0\mathcal{K}}\right\|>C_1\left\{\Big(\frac{t+\log r_n+q_n}{n}\Big)^{1/2}+\frac{q_n(\log n)^{2/\xi}(t+\log r_n+q_n)^{2/\xi}}{n}\right\},$$

$$\sup_{\mathcal{K}\in\Omega_n}\left\|\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}\widehat{\boldsymbol{I}}_{\beta\gamma}-\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{I}_{\beta\gamma}\right\|>C_1\left\{\Big(\frac{t+\log r_n+q_n}{n}\Big)^{1/2}+\frac{q_n(\log n)^{2/\xi}(t+\log r_n+q_n)^{2/\xi}}{n}\right\},\ \textit{and}$$

$$\sup_{\mathcal{K}\in\Omega_n}\left\|\widehat{\boldsymbol{I}}_{\beta\alpha}-\boldsymbol{I}_{\beta\alpha}\right\|>C_1\left\{\Big(\frac{t+\log r_n}{n}\Big)^{1/2}+\frac{q_n^{1/2}(\log n)^{1/\xi}(t+\log r_n)^{1/\min(1,\xi)}}{n}\right\}$$

hold with probability at most C_2e^{-t} for large enough n and t, where C_1 and C_2 are positive constants.

Let

$$\begin{aligned}
\widehat{\sigma}_1^2(\mathcal{K})=&\frac{1}{n}\sum_{i=1}^n\mathrm{Var}(\epsilon\mid R_i,\boldsymbol{X}_i)\big\{R_iS_i+(1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}+\boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\}^2\\
\widehat{\sigma}_2^2(\mathcal{K})=&\frac{1}{n}\sum_{i=1}^n\mathrm{Var}\Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}}+\boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\big\{\mathrm{E}(\epsilon\mid R_i,\boldsymbol{X})\boldsymbol{X}-\mathrm{E}(\epsilon\boldsymbol{X}\mid R_i)\big\}\\
&+\big\{\mathrm{E}(\epsilon\mid R_i,\boldsymbol{X})-\mathrm{E}(\epsilon\mid R_i)\big\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\\
&+\big\{\mathrm{E}(\epsilon\mid R_i,\boldsymbol{X})+\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}\big\}R_i(S_i-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\mid R_i,\boldsymbol{A}_i,S_i-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i\Big]\\
\widehat{\sigma}_3^2(\mathcal{K})=&\frac{1}{n}\sum_{i=1}^n\mathrm{Var}\Big\{(\boldsymbol{\gamma}_{0X}^{\mathrm{T}}+\boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\mathrm{E}(\epsilon\boldsymbol{X}\mid R)+\mathrm{E}(\epsilon\mid R)\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\mid\boldsymbol{A}_i\Big\}.
\end{aligned}$$

Lemma 2.2. *Under conditions (C1)–(C4), for large enough n and t,*

$$P\Big[\sup_{\mathcal{K}\in\Omega_n}\sum_{k=1}^3\big|\widehat{\sigma}_k^2(\mathcal{K})-\sigma_k^2(\mathcal{K})\big|>C_1\left\{\Big(\frac{t+\log r_n}{n}\Big)^{1/2}+\frac{q_n(\log n)^{2/\xi}(t+\log r_n)^{2/\xi}}{n}\right\}\Big]\le C_2e^{-t},$$

where C_1 and C_2 are positive constants.

Lemma 2.3. *Under conditions (C1)–(C4),*

$$\mathrm{E}\left\{\sup_{\mathcal{K}\in\Omega_n}\left|\frac{1}{n^{1/2}}\sum_{i=1}^n(\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}-\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1})\boldsymbol{W}_{\mathcal{K},i}R_i(S_i-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\right|\right\}=o(1).$$

Lemma 2.4. *Assume that conditions (C1)–(C3) hold. For U_{ki} and $\widetilde{U}_{ki}$ ($k = 1, 2, 3$; $i = 1, \ldots, n$) defined in the proof of Theorem 2.1,*

$$P\left[\sum_{k=1}^{3} \sup_{\mathcal{K}\in\Omega_n} \frac{1}{n^{3/2}} \sum_{i=1}^{n} \left(|U_{ki}|^3 + |\widetilde{U}_{ki}|^3\right) > C_1\left\{\frac{(t+\log r_n)^{1/2}}{n} + \frac{q_n^{3/2}(\log n)^{6/\xi}(t+\log r_n)^{6/\xi}}{n^{3/2}}\right\}\right]$$

is smaller than $C_2 e^{-t}$ for large enough n and t, where C_1 and C_2 are positive constants.

The proofs of Lemmas 2.1–2.4 and Theorem 2.2 involve Theorem A.1 of Kuchibhotla et al. (2018), which is restated here for the convenience of the reader. We have relabelled some quantities to be consistent with the notations in this chapter.

Lemma 2.5. *Suppose $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_n$ are mean zero independent random vectors in $\mathbb{R}^k$ such that for some $a > 0$ and $K_{n,k} > 0$, $\max_{1\le i\le n} \max_{1\le j\le k} \|Q_{ij}\|_{\psi_a} \le K_{n,k}$, where Q_{ij} is the jth component of $\boldsymbol{Q}_i$. Define*

$$\Gamma_{n,k} \equiv \max_{1\le j\le k} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(Q_{ij}^2).$$

Then for any $t \ge 0$, with probability at least $1 - 3e^{-t}$,

$$\max_{1\le j\le k}\left|\frac{1}{n}\sum_{i=1}^{n} Q_{ij}\right| \le 7\Big[\frac{\Gamma_{n,k}\{t+\log(2k)\}}{n}\Big]^{1/2} + \frac{C_a K_{n,k}\{\log(2n)\}^{1/a}\{t+\log(2k)\}^{1/T_1(a)}}{n},$$

where $T_1(a) = \min\{a, 1\}$ and C_a is a constant depending only on a.

Proof of Lemma 2.1. Let $\boldsymbol{W} = (\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{A}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{W}_i = (\boldsymbol{X}_i^{\mathrm{T}}, \boldsymbol{A}_i^{\mathrm{T}})^{\mathrm{T}}$ for $i = 1, \ldots, n$. Let $\widehat{\boldsymbol{I}}_{C,\gamma\gamma} = n^{-1}\sum_{i=1}^{n} R_i \boldsymbol{W}_i \boldsymbol{W}_i^{\mathrm{T}}$, $\widehat{\boldsymbol{U}}_{C,\gamma} = n^{-1}\sum_{i=1}^{n} R_i S_i \boldsymbol{W}_i$, $\boldsymbol{I}_{C,\gamma\gamma} = \mathrm{E}(R\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}})$, and $\boldsymbol{U}_{C,\gamma} = \mathrm{E}(RS\boldsymbol{W})$. Let p be the dimension of $\boldsymbol{W}$ and $\Theta_n = \{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}\| \le 1, \boldsymbol{b}_{\mathcal{K}^{\mathrm{C}}} = \boldsymbol{0}$ for some $\mathcal{K} \in \Omega_n\}$, where $\boldsymbol{b}_{\mathcal{K}^{\mathrm{C}}}$ denotes the subvector of $\boldsymbol{b}$ that corresponds to the com-

ponents of $\boldsymbol{A}$ not in $\mathcal{K}$. Note that

$$\begin{aligned}\sup_{\mathcal{K}\in\Omega_n}\left\|\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}-\boldsymbol{\gamma}_{0\mathcal{K}}\right\| &= \sup_{\boldsymbol{b}\in\Theta_n}\left|\boldsymbol{b}^{\mathrm{T}}(\widehat{\boldsymbol{I}}_{C,\gamma\gamma}^{-1}\widehat{\boldsymbol{U}}_{C,\gamma}-\boldsymbol{I}_{C,\gamma\gamma}^{-1}\boldsymbol{U}_{C,\gamma})\right| \\ &\leq \frac{\sup_{\boldsymbol{b}\in\Theta_n}\left|\boldsymbol{b}^{\mathrm{T}}(\widehat{\boldsymbol{U}}_{C,\gamma}-\boldsymbol{U}_{C,\gamma})\right|+\sup_{\boldsymbol{b}\in\Theta_n}\left|\boldsymbol{b}^{\mathrm{T}}(\widehat{\boldsymbol{I}}_{C,\gamma\gamma}-\boldsymbol{I}_{C,\gamma\gamma})\boldsymbol{b}\right|\|\boldsymbol{\gamma}_{0\mathcal{K}}\|}{\inf_{\mathcal{K}\in\Omega_n}\lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})\}-\sup_{\boldsymbol{b}\in\Theta_n}\left|\boldsymbol{b}^{\mathrm{T}}(\widehat{\boldsymbol{I}}_{C,\gamma\gamma}-\boldsymbol{I}_{C,\gamma\gamma})\boldsymbol{b}\right|}\end{aligned}$$

whenever $\inf_{\mathcal{K}\in\Omega_n}\lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})\}>\sup_{\boldsymbol{b}\in\Theta_n}\left|\boldsymbol{b}^{\mathrm{T}}(\widehat{\boldsymbol{I}}_{C,\gamma\gamma}-\boldsymbol{I}_{C,\gamma\gamma})\boldsymbol{b}\right|$; this inequality follows from the proof of Theorem 3.1 in Kuchibhotla et al. (2018). Let $\mathcal{N}(\epsilon,\Theta_n)$ be an ϵ-net of Θ_n and $\mathcal{N}_s(\epsilon)$ be an ϵ-net of $\{\boldsymbol{b}\in\mathbb{R}^s:\|\boldsymbol{b}\|\leq 1\}$ for $s\geq 1$. We have $|\mathcal{N}(\epsilon,\Theta_n)|\leq r_n|\mathcal{N}_{q_n}(\epsilon)|$, which is in turn smaller than $r_n(1+\epsilon^{-1})^{q_n}$ by Lemma 4.1 of Pollard (1990). From expression (21) of Kuchibhotla et al. (2018), we have

$$\sup_{\boldsymbol{b}\in\Theta_n}\left|\boldsymbol{b}^{\mathrm{T}}(\widehat{\boldsymbol{U}}_{C,\gamma}-\boldsymbol{U}_{C,\gamma})\right|\leq 2\sup_{\boldsymbol{b}\in\mathcal{N}(1/2,\Theta_n)}\left|\boldsymbol{b}^{\mathrm{T}}\widehat{\boldsymbol{U}}_{C,\gamma}-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{U}_{C,\gamma}\right|.$$

We place a probability bound on the right-hand side above using Lemma 2.5, with $Q_{ij}=R_iS_i\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{W}_i$, $\boldsymbol{b}_j$ being the jth element of $\mathcal{N}(1/2,\Theta_n)$, $k=|\mathcal{N}(1/2,\Theta_n)|$, and $a=\xi/2$. Note that

$$\|R_iS_i\boldsymbol{b}^{\mathrm{T}}\boldsymbol{W}_i\|_{\psi_{\xi/2}}\lesssim\|S_i\|_{\psi_\xi}\|\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{W}_i\|_{\psi_\xi}\lesssim q_n^{1/2},$$

where the first inequality follows from (3.5) of Kuchibhotla and Chakrabortty (2018). Therefore, with probability at most $3e^{-t}$,

$$\begin{aligned}&\sup_{\boldsymbol{b}\in\mathcal{N}(1/2,\Theta_n)}\left|\boldsymbol{b}^{\mathrm{T}}\widehat{\boldsymbol{U}}_{C,\gamma}-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{U}_{C,\gamma}\right| \\ &>M_1\Big[\Big\{\frac{t+\log(2r_n)+q_n}{n}\Big\}^{1/2}+\frac{q_n^{1/2}\{\log(2n)\}^{2/\xi}\{t+\log(2r_n)+q_n\}^{2/\xi}}{n}\Big]\end{aligned}$$

for any $t > 0$ and some positive constant M_1. Likewise,

$$\sup_{\boldsymbol{b}\in\Theta_n} \left|\boldsymbol{b}^{\mathrm{T}}(\widehat{\boldsymbol{I}}_{C,\gamma\gamma} - \boldsymbol{I}_{C,\gamma\gamma})\boldsymbol{b}\right| \leq 2 \sup_{\boldsymbol{b}\in\mathcal{N}(1/4,\Theta_n)} \left|\boldsymbol{b}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{C,\gamma\gamma}\boldsymbol{b} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{I}_{C,\gamma\gamma}\boldsymbol{b}\right|,$$

and with probability at most $3e^{-t}$,

$$\begin{aligned}
&\sup_{\boldsymbol{b}\in\mathcal{N}(1/4,\Theta_n)} \left|\boldsymbol{b}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{C,\gamma\gamma}\boldsymbol{b} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{I}_{C,\gamma\gamma}\boldsymbol{b}\right| \\
&> M_2\Big[\Big\{\frac{t + \log(2r_n) + q_n}{n}\Big\}^{1/2} + \frac{q_n\{\log(2n)\}^{2/\xi}\{t + \log(2r_n) + q_n\}^{2/\xi}}{n}\Big]
\end{aligned}$$

for any $t > 0$ and some positive constant M_2. By condition (C2), $\inf_{\mathcal{K}\in\Omega_n} \lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})\}$ is bounded away from 0. Also,

$$\mathrm{Var}(RS) \geq \mathrm{Var}(R\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}) \geq \lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})\}\|\boldsymbol{\gamma}_{0\mathcal{K}}\|^2 - \{\mathrm{E}(R\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})\}^2$$

uniformly over $\mathcal{K}$. Because $|\mathrm{E}(R\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}})|$ and $\mathrm{Var}(RS)$ are bounded above and the eigenvalues of $\mathrm{E}(R\boldsymbol{W}_{\mathcal{K}}\boldsymbol{W}_{\mathcal{K}}^{\mathrm{T}})$ are bounded below over $\mathcal{K} \in \Omega_n$, $\sup_{\mathcal{K}\in\Omega_n} \|\boldsymbol{\gamma}_{0\mathcal{K}}\|^2$ is bounded. We conclude that for any $t > 0$, the probability of the event

$$\sup_{\mathcal{K}\in\Omega_n} \left\|\widehat{\boldsymbol{\gamma}}_{\mathcal{K}} - \boldsymbol{\gamma}_{0\mathcal{K}}\right\| > M_3\Big[\Big\{\frac{t + \log(2r_n) + q_n}{n}\Big\}^{1/2} + \frac{q_n\{\log(2n)\}^{2/\xi}\{t + \log(2r_n) + q_n\}^{2/\xi}}{n}\Big]$$

is bounded by M_4e^{-t} for some positive constants M_3 and M_4, and the first result follows. The second result can be proved analogously.

For the third result, let $\widehat{\boldsymbol{I}}_{2,\beta\alpha} = -n^{-1}\sum_{i=1}^n \mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)(1 - R_i)\boldsymbol{W}_i\boldsymbol{X}_i^{\mathrm{T}}$ and $\boldsymbol{I}_{2,\beta\alpha}$ be its expected value. Let $\Xi_n = \{\boldsymbol{\gamma} \in \mathbb{R}^p : \boldsymbol{\gamma}_{\mathcal{K}} = \boldsymbol{\gamma}_{0\mathcal{K}}, \boldsymbol{\gamma}_{\mathcal{K}^{\mathrm{C}}} = \boldsymbol{0} \text{ for some } \mathcal{K} \in \Omega_n\}$. We have

$$\sup_{\mathcal{K}\in\Omega_n} \left\|\widehat{\boldsymbol{I}}_{\beta\alpha} - \boldsymbol{I}_{\beta\alpha}\right\|$$

$$\leq \Big\| \frac{1}{n} \sum_{i=1}^{n} \mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}_i) \boldsymbol{X}_i R_i S_i - \mathrm{E}\big\{ \mu'(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \boldsymbol{X} RS \big\} \Big\| + \sup_{\boldsymbol{\gamma} \in \Xi_n} \big\| \boldsymbol{\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{2,\beta\alpha} - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{I}_{2,\beta\alpha} \big\|.$$

Clearly, the first term on the right-hand side above is of order $n^{-1/2}$. By Lemma 2.5,

$$\begin{aligned} &\sup_{\boldsymbol{\gamma} \in \Xi_n} \big\| \boldsymbol{\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{2,\beta\alpha} - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{I}_{2,\beta\alpha} \big\| \\ &> M_5 \bigg[\Big\{ \frac{t + \log(2r_n)}{n} \Big\}^{1/2} + \frac{q_n^{1/2} \{\log(2n)\}^{1/\xi} \{t + \log(2r_n)\}^{1/\min(1,\xi)}}{n} \bigg] \end{aligned}$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_5. The desired result follows. □

Proof of Lemma 2.2. Let $\widehat{\sigma}_{1i}^2(\mathcal{K})$ be the ith term in the summation in $\widehat{\sigma}_1^2(\mathcal{K})$ $(i = 1, \ldots, n)$. Because

$$\begin{aligned} \sigma_1^2(\mathcal{K}) &= \mathrm{E}\Big(\mathrm{Var}\Big[\epsilon \big\{ RS + (1 - R) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1} \boldsymbol{X} \big\} \mid R, S, \boldsymbol{X}, \boldsymbol{A} \Big] \Big) \\ &= \mathrm{E}\Big[\mathrm{Var}\big(\epsilon \mid R, \boldsymbol{X} \big) \big\{ RS + (1 - R) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1} \boldsymbol{X} \big\}^2 \Big], \end{aligned}$$

the expectation of $\widehat{\sigma}_{1i}^2(\mathcal{K})$ is equal to $\sigma_1^2(\mathcal{K})$ for $i = 1, \ldots, n$. Under condition (C1), $\|\widehat{\sigma}_{1i}^2(\mathcal{K})\|_{\psi_{\xi/2}} \lesssim q_n$ uniformly over $\mathcal{K} \in \Omega_n$. By Lemma 2.5,

$$\begin{aligned} &\sup_{\mathcal{K} \in \Omega_n} \Big| \frac{1}{n} \sum_{i=1}^{n} \widehat{\sigma}_{1i}^2(\mathcal{K}) - \sigma_1^2(\mathcal{K}) \Big| \\ &> M_1 \bigg[\Big\{ \frac{t + \log(2r_n)}{n} \Big\}^{1/2} + \frac{q_n \{\log(2n)\}^{2/\xi} \{t + \log(2r_n)\}^{2/\xi}}{n} \bigg] \end{aligned}$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_1. Therefore, the desired probability bound holds for $\sup_{\mathcal{K} \in \Omega_n} |\widehat{\sigma}_1^2(\mathcal{K}) - \sigma_1^2(\mathcal{K})|$.

To show the result for $\widehat{\sigma}_2^2(\mathcal{K})$, we note that the term in the conditional variance in the definition of $\widehat{\sigma}_2^2(\mathcal{K})$ has conditional expectation zero given $(R, \boldsymbol{A}, S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}} \boldsymbol{X})$. By

condition (C4), $\boldsymbol{X}$ is independent of $(S-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$, so that $\mathrm{E}\{\mathrm{E}(\epsilon \mid R,\boldsymbol{X})\boldsymbol{X}-\mathrm{E}(\epsilon\boldsymbol{X} \mid R) \mid R,\boldsymbol{A},S-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}\}=0$ and $\mathrm{E}\{\mathrm{E}(\epsilon \mid R,\boldsymbol{X})-\mathrm{E}(\epsilon \mid R) \mid R,\boldsymbol{A},S-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}\}=0$. Also,

$$\mathrm{E}\big\{\mathrm{E}(\epsilon \mid R=1,\boldsymbol{X})+\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K}} \mid R=1,\boldsymbol{A},S-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}\big\}=0. \tag{2.8}$$

To see this, let $\widetilde{\boldsymbol{W}}_{\mathcal{K}}=\boldsymbol{W}_{\mathcal{K}}-(\boldsymbol{0}^{\mathrm{T}},\mathrm{E}(\boldsymbol{A}_{\mathcal{K}}^{\mathrm{T}}))^{\mathrm{T}}$, and note that

$$\begin{aligned}\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{w} &= \mathrm{E}\{\epsilon(1-R)\widetilde{\boldsymbol{W}}_{\mathcal{K}}\}^{\mathrm{T}}\mathrm{E}(R\widetilde{\boldsymbol{W}}_{\mathcal{K}}\widetilde{\boldsymbol{W}}_{\mathcal{K}}^{\mathrm{T}})^{-1}\{\boldsymbol{w}-(\boldsymbol{0}^{\mathrm{T}},\mathrm{E}(\boldsymbol{A}_{\mathcal{K}}^{\mathrm{T}}))^{\mathrm{T}}\}\\ &\equiv \widetilde{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widetilde{\boldsymbol{I}}_{\gamma\gamma}^{-1}\{\boldsymbol{w}-(\boldsymbol{0}^{\mathrm{T}},\mathrm{E}(\boldsymbol{A}_{\mathcal{K}}^{\mathrm{T}}))^{\mathrm{T}}\}\end{aligned}$$

for any vector $\boldsymbol{w}$ of appropriate dimension. With $\boldsymbol{X}=(1,\widetilde{\boldsymbol{X}}^{\mathrm{T}})^{\mathrm{T}}$, we have $\widetilde{\boldsymbol{I}}_{\beta\gamma}=-P(R=1)(\mathrm{E}_1(\epsilon),\mathrm{E}_1(\epsilon\widetilde{\boldsymbol{X}}^{\mathrm{T}}),\boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}$, where E_1 denotes expectation given $R=1$, and

$$\widetilde{\boldsymbol{I}}_{\gamma\gamma}=P(R=1)\begin{pmatrix}1 & \mathrm{E}_1(\widetilde{\boldsymbol{X}}^{\mathrm{T}}) & \boldsymbol{0}\\ \mathrm{E}_1(\widetilde{\boldsymbol{X}}) & \mathrm{E}_1(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \mathrm{E}_1[\{\boldsymbol{A}_{\mathcal{K}}-\mathrm{E}(\boldsymbol{A}_{\mathcal{K}})\}\{\boldsymbol{A}_{\mathcal{K}}-\mathrm{E}(\boldsymbol{A}_{\mathcal{K}})\}^{\mathrm{T}}]\end{pmatrix}.$$

Therefore,

$$\begin{aligned}&\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\mathrm{E}(\boldsymbol{W}_{\mathcal{K}} \mid R=1,\boldsymbol{A},S-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})\\ &=-\begin{pmatrix}\mathrm{E}_1(\epsilon) & \mathrm{E}_1(\epsilon\widetilde{\boldsymbol{X}}^{\mathrm{T}})\end{pmatrix}\begin{pmatrix}1 & \mathrm{E}_1(\widetilde{\boldsymbol{X}}^{\mathrm{T}})\\ \mathrm{E}_1(\widetilde{\boldsymbol{X}}) & \mathrm{E}_1(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}})\end{pmatrix}^{-1}\begin{pmatrix}1\\ \mathrm{E}_1(\widetilde{\boldsymbol{X}})\end{pmatrix}=-\mathrm{E}_1(\epsilon),\end{aligned}$$

and that (2.8) holds. We have

$$\mathrm{E}\Big[\big\{\mathrm{E}(\epsilon \mid R,\boldsymbol{X})+\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K}}\big\}R(S-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}) \mid R,\boldsymbol{A},S-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}\Big]=0,$$

so that $\mathrm{E}\{\widehat{\sigma}_{2i}^2(\mathcal{K})\}=\sigma_2^2(\mathcal{K})$, where $\widehat{\sigma}_{2i}^2(\mathcal{K})$ is the ith term in the summation in $\widehat{\sigma}_2^2(\mathcal{K})$ ($i=$

$1, \ldots, n$). The probability bound for $\sup_{\mathcal{K}\in\Omega_n} |\widehat{\sigma}_2^2(\mathcal{K}) - \sigma_2^2(\mathcal{K})|$ can be established using a similar argument as the above. Likewise, we can establish the bound for $\sup_{\mathcal{K}\in\Omega_n} |\widehat{\sigma}_3^2(\mathcal{K}) - \sigma_3^2(\mathcal{K})|$ using a similar argument. □

Proof of Lemma 2.3. Let $\Psi_n = \{(\boldsymbol{b}_1, \boldsymbol{b}_2) \in \mathbb{R}^p \times \mathbb{R}^p : \|\boldsymbol{b}_1\| \le 1, (\boldsymbol{b}_1)_{\mathcal{K}^{\mathrm{C}}} = \boldsymbol{0}, \boldsymbol{b}_2 = \boldsymbol{\gamma}_{0\mathcal{K}} \text{ for some } \mathcal{K} \in \Omega_n\}$. We have

$$\begin{aligned}
&\sup_{\mathcal{K}\in\Omega_n} \left| \frac{1}{n^{1/2}} \sum_{i=1}^{n} (\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} - \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1}) \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \right| \\
&\le n^{1/2} \sup_{\mathcal{K}\in\Omega_n} \left\| \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} - \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \right\| \sup_{(\boldsymbol{b}_1,\boldsymbol{b}_2)\in\Psi_n} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{b}_1^{\mathrm{T}} \boldsymbol{W}_i R_i (S_i - \boldsymbol{b}_2^{\mathrm{T}} \boldsymbol{W}_i) \right| \\
&\le 2n^{1/2} \sup_{\mathcal{K}\in\Omega_n} \left\| \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} - \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \right\| \sup_{(\boldsymbol{b}_1,\boldsymbol{b}_2)\in\mathcal{N}(1/2,\Psi_n)} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{b}_1^{\mathrm{T}} \boldsymbol{W}_i R_i (S_i - \boldsymbol{b}_2^{\mathrm{T}} \boldsymbol{W}_i) \right|, \qquad (2.9)
\end{aligned}$$

where $\mathcal{N}(1/2, \Psi_n)$ is a $(1/2)$-net of Ψ_n under the distance $d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a}_1 - \boldsymbol{b}_1\| + I(\boldsymbol{a}_2 \neq \boldsymbol{b}_2)$ for $\boldsymbol{a} \equiv (\boldsymbol{a}_1, \boldsymbol{a}_2)$ and $\boldsymbol{b} \equiv (\boldsymbol{b}_1, \boldsymbol{b}_2)$ in Ψ_n; by the arguments in the proof of Lemma 2.1, $|\mathcal{N}(\epsilon, \Psi_n)| \le r_n(1 + \epsilon^{-1})^{q_n}$. For any $(\boldsymbol{b}_1, \boldsymbol{b}_2) \in \Psi_n$, we have $\mathrm{E}\{\boldsymbol{b}_1^{\mathrm{T}} \boldsymbol{W}_i R_i (S_i - \boldsymbol{b}_2^{\mathrm{T}} \boldsymbol{W}_i)\} = 0$ for $i = 1, \ldots, n$. By Lemma 2.5,

$$\begin{aligned}
&\sup_{(\boldsymbol{b}_1,\boldsymbol{b}_2)\in\mathcal{N}(1/2,\Psi_n)} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{b}_1^{\mathrm{T}} \boldsymbol{W}_i R_i (S_i - \boldsymbol{b}_2^{\mathrm{T}} \boldsymbol{W}_i) \right| \\
&> M_1 \Big[\Big\{ \frac{t + \log(2r_n) + q_n}{n} \Big\}^{1/2} + \frac{q_n \{\log(2n)\}^{2/\xi} \{t + \log(2r_n) + q_n\}^{2/\xi}}{n} \Big]
\end{aligned}$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_1. Therefore, by Lemma 2.1, the right-hand side of (2.9) is bounded above by

$$M_2 n^{1/2} \Big[\Big\{ \frac{t + \log(2r_n) + q_n}{n} \Big\}^{1/2} + \frac{q_n \{\log(2n)\}^{2/\xi} \{t + \log(2r_n) + q_n\}^{2/\xi}}{n} \Big]^2$$

with probability at least $1 - M_3 e^{-t}$ for any $t > 0$ and some positive constants M_2 and M_3. By condition (C2), the above expression tends to 0, and thus the desired result holds. □

Proof of Lemma 2.4. Recall that $U_{1i} = \{\epsilon_i - \mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\}\{R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\}$, and note that $\sup_{\mathcal{K}\in\Omega_n}\|U_{1i}^3\|_{\psi_{\xi/6}} \lesssim q_n^{3/2}$. By Lemma 2.5,

$$\begin{aligned}&\sup_{\mathcal{K}\in\Omega_n}\Big\{\frac{1}{n}\sum_{i=1}^{n}|U_{1i}|^3 - \mathrm{E}(|U_{11}|^3)\Big\}\\ &> M_1\Big[\Big\{\frac{t+\log(2r_n)}{n}\Big\}^{1/2} + \frac{q_n^{3/2}\{\log(2n)\}^{6/\xi}\{t+\log(2r_n)\}^{6/\xi}}{n}\Big]\end{aligned}$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_1. Because $\mathrm{E}(|U_{11}|^3)$ is uniformly bounded over $\mathcal{K} \in \Omega_n$,

$$\frac{1}{n^{3/2}}\sup_{\mathcal{K}\in\Omega_n}\sum_{i=1}^{n}|U_{1i}|^3 > M_2\Big[\frac{\{t+\log(2r_n)\}^{1/2}}{n} + \frac{q_n^{3/2}\{\log(2n)\}^{6/\xi}\{t+\log(2r_n)\}^{6/\xi}}{n^{3/2}}\Big]$$

with probability at most M_3e^{-t} for any $t > 0$ and some positive constants M_2 and M_3. Recall that U_{2i} is equal to

$$\begin{aligned}&(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\boldsymbol{X}_i - \mathrm{E}(\epsilon\boldsymbol{X} \mid R_i)\big\} + \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) - \mathrm{E}(\epsilon \mid R_i)\big\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\\ &+ \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}\big\}R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}),\end{aligned}$$

and note that $\|\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K}}\|$ is bounded, so

$$\big\|U_{2i}^3\big\|_{\psi_{\xi/3}} = O\big(1 + \big\|\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\big\|_{\psi_\xi}^3 + \big\|S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\big\|_{\psi_\xi}^3\big) \lesssim q_n^{3/2}.$$

By Lemma 2.5,

$$\begin{aligned}&\sup_{\mathcal{K}\in\Omega_n}\Big\{\frac{1}{n}\sum_{i=1}^{n}|U_{2i}|^3 - \mathrm{E}(|U_{21}|^3)\Big\}\\ &> M_4\Big[\Big\{\frac{t+\log(2r_n)}{n}\Big\}^{1/2} + \frac{q_n^{3/2}\{\log(2n)\}^{3/\xi}\{t+\log(2r_n)\}^{3/\xi}}{n}\Big]\end{aligned}$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_4. Because $\mathrm{E}(|U_{21}|^3)$ is uniformly bounded over $\mathcal{K} \in \Omega_n$,

$$\frac{1}{n^{3/2}} \sup_{\mathcal{K}\in\Omega_n} \sum_{i=1}^{n} |U_{2i}|^3 > M_5 \Big[\frac{\{t + \log(2r_n)\}^{1/2}}{n} + \frac{q_n^{3/2}\{\log(2n)\}^{3/\xi}\{t + \log(2r_n)\}^{3/\xi}}{n^{3/2}} \Big]$$

with probability at most $M_6 e^{-t}$ for any $t > 0$ and some positive constants M_5 and M_6. Similar arguments show that the same bound applies to the terms involving $\widetilde{U}_{1i}$, $\widetilde{U}_{2i}$, U_{3i}, and $\widetilde{U}_{3i}$. □

2.6.5 Proofs of Theorems 2.1 and 2.2

Proof of Theorem 2.1. We first prove the theorem under conditions (C1)–(C6) and then consider the relaxation of condition (C4) to condition (C4'). Let $\epsilon_i = Y_i - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}_i)$ for $i = 1, \ldots, n$. For any fixed $\mathcal{K}$, we can write

$$\begin{aligned}
&U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_\mathcal{K}) \\
&= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \big\{ Y_i - \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{X}_i) \big\} \big\{ R_i S_i + (1 - R_i) \widehat{\boldsymbol{\gamma}}_\mathcal{K}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} \big\} \\
&= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[\epsilon_i \big\{ R_i S_i + (1 - R_i) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} + \widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1} \boldsymbol{X}_i \big\} + \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \Big] \\
&\quad - \frac{1}{2n^{1/2}} \sum_{i=1}^{n} \mu''(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{X}_i) (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \big\{ R_i S_i + (1 - R_i) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} \big\} \\
&\quad - \frac{1}{n^{1/2}} \sum_{i=1}^{n} \{ \mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{X}_i) - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}_i) \} (1 - R_i) \boldsymbol{W}_{\mathcal{K},i}^{\mathrm{T}} (\widehat{\boldsymbol{\gamma}}_\mathcal{K} - \boldsymbol{\gamma}_{0\mathcal{K}}) \\
&= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[\epsilon_i \big\{ R_i S_i + (1 - R_i) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1} \boldsymbol{X}_i \big\} \\
&\qquad + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \Big] + o_p(1),
\end{aligned}$$

where $\widetilde{\boldsymbol{\alpha}}$ is some value between $\widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_0$. The third equality follows from the convergence of $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{I}}_{\alpha\alpha}$, and $\widehat{\boldsymbol{I}}_{\beta\alpha}$ to the true values (by Lemma 2.1 and condition (C1)) and the convergence of $n^{-1/2}\sum_{i=1}^{n}(\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} - \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1})\boldsymbol{W}_{\mathcal{K},i}R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})$ to zero (by Lemma 2.3). Note that the $o_p(1)$ term converges in mean to zero uniformly over $\mathcal{K} \in \Omega_n$. The first term on the right-hand side above can be written as

$$
\begin{aligned}
&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\big\{\epsilon_i - \mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\big\}\big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\} \\
&+ \frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i)\boldsymbol{X}_i - \mathrm{E}(\epsilon\boldsymbol{X} \mid R_i)\big\} \\
&+ \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) - \mathrm{E}(\epsilon \mid R_i)\big\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i} \\
&+ \big\{\mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}_i) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}\big\}R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\Big] \\
&+ \frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big\{(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\mathrm{E}(\epsilon\boldsymbol{X} \mid R_i) + \mathrm{E}(\epsilon \mid R_i)\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\Big\} \\
\equiv\ &\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{1i} + \frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{2i} + \frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{3i}.
\end{aligned}
$$

Note that U_{1i}, U_{2i}, and U_{3i} generally depend on the selected model $\mathcal{K}$.

By a version of the portmanteau theorem (Pollard, 2002, p. 177), it suffices to show that for any $g \in \mathcal{C}_{\mathrm{B}}^3$,

$$
\mathrm{E}\Big[g\Big\{\frac{U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})}{\sigma(\mathcal{K}^*)}\Big\}\Big] \to \mathrm{E}\{g(Z)\}, \tag{2.10}
$$

where Z is a standard normal random variable. Based on the above results and the mean-value theorem,

$$
\mathrm{E}\Big[g\Big\{\frac{U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}^*})}{\sigma(\mathcal{K}^*)}\Big\}\Big] = \int \mathrm{E}\Big[g\Big\{\frac{U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})}{\sigma(\mathcal{K})}\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big]\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K})
$$

$$= \int_{\mathcal{K}\in\Omega_n} \mathrm{E}\Big[g\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{U_{1i}+U_{2i}+U_{3i}}{\sigma(\mathcal{K})}\Big\} \mid \mathcal{K}^*=\mathcal{K}\Big]\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) + o(1), \tag{2.11}$$

where $\mathcal{P}_{\mathcal{K}^*}$ is the probability measure of $\mathcal{K}^*$. We adopt an argument similar to Lindeberg's telescoping argument for the central limit theorem (Chung, 2001, p. 211). For $i = 1,\ldots,n$, let

$$\widetilde{U}_{1i} = \mathrm{Var}(\epsilon \mid R_i, \boldsymbol{X}_i)^{1/2}\big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\}Z_{1i},$$

where $Z_{11},\ldots,Z_{1n}$ are i.i.d. standard normal random variables that are independent of the observed data. Let $V_{1i} = \widetilde{U}_{11}+\cdots+\widetilde{U}_{1,i-1}+U_{1,i+1}+\cdots+U_{1n}$ for $i=1,\ldots,n$. Note that

$$\begin{aligned}
&\mathrm{E}\Big[g\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{U_{1i}+U_{2i}+U_{3i}}{\sigma(\mathcal{K})}\Big\} - g\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{\widetilde{U}_{1i}+U_{2i}+U_{3i}}{\sigma(\mathcal{K})}\Big\} \mid \mathcal{K}^*=\mathcal{K}\Big] \\
&= \sum_{i=1}^{n}\mathrm{E}\Big[g\Big\{\frac{V_{1i}+\sum_j U_{2j}+\sum_j U_{3j}+U_{1i}}{\sigma(\mathcal{K})n^{1/2}}\Big\} - g\Big\{\frac{V_{1i}+\sum_j U_{2j}+\sum_j U_{3j}+\widetilde{U}_{1i}}{\sigma(\mathcal{K})n^{1/2}}\Big\} \mid \mathcal{K}^*=\mathcal{K}\Big] \\
&= \frac{1}{\sigma(\mathcal{K})n^{1/2}}\sum_{i=1}^{n}\mathrm{E}\Big[g'\Big\{\frac{V_{1i}+\sum_j U_{2j}+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\}(U_{1i}-\widetilde{U}_{1i}) \mid \mathcal{K}^*=\mathcal{K}\Big] \\
&\quad + \frac{1}{2\sigma(\mathcal{K})^2 n}\sum_{i=1}^{n}\mathrm{E}\Big[g''\Big\{\frac{V_{1i}+\sum_j U_{2j}+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\}(U_{1i}^2-\widetilde{U}_{1i}^2) \mid \mathcal{K}^*=\mathcal{K}\Big] \\
&\quad + \frac{1}{6\sigma(\mathcal{K})^3 n^{3/2}}\sum_{i=1}^{n}\mathrm{E}\Big\{g'''(a)U_{1i}^3 - g'''(\widetilde{a})\widetilde{U}_{1i}^3 \mid \mathcal{K}^*=\mathcal{K}\Big\}
\end{aligned} \tag{2.12}$$

for some variables a and $\widetilde{a}$. By construction, U_{1i} and $\widetilde{U}_{1i}$ are independent of V_{1i} and $\sum_j(U_{2j}+U_{3j})$ given $\mathcal{O}_1 \equiv (R_i, S_i, \boldsymbol{X}_i, \boldsymbol{A}_i)_{i=1,\ldots,n}$. The expectation in the first term on the right-hand side of (2.12) is

$$\mathrm{E}\bigg(\mathrm{E}\Big[g'\Big\{\frac{V_{1i}+\sum_j U_{2j}+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\} \mid \mathcal{O}_1, \mathcal{K}^*=\mathcal{K}\Big]\mathrm{E}(U_{1i}-\widetilde{U}_{1i} \mid \mathcal{O}_1, \mathcal{K}^*=\mathcal{K}) \mid \mathcal{K}^*=\mathcal{K}\bigg) = 0,$$

because $\mathrm{E}(U_{1i} - \widetilde{U}_{1i} \mid \mathcal{O}_1, \mathcal{K}^* = \mathcal{K}) = \mathrm{E}(U_{1i} - \widetilde{U}_{1i} \mid \mathcal{O}_1) = 0$. Likewise, the second term on the right-hand side of (2.12) is 0, because the conditional second moments of U_{1i} and $\widetilde{U}_{1i}$ given $\mathcal{O}_1$ match ($i = 1, \ldots, n$). For $\mathcal{K} \in \Omega_n$, the right-hand side of (2.12) is bounded above by

$$\zeta_{1n} \equiv M n^{-3/2} \sum_{i=1}^{n} \mathrm{E}\Big\{ \sup_{\mathcal{K} \in \Omega_n} \Big(|U_{1i}|^3 + |\widetilde{U}_{1i}|^3 \Big) \mid \mathcal{K}^* = \mathcal{K} \Big\}$$

for some positive constant M. By Lemma 2.4, $\int_{\Omega_n} \zeta_{1n} \, \mathrm{d}\mathcal{P}_{\mathcal{K}^*} \to 0$.

Next, we show that U_{2i}'s in (2.11) can be similarly replaced by normal random variables. Let $\widetilde{U}_{2i}$ be

$$\mathrm{Var}\Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1}) \big\{ \mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}) \boldsymbol{X} - \mathrm{E}(\epsilon \boldsymbol{X} \mid R_i) \big\} + \big\{ \mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}) - \mathrm{E}(\epsilon \mid R_i) \big\} \boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}} \boldsymbol{A}_{\mathcal{K},i}$$
$$+ \big\{ \mathrm{E}(\epsilon \mid R_i, \boldsymbol{X}) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} \big\} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \mid R_i, \boldsymbol{A}_i, S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}} \boldsymbol{X}_i \Big]^{1/2} Z_{2i}$$

for $i = 1, \ldots, n$, where $Z_{21}, \ldots, Z_{2n}$ are i.i.d. standard normal random variables that are independent of the observed data and $Z_{11}, \ldots, Z_{1n}$. Note that the above conditional variance is taken with respect to $\boldsymbol{X}$. We wish to show that

$$\int_{\Omega_n} \mathrm{E}\Big[g\Big\{ \sum_{i=1}^{n} \frac{\widetilde{U}_{1i} + U_{2i} + U_{3i}}{n^{1/2} \sigma(\mathcal{K})} \Big\} - g\Big\{ \sum_{i=1}^{n} \frac{\widetilde{U}_{1i} + \widetilde{U}_{2i} + U_{3i}}{n^{1/2} \sigma(\mathcal{K})} \Big\} \mid \mathcal{K}^* = \mathcal{K} \Big] \, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) = o(1). \tag{2.13}$$

Note that $n^{-1/2} \sum_{i=1}^{n} \widetilde{U}_{1i}$ can be written as $\widehat{\sigma}_1(\mathcal{K}) Z_1$, where Z_1 is a standard normal random variable independent of the observed data. By linear expansion of $\widehat{\sigma}_1(\mathcal{K})$ at $\sigma_1(\mathcal{K})$, the left-hand side of (2.13) is

$$\int_{\Omega_n} \mathrm{E}\Big[g\Big\{ \frac{n^{1/2} \sigma_1(\mathcal{K}) Z_1 + \sum_i (U_{2i} + U_{3i})}{n^{1/2} \sigma(\mathcal{K})} \Big\}$$
$$- g\Big\{ \frac{n^{1/2} \sigma_1(\mathcal{K}) Z_1 + \sum_i (\widetilde{U}_{2i} + U_{3i})}{n^{1/2} \sigma(\mathcal{K})} \Big\} \mid \mathcal{K}^* = \mathcal{K} \Big] \, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) \tag{2.14}$$

up to an additive term bounded above by $\sup|\sigma(\mathcal{K})^{-1}g'|\mathrm{E}\{\sup_{\mathcal{K}\in\Omega_n}|\widehat{\sigma}_1(\mathcal{K})-\sigma_1(\mathcal{K})|Z_1\}$, which tends to 0 by Lemma 2.2. For any bounded variable B, we have

$$\begin{aligned}
&\int_{\Omega_n}\mathrm{E}(B\mid\mathcal{K}^*=\mathcal{K})-\mathrm{E}(B\mid\mathcal{K}_0^*=\mathcal{K})\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K})\\
&=\int_{\Omega_n}\mathrm{E}\{BI(\mathcal{K}_0^*\neq\mathcal{K}^*)\mid\mathcal{K}^*=\mathcal{K}\}\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K})-\int_{\Omega_n}\mathrm{E}\{BI(\mathcal{K}_0^*\neq\mathcal{K}^*)\mid\mathcal{K}_0^*=\mathcal{K}\}\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K})\\
&\leq\sup|B|\Big\{1+\sup_{\mathcal{K}\in\Omega_n}\frac{P(\mathcal{K}^*=\mathcal{K})}{P(\mathcal{K}_0^*=\mathcal{K})}\Big\}P(\mathcal{K}_0^*\neq\mathcal{K}^*)=o(1),
\end{aligned}$$

where $\mathcal{K}_0^*=\mathcal{K}^*(\mathcal{S}-\mathcal{X}\boldsymbol{\gamma}_{0X},\mathcal{A})$, and the last equality follows from condition (C5). Therefore, the event $\{\mathcal{K}^*=\mathcal{K}\}$ in the conditional expectation in (2.14) can be replaced by $\{\mathcal{K}_0^*=\mathcal{K}\}$.

Let $V_{2i}=\widetilde{U}_{21}+\cdots+\widetilde{U}_{2,i-1}+U_{2,i+1}+\cdots+U_{2n}$ for $i=1,\ldots,n$. The term inside the integration of the left-hand side of (2.13) is up to a vanishing term equal to

$$\begin{aligned}
&\sum_{i=1}^{n}\mathrm{E}\Big[g\Big\{\frac{V_{2i}+n^{1/2}\sigma_1(\mathcal{K})Z_1+U_{2i}+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\}\\
&\qquad-g\Big\{\frac{V_{2i}+n^{1/2}\sigma_1(\mathcal{K})Z_1+\widetilde{U}_{2i}+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\}\mid\mathcal{K}_0^*=\mathcal{K}\Big]\\
&=\frac{1}{\sigma(\mathcal{K})n^{1/2}}\sum_{i=1}^{n}\mathrm{E}\Big[g'\Big\{\frac{V_{2i}+n^{1/2}\sigma_1(\mathcal{K})Z_1+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\}(U_{2i}-\widetilde{U}_{2i})\mid\mathcal{K}_0^*=\mathcal{K}\Big]\\
&\quad+\frac{1}{2\sigma(\mathcal{K})^2n}\sum_{i=1}^{n}\mathrm{E}\Big[g''\Big\{\frac{V_{2i}+n^{1/2}\sigma_1(\mathcal{K})Z_1+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\}(U_{2i}^2-\widetilde{U}_{2i}^2)\mid\mathcal{K}_0^*=\mathcal{K}\Big]\\
&\quad+\frac{1}{6\sigma(\mathcal{K})^3n^{3/2}}\sum_{i=1}^{n}\mathrm{E}\Big\{g'''(b)U_{2i}^3-g'''(\widetilde{b})\widetilde{U}_{2i}^3\mid\mathcal{K}_0^*=\mathcal{K}\Big\}
\end{aligned}\tag{2.15}$$

for some variables b and $\widetilde{b}$. Let $\mathcal{O}_2=(R_i,\boldsymbol{A}_i,S_i-\boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)_{i=1,\ldots,n}$. Since the event $\{\mathcal{K}_0^*=\mathcal{K}\}$ is implied by $\mathcal{O}_2$, we have

$$\mathrm{E}\Big[g'\Big\{\frac{V_{2i}+n^{1/2}\sigma_1(\mathcal{K})Z_1+\sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\}(U_{2i}-\widetilde{U}_{2i})\mid\mathcal{K}_0^*=\mathcal{K}\Big]$$

$$= \mathrm{E}\Big(\mathrm{E}\Big[g'\Big\{\frac{V_{2i} + n^{1/2}\sigma_1(\mathcal{K})Z_1 + \sum_j U_{3j}}{\sigma(\mathcal{K})n^{1/2}}\Big\} \mid \mathcal{O}_2, \mathcal{K}_0^* = \mathcal{K}\Big]\mathrm{E}(U_{2i} - \widetilde{U}_{2i} \mid \mathcal{O}_2) \mid \mathcal{K}_0^* = \mathcal{K}\Big) = 0.$$

Likewise, the second term on the right-hand side of (2.15) is zero because the conditional second moments of U_{2i} and $\widetilde{U}_{2i}$ match. By Lemma 2.4, the third term on the right-hand side of (2.15) is bounded by some positive variable ζ_{2n} such that $\int_{\Omega_n} \zeta_{2n}\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*} \to 0$, and (2.13) holds.

Let $\widetilde{U}_{3i} = \mathrm{Var}\{(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1})\mathrm{E}(\epsilon\boldsymbol{X} \mid R) + \mathrm{E}(\epsilon \mid R)\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i} \mid \boldsymbol{A}_{\mathcal{K},i}\}^{1/2}Z_{3i}$ for $i = 1, \ldots, n$, where $Z_{31}, \ldots, Z_{3n}$ are i.i.d. standard normal variables that are independent of the observed data and $(Z_{1i}, Z_{2i})_{i=1,\ldots,n}$. Similarly, we wish to show that

$$\int_{\Omega_n} \mathrm{E}\Big[g\Big\{\sum_{i=1}^n \frac{\widetilde{U}_{1i} + \widetilde{U}_{2i} + U_{3i}}{n^{1/2}\sigma(\mathcal{K})}\Big\} - g\Big\{\sum_{i=1}^n \frac{\widetilde{U}_{1i} + \widetilde{U}_{2i} + \widetilde{U}_{3i}}{n^{1/2}\sigma(\mathcal{K})}\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big]\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) = o(1). \tag{2.16}$$

Write $n^{-1/2}\sum_{i=1}^n \widetilde{U}_{2i} = \widehat{\sigma}_2(\mathcal{K})Z_2$, where Z_2 is a standard normal random variable independent of the observed data and Z_1. By the mean-value theorem and Lemma 2.2, the left-hand side of (2.16) is

$$\int_{\Omega_n} \mathrm{E}\Big[g\Big\{\frac{n^{1/2}\sigma_1(\mathcal{K})Z_1 + n^{1/2}\sigma_2(\mathcal{K})Z_2 + \sum_{i=1}^n U_{3i}}{n^{1/2}\sigma(\mathcal{K})}\Big\} - g\Big\{\frac{n^{1/2}\sigma_1(\mathcal{K})Z_1 + n^{1/2}\sigma_2(\mathcal{K})Z_2 + \sum_{i=1}^n \widetilde{U}_{3i}}{n^{1/2}\sigma(\mathcal{K})}\Big\} \mid \mathcal{K}_0^* = \mathcal{K}\Big]\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) + o(1).$$

Then, we show that the conditional expectation above can be replaced by a marginal expectation. For any bounded function h of $(\mathcal{R}, \widetilde{\mathcal{A}})$, where $\mathcal{R} \equiv (R_1, \ldots, R_n)$ and $\widetilde{\mathcal{A}} \equiv (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n)$, we have

$$\begin{aligned}&\mathrm{E}\{h(\mathcal{R}, \widetilde{\mathcal{A}}) \mid \mathcal{K}_0^* = \mathcal{K}\} \\ &= \int \sum_{\boldsymbol{r}} h(\boldsymbol{r}, \boldsymbol{a}) P(\mathcal{R} = \boldsymbol{r} \mid \widetilde{\mathcal{A}} = \boldsymbol{a}, \mathcal{K}_0^* = \mathcal{K}) f_{\mathcal{A}}(\boldsymbol{a} \mid \mathcal{K}_0^* = \mathcal{K})\,\mathrm{d}\lambda(\boldsymbol{a})\end{aligned}$$

$$= \int \sum_{\boldsymbol{r}} h(\boldsymbol{r}, \boldsymbol{a}) \frac{P(\mathcal{K}_0^* = \mathcal{K} \mid \widetilde{\mathcal{A}} = \boldsymbol{a}, \mathcal{R} = \boldsymbol{r}) P(\mathcal{R} = \boldsymbol{r}) f_{\mathcal{A}}(\boldsymbol{a})}{P(\mathcal{K}_0^* = \mathcal{K})} \, \mathrm{d}\lambda(\boldsymbol{a})$$

$$= \mathrm{E}\{h(\mathcal{R}, \widetilde{\mathcal{A}})\} + \int \sum_{\boldsymbol{r}} h(\boldsymbol{r}, \boldsymbol{a}) \Big\{ \frac{P(\mathcal{K}_0^* = \mathcal{K} \mid \widetilde{\mathcal{A}} = \boldsymbol{a}, \mathcal{R} = \boldsymbol{r})}{P(\mathcal{K}_0^* = \mathcal{K})} - 1 \Big\} P(\mathcal{R} = \boldsymbol{r}) f_{\mathcal{A}}(\boldsymbol{a}) \, \mathrm{d}\lambda(\boldsymbol{a}),$$

where $f_{\mathcal{A}}$ is the density of $\widetilde{\mathcal{A}}$ with respect to some dominating measure λ. By condition (C6), the second term on the right-hand side above converges to 0 uniformly over $\mathcal{K} \in \Omega_n$. Let $V_{3i} = \widetilde{U}_{31} + \cdots + \widetilde{U}_{3,i-1} + U_{3,i+1} + \cdots + U_{3n}$ for $i = 1, \ldots, n$. The term inside the integration of the left-hand side of (2.16) is up to a vanishing term equal to

$$\sum_{i=1}^{n} \mathrm{E}\Big[g\Big\{ \frac{V_{3i} + n^{1/2}\sigma_1(\mathcal{K}) Z_1 + n^{1/2}\sigma_2(\mathcal{K}) Z_2 + U_{3i}}{\sigma(\mathcal{K}) n^{1/2}} \Big\} - g\Big\{ \frac{V_{3i} + n^{1/2}\sigma_1(\mathcal{K}) Z_1 + n^{1/2}\sigma_2(\mathcal{K}) Z_2 + \widetilde{U}_{3i}}{\sigma(\mathcal{K}) n^{1/2}} \Big\} \Big].$$

Based on an expansion similar to (2.12) and (2.15), we can show that the above expression tends to 0 by Lemma 2.4 and the fact that the first two moments of U_{3i} and $\widetilde{U}_{3i}$ match. Combining the above results, we have

$$\begin{aligned} \mathrm{E}\Big[g\Big\{ \sum_{i=1}^{n} \frac{U_{1i} + U_{2i} + U_{3i}}{n^{1/2}\sigma(\mathcal{K})} \Big\} \mid \mathcal{K}^* = \mathcal{K} \Big] &= \mathrm{E}\Big[g\Big\{ \sum_{i=1}^{n} \frac{\widetilde{U}_{1i} + \widetilde{U}_{2i} + \widetilde{U}_{3i}}{n^{1/2}\sigma(\mathcal{K})} \Big\} \mid \mathcal{K}^* = \mathcal{K} \Big] + o(1) \\ &= \mathrm{E}\Big[g\Big\{ \frac{\sigma_1(\mathcal{K}) Z_1 + \sigma_2(\mathcal{K}) Z_2 + \sigma_3(\mathcal{K}) Z_3}{\sigma(\mathcal{K})} \Big\} \Big] + o(1) \end{aligned}$$

uniformly over $\mathcal{K} \in \Omega_n$, where Z_3 is a standard normal random variable independent of Z_1, Z_2, and the observed data. Because $\sigma_1(\mathcal{K}) Z_1 + \sigma_2(\mathcal{K}) Z_2 + \sigma_3(\mathcal{K}) Z_3$ is normal with mean 0 and variance $\sigma^2(\mathcal{K})$, the desired convergence (2.10) follows.

We consider the relaxation of condition (C4). Under condition (C4'), $(Y, \boldsymbol{X})$ may depend on some auxiliary variables, and $\mathrm{E}(U_{1i} - \widetilde{U}_{1i} \mid \mathcal{O}_1)$, $\mathrm{E}(U_{1i}^2 - \widetilde{U}_{1i}^2 \mid \mathcal{O}_1)$, $\mathrm{E}(U_{2i} - \widetilde{U}_{2i} \mid \mathcal{O}_2)$, and $\mathrm{E}(U_{2i}^2 - \widetilde{U}_{2i}^2 \mid \mathcal{O}_2)$ may be nonzero. Nevertheless, the selection probability of the auxiliary variables that are associated with $(Y, \boldsymbol{X})$ vanishes, so that similar arguments

apply with $\mathcal{O}_1$ and $\mathcal{O}_2$ redefined to not include those dependent components of $\boldsymbol{A}$. For any bounded random variable B,

$$\begin{aligned}
&\int_{\Omega_n} \mathrm{E}(B \mid \mathcal{K}^* = \mathcal{K}) - \mathrm{E}(B \mid \mathcal{K}^* \in \overline{\mathcal{K}})\, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) \\
&= \int_{\Omega_n} \big\{\mathrm{E}(B \mid \mathcal{K}^* = \mathcal{K}) - \mathrm{E}(B \mid \mathcal{K}^* \in \overline{\mathcal{K}}, \mathcal{K}^* \neq \mathcal{K})\big\} P(\mathcal{K}^* \neq \mathcal{K} \mid \mathcal{K}^* \in \overline{\mathcal{K}})\, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) \\
&\leq 2 \sup |B| \sum_{\mathcal{K} \in \Omega_n} P(\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}) \frac{P(\mathcal{K}^* = \mathcal{K})}{P(\mathcal{K}^* \in \overline{\mathcal{K}})} \\
&\leq 2 \sup |B| \sum_{\mathcal{K} \in \Omega_n} P(\mathcal{K}^* \neq \mathcal{K}, \mathcal{K}^* \in \overline{\mathcal{K}}) = o(1).
\end{aligned}$$

Therefore, the event $\{\mathcal{K}^* = \mathcal{K}\}$ in the conditional expectation on the right-hand side of (2.11) can be replaced by $\mathcal{K}^* \in \overline{\mathcal{K}}$. The original arguments can then be applied to this updated version of (2.11), with $\mathcal{O}_1$ replaced by $(R_i, S_i, \boldsymbol{X}_i, \boldsymbol{A}_{\mathcal{M}_n,i})_{i=1,\ldots,n}$ and $\mathcal{O}_2$ replaced by $(R_i, \boldsymbol{A}_{\mathcal{M}_n,i}, S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}} \boldsymbol{X}_i)_{i=1,\ldots,n}$. □

Proof of Theorem 2.2. Let $\widehat{\sigma}_{0i}(\mathcal{K}) = \epsilon_i\{R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} \boldsymbol{I}_{\alpha\alpha}^{-1} \boldsymbol{X}_i\} + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i})$ for $i = 1, \ldots, n$ and $\widehat{\sigma}_0^2(\mathcal{K}) = n^{-1} \sum_{i=1}^n \widehat{\sigma}_{0i}^2(\mathcal{K})$. Note that $n^{-1} \sum_{i=1}^n \{\widehat{\sigma}_i(\mathcal{K}) - \overline{\sigma}(\mathcal{K})\}^2$ is equal to

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n \big\{\widehat{\sigma}_i(\mathcal{K}) - \widehat{\sigma}_{0i}(\mathcal{K}) + \widehat{\sigma}_{0i}(\mathcal{K})\big\}^2 - \overline{\sigma}(\mathcal{K})^2 \\
&= \frac{1}{n} \sum_{i=1}^n \widehat{\sigma}_{0i}^2(\mathcal{K}) + \frac{2}{n} \sum_{i=1}^n \big\{\widehat{\sigma}_i(\mathcal{K}) - \widehat{\sigma}_{0i}(\mathcal{K})\big\} \widehat{\sigma}_{0i}(\mathcal{K}) + \frac{1}{n} \sum_{i=1}^n \big\{\widehat{\sigma}_i(\mathcal{K}) - \widehat{\sigma}_{0i}(\mathcal{K})\big\}^2 - \overline{\sigma}(\mathcal{K})^2.
\end{aligned} \tag{2.17}$$

Using the arguments in the proof of Lemma 2.2, we can show that the first term on the right-hand side of (2.17) converges in mean to $\sigma_0^2(\mathcal{K})$ uniformly over $\mathcal{K} \in \Omega_n$. We then show that the remaining terms in the expression converge in mean to 0 uniformly. Note

that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\big\{\widehat{\sigma}_i(\mathcal{K})-\widehat{\sigma}_{0i}(\mathcal{K})\big\}\widehat{\sigma}_{0i}(\mathcal{K})\\
&=\frac{1}{n}\sum_{i=1}^{n}\Big[\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\big(\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}-\boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\big)\boldsymbol{X}_i+\big(\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}-\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\big)\boldsymbol{W}_{\mathcal{K},i}R_i(S_i-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\\
&\quad-\mu'(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)^{\mathrm{T}}\boldsymbol{X}_i\big\{R_iS_i+(1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}+\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\}\\
&\quad+\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}(\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}-\boldsymbol{\gamma}_{0\mathcal{K}})^{\mathrm{T}}(1-R_i)\boldsymbol{W}_{\mathcal{K},i}-\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}R_i(\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}-\boldsymbol{\gamma}_{0\mathcal{K}})^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\\
&\quad-\frac{1}{2}\mu''(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{X}_i)(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)\big\{R_iS_i+(1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}+\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}\boldsymbol{X}_i\big\}\\
&\quad-\{\mu(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{X}_i)-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}(1-R_i)\boldsymbol{W}_{\mathcal{K},i}^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}-\boldsymbol{\gamma}_{0\mathcal{K}})\Big]\widehat{\sigma}_{0i}(\mathcal{K}).
\end{aligned}
$$

Consider $n^{-1}\sum_{i=1}^{n}\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\big(\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}-\boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\big)\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})$. We have

$$
\begin{aligned}
&\sup_{\mathcal{K}\in\Omega_n}\left|\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\big(\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}-\boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\big)\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})\right|\\
&\le\sup_{\mathcal{K}\in\Omega_n}\left\|\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\alpha\alpha}^{-1}-\boldsymbol{I}_{\beta\alpha}^{\mathrm{T}}\boldsymbol{I}_{\alpha\alpha}^{-1}\right\|\sup_{\mathcal{K}\in\Omega_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})\right\|. \qquad (2.18)
\end{aligned}
$$

By Lemma 2.1, the first term on the right-hand side above is bounded by

$$
M_1\left\{\Big(\frac{t+\log r_n}{n}\Big)^{1/2}+\frac{q_n^{1/2}(\log n)^{1/\xi}(t+\log r_n)^{1/\min(1,\xi)}}{n}\right\}
$$

with probability at least $1-M_2e^{-t}$ for any $t>0$ and some positive constants M_1 and M_2. Also, we have

$$
\begin{aligned}
&\sup_{\mathcal{K}\in\Omega_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})\right\|\\
&\le\sup_{\mathcal{K}\in\Omega_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})-\mathrm{E}\big[\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})\big]\right\|\\
&\quad+\sup_{\mathcal{K}\in\Omega_n}\left\|\mathrm{E}\big[\{Y_i-\mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})\big]\right\|.
\end{aligned}
$$

Under conditions (C1) and (C2), $\sup_{\mathcal{K}\in\Omega_n} \|\mathrm{E}[\{Y_i - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})]\| = O(1)$. Since $\|\{Y_i - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})\|_{\psi_{\xi/3}} \lesssim q_n^{1/2}$, by Lemma 2.5,

$$\begin{aligned}
&\sup_{\mathcal{K}\in\Omega_n} \Big\| \frac{1}{n}\sum_{i=1}^{n} \{Y_i - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K}) - \mathrm{E}\big[\{Y_i - \mu(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\}\boldsymbol{X}_i\widehat{\sigma}_{0i}(\mathcal{K})\big] \Big\| \\
&\le M_3 \left[\left\{ \frac{t + \log(2r_n)}{n} \right\}^{1/2} + \frac{q_n^{1/2}\{\log(2n)\}^{3/\xi}\{t + \log(2r_n)\}^{3/\xi}}{n} \right]
\end{aligned}$$

with probability at least $1 - 3e^{-t}$ for large enough n and t and some positive constant M_3. Under the rates of q_n and r_n given in condition (C2), we conclude that the left-hand side of (2.18) converges to 0 in mean uniformly over $\mathcal{K} \in \Omega_n$. Similar arguments show that

$$\begin{aligned}
&\sup_{\mathcal{K}\in\Omega_n} \left| \frac{1}{n}\sum_{i=1}^{n} \{\widehat{\sigma}_i(\mathcal{K}) - \widehat{\sigma}_{0i}(\mathcal{K})\}\widehat{\sigma}_{0i}(\mathcal{K}) \right| \\
&\le M_4 \left[\frac{t + \log(2r_n) + q_n}{n} + \frac{q_n^{5/2}\{\log(2n)\}^{6/\xi}\{t + \log(2r_n) + q_n\}^{6/\xi}}{n^2} \right]
\end{aligned}$$

with probability at least $1 - M_5 e^{-t}$ for large enough n and t, where M_4 and M_5 are some positive constants. Therefore, by condition (C2), the second term on the right-hand side of (2.17) converges to 0 in mean uniformly over $\mathcal{K} \in \Omega_n$. Similar arguments show that the third and fourth terms in that expression also converge to 0 in mean uniformly over $\mathcal{K} \in \Omega_n$. The desired result follows. □

2.6.6 Additional Numerical Results

Table 2.1: Rejection probabilities and references of significant proteins in the TCGA colorectal adenocarcinoma analysis

Protein	**Proposed Method**	**Complete-case**	**Covariate-only**	**Reference**
`IRS1`	4.53E−08	2.17E−06	1.57E−06	Esposito et al. (2012)
`Caspase-7_cleavedD198`	5.87E−07	1.17E−05	3.24E−06	N/A

Continued on next page

Table 2.1 – *Continued from previous page*

Protein	Proposed Method	Complete-case	Covariate-only	Reference
eIF4E	9.60E−06	1.26E−04	8.45E−05	Diab-Assaf et al. (2015)
c-Myc	1.14E−05	1.80E−04	1.42E−04	Erisman et al. (1988)
Cyclin_E1	8.89E−05	3.21E−04	2.28E−04	Qi et al. (2015)
p38_MAPK	2.50E−04	7.13E−04	7.76E−04	Thyagarajan et al. (2010)
XRCC1	4.45E−04	2.31E−03	2.29E−03	Huang et al. (2013)
GAB2	4.79E−04	8.98E−04	7.46E−04	Ding et al. (2015)
Paxillin	6.78E−04	5.02E−03	4.83E−03	Zhao et al. (2015)
PREX1	7.84E−04	9.32E−04	8.09E−04	N/A
Bcl-2	8.22E−04	1.14E−03	9.82E−04	Hague et al. (1994)
Bax	8.53E−04	3.47E−04	2.87E−04	Pryczynicz et al. (2014)
PKC-delta_pS664	1.19E−03	2.95E−03	3.27E−03	N/A
YB-1	1.53E−03	6.36E−03	6.02E−03	Tsofack et al. (2011)
NF-kB-p65_pS536	1.78E−03	3.64E−03	3.48E−03	N/A
GATA3	1.93E−03	4.52E−03	4.42E−03	Wang et al. (2020)
Rictor_pT1135	3.64E−03	1.23E−02	1.17E−02	N/A
HER3	4.81E−03	1.60E−02	1.58E−02	Kountourakis et al. (2006)
PRAS40_pT246	5.94E−03	9.35E−03	9.12E−03	N/A
GAPDH	6.85E−03	1.38E−02	1.30E−02	Tang et al. (2012)
INPP4B	7.15E−03	1.51E−02	1.80E−02	Yang et al. (2020)
YAP_pS127	7.70E−03	3.43E−02	3.81E−02	N/A
4E-BP1	7.93E−03	1.69E−02	1.70E−02	Diab-Assaf et al. (2015)
FASN	8.06E−03	5.14E−02	5.18E−02	N/A
Tuberin	8.20E−03	8.90E−03	9.46E−03	N/A
CDK1_pY15	8.29E−03	7.78E−02	8.36E−02	N/A
p53	1.36E−02	5.46E−02	5.30E−02	Rodrigues et al. (1990)
Dvl3	1.40E−02	1.07E−02	1.01E−02	N/A
MAPK_pT202_Y204	1.56E−02	4.97E−02	5.03E−02	N/A
S6_pS240_S244	1.96E−02	4.53E−02	4.27E−02	N/A
PKC-alpha	2.04E−02	4.72E−02	5.00E−02	N/A
Rab25	2.27E−02	2.59E−02	2.52E−02	Nam et al. (2010)
Rb	2.69E−02	1.40E−02	1.57E−02	Yamamoto et al. (1999)
Chk1	3.10E−02	6.42E−02	7.11E−02	Bertoni et al. (1999)
Src_pY527	3.13E−02	6.58E−02	7.24E−02	N/A

Continued on next page

Table 2.1 – *Continued from previous page*

Protein	**Proposed Method**	**Complete-case**	**Covariate-only**	**Reference**
ARID1A	3.29E−02	1.59E−02	1.40E−02	Wei et al. (2014)
GATA6	3.51E−02	4.43E−02	4.32E−02	Belaguli et al. (2010)
p70S6K_pT389	3.62E−02	3.30E−02	3.73E−02	N/A
TSC1	4.10E−02	3.48E−02	3.78E−02	N/A
4E-BP1_pT70	4.23E−02	2.31E−02	2.34E−02	N/A
CDK1	4.36E−02	7.11E−02	7.73E−02	Gan et al. (2017)
Akt	4.52E−02	5.09E−02	4.99E−02	Agarwal et al. (2013)
PKC-pan_BetaII_pS660	4.87E−02	4.45E−02	4.74E−02	N/A
LKB1	5.23E−02	1.79E−02	2.07E−02	He et al. (2014)
eEF2K	5.31E−02	1.07E−02	1.05E−02	N/A
MEK1_pS217_S221	6.58E−02	4.17E−02	4.42E−02	N/A

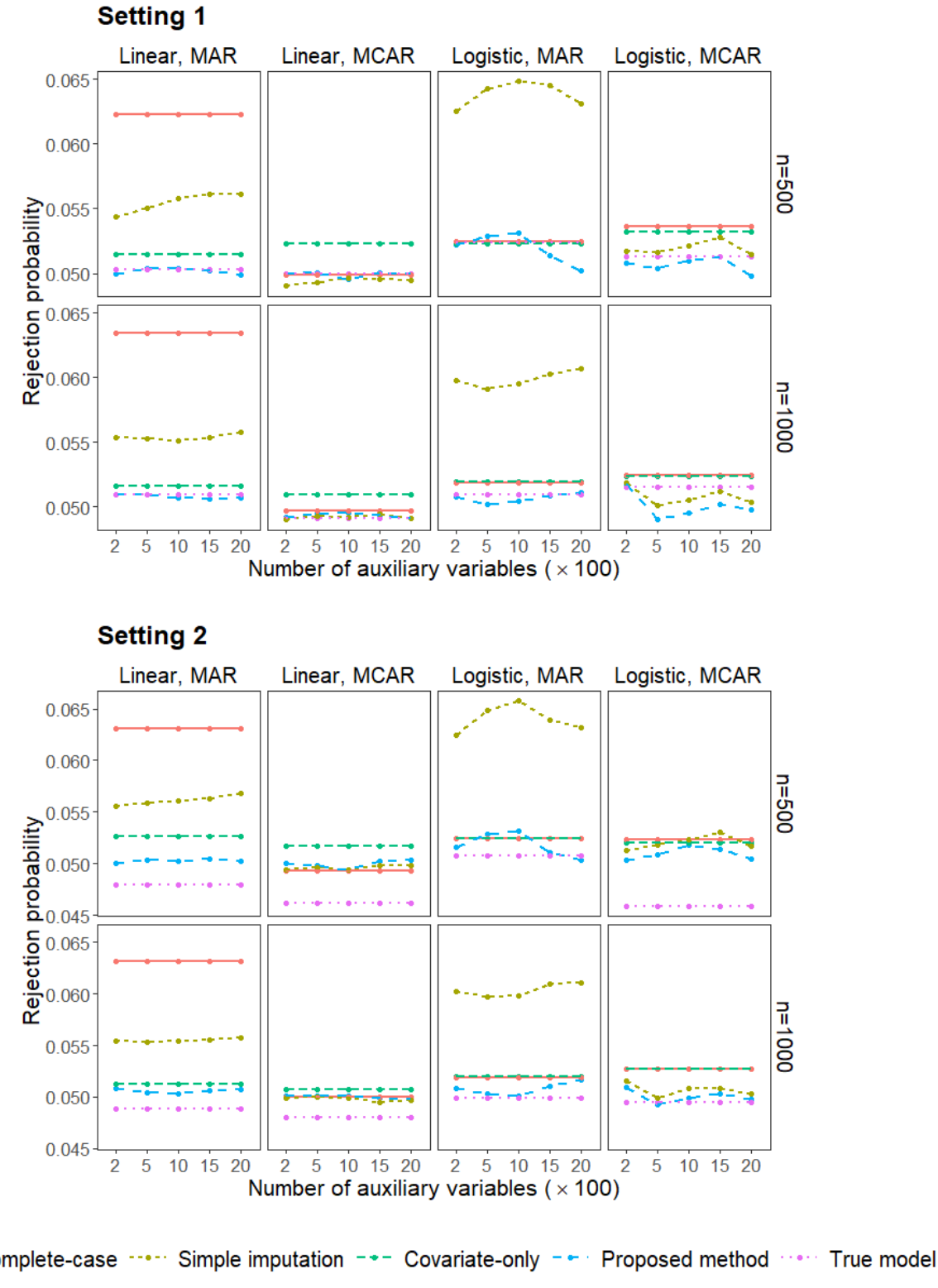


Figure 2.4: Rejection probabilities under a missing proportion of 30% and the null hypothesis.

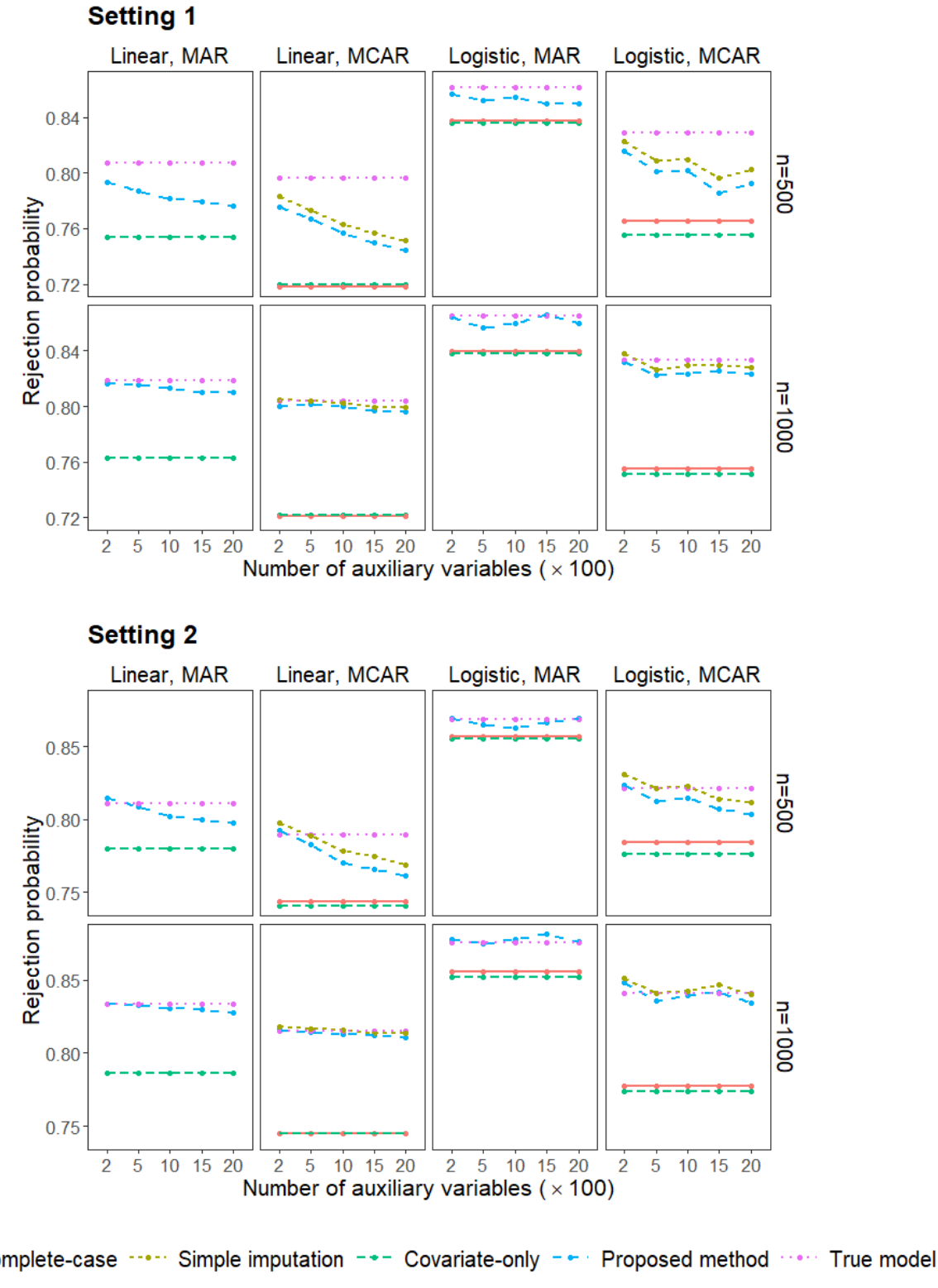
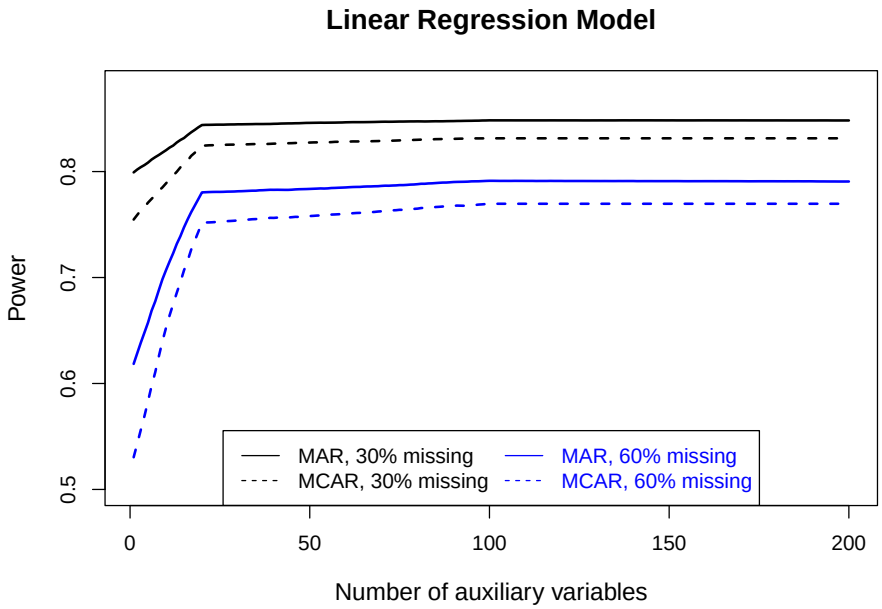


Figure 2.5: Rejection probabilities under a missing proportion of 30% and the alternative hypothesis.



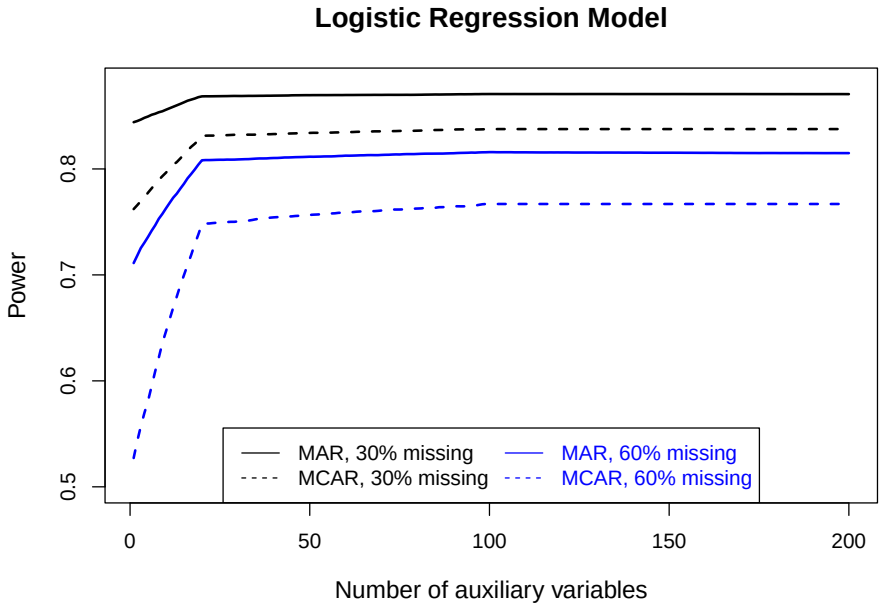


Figure 2.6: Asymptotic power over different numbers of auxiliary variables.

Chapter 3

Score Tests with an Incomplete Covariate in Semiparametric Models for Censored Data

3.1 Methodology

Let T denote an event time, S a covariate of interest, $\boldsymbol{X}$ a vector of other covariates, and $\boldsymbol{A}$ a potentially high-dimensional vector of auxiliary variables. Assume that the cumulative hazard function of T conditional on $(\boldsymbol{X}, S)$ takes the form

$$\Lambda(t \mid \boldsymbol{X}, S) = G\{\Lambda(t) \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \beta S)\}, \tag{3.1}$$

where $\Lambda(\cdot)$ is an unknown increasing function in $[0, \tau]$ with $\Lambda(0) = 0$, $G(\cdot)$ is a prespecified transformation function that is strictly increasing with $G(0) = 0$, and $\boldsymbol{\alpha}$ and β are regression parameters, and τ is the end-of-study time. Under (3.1), the model of T can

be expressed as a linear transformation model, with

$$\log \Lambda(T) = -\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} - \beta S + \epsilon,$$

where ϵ is an error term with $P(\epsilon < x) = 1 - \exp[-G\{\exp(x)\}]$. Thus, β can be interpreted as the linear effect of the covariate S on a transformation of T.

Suppose that the survival time is possibly right-censored at C, which is assumed to be independent with $(T, S, \boldsymbol{A})$ given $\boldsymbol{X}$. Let $Y = \min(T, C)$ and $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Also, suppose that S may be missing, and let R be the indicator of whether S is observed, i.e., $R = 1$ if S is observed, and $R = 0$ otherwise. We assume that R is conditionally independent of $(S, \boldsymbol{A})$ given $(T, C, \boldsymbol{X})$. Under this assumption, S is missing at random. The observed data from a random sample of n subjects consist of $(Y_i, \Delta_i, \boldsymbol{X}_i, \boldsymbol{A}_i, R_i S_i, R_i)$ for $i = 1, \ldots, n$.

3.1.1 Imputation Score Test

Under the transformation model (3.1), we are interested in testing the null hypothesis $H_0 : \beta = 0$, that the covariate of interest S does not have an effect on the (transformed) hazard of T. Since S is subject to missing, we propose to first fit a working model of S against $(\boldsymbol{X}, \boldsymbol{A})$ and impute the missing values of S based on the model. Because the auxiliary variables $\boldsymbol{A}$ may be high-dimensional, we propose to select a low-dimensional subset of the components of $\boldsymbol{A}$ to construct the model of S. Correct specification of the model of S is not necessary, since the model is only of secondary interest. Suppose $\boldsymbol{A}$ is p-dimensional. We choose a subset $\mathcal{K} \subset \{1, \ldots, p\}$, and let $\boldsymbol{W}_{\mathcal{K}}$ denote the vector that consists of $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ indexed by $\mathcal{K}$. We fit a working model of $S = \boldsymbol{\gamma}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}} + \delta$, where δ is a mean-zero error term, and $\boldsymbol{\gamma}_{\mathcal{K}}$ is a vector of regression parameters.

We then perform a score test based on the outcome model (3.1) and the selected working model of S. The reason that we choose the score test over the Wald test and the likelihood ratio test is that the score test is performed under the null hypothesis, and the estimating procedure is much simpler. Let $f(Y \mid \boldsymbol{X}, S; \beta, \boldsymbol{\alpha}, \Lambda)$ and $g(S \mid \boldsymbol{W}_{\mathcal{K}}; \boldsymbol{\gamma}_{\mathcal{K}})$ denote the density functions of Y and S, respectively. The log-likelihood function concerning parameters $(\beta, \boldsymbol{\alpha}, \Lambda, \boldsymbol{\gamma}_{\mathcal{K}})$ is

$$\begin{aligned}\log L(\beta, \boldsymbol{\alpha}, \Lambda, \boldsymbol{\gamma}_{\mathcal{K}}) = &\sum_{i=1}^{n} R_i \big\{ \log f(Y_i \mid \boldsymbol{X}_i, S_i; \beta, \boldsymbol{\alpha}, \Lambda) + \log g(S_i \mid \boldsymbol{W}_{\mathcal{K},i}; \boldsymbol{\gamma}_{\mathcal{K}}) \big\} \\ &+ \sum_{i=1}^{n} (1 - R_i) \log \int f(Y_i \mid \boldsymbol{X}_i, s; \beta, \boldsymbol{\alpha}, \Lambda) g(s \mid \boldsymbol{W}_{\mathcal{K},i}; \boldsymbol{\gamma}_{\mathcal{K}}) \, \mathrm{d}s. \quad (3.2)\end{aligned}$$

We propose to estimate $\boldsymbol{\gamma}_{\mathcal{K}}$ by solving $\sum_{i=1}^{n} R_i (S_i - \boldsymbol{\gamma}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \boldsymbol{W}_{\mathcal{K},i} = \mathbf{0}$, and let $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$ denote the estimator. We only use the subjects with $R_i = 1$ because $f(Y \mid \boldsymbol{X}, S)$ does not involve S under H_0, so subjects with $R_i = 0$ do not contribute to the estimation of $\boldsymbol{\gamma}_{\mathcal{K}}$. We adopt the nonparametric maximum likelihood approach of Zeng and Lin (2007) to estimate $\boldsymbol{\alpha}$ and Λ under H_0. Here, we treat Λ as a step function with jumps only at the observed survival times. Let $t_1 < \cdots < t_m$ denote the set of observed survival times with m being the number of unique observed survival times, and λ_k be the jump size at t_k for $k = 1, \ldots, m$. The log-likelihood function pertaining to $(\boldsymbol{\alpha}, \Lambda)$ is

$$\sum_{i=1}^{n} \Delta_i \left[\log G' \Big\{ \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i) \sum_{t_k \le Y_i} \lambda_k \Big\} + \log \lambda_{k(i)} + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i \right] - G \Big\{ \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i) \sum_{t_k \le Y_i} \lambda_k \Big\}, \quad (3.3)$$

where $k(\cdot)$ is a map defined on $\{i = 1, \ldots, n : \Delta_i = 1\}$ such that $\lambda_{k(i)}$ is the jump size at time Y_i, and $G'(\cdot)$ is the first derivative of $G(\cdot)$.

Let $\widehat{\boldsymbol{\zeta}} \equiv (\widehat{\boldsymbol{\alpha}}, \widehat{\lambda}_1, \ldots, \widehat{\lambda}_m)$ be the maximizer of (3.3), and let $\boldsymbol{\zeta} = (\boldsymbol{\alpha}, \lambda_1, \ldots, \lambda_m)$. Also, let $\xi_i(\boldsymbol{\zeta}) = \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}_i) \sum_{t_k \le Y_i} \lambda_k$, $G_i(\boldsymbol{\zeta}) = G\{\xi_i(\boldsymbol{\zeta})\}$, and $G''(\cdot)$ denote the second deriva-

tive of $G(\cdot)$. The (scaled) score statistic for β is

$$U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = n^{-1/2} \sum_{i=1}^{n} \Big\{ \Delta_i + \Delta_i \frac{G_i''(\widehat{\boldsymbol{\zeta}})}{G_i'(\widehat{\boldsymbol{\zeta}})} \xi_i(\widehat{\boldsymbol{\zeta}}) - G_i'(\widehat{\boldsymbol{\zeta}}) \xi_i(\widehat{\boldsymbol{\zeta}}) \Big\} \big\{ R_i S_i + (1 - R_i) \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} \big\}.$$

Let $\boldsymbol{\alpha}_0$ and Λ_0 denote the true values of $\boldsymbol{\alpha}$ and Λ, respectively. Let $\boldsymbol{\zeta}_0 = (\boldsymbol{\alpha}_0, \lambda_{0,1}, \ldots, \lambda_{0,m})$, where $\lambda_{0,k} = \Lambda_0(t_k) - \Lambda_0(t_{k-1})$ for $k = 1, \ldots, m$ with $t_0 = 0$. For a given $\mathcal{K}$, define $\boldsymbol{\gamma}_{0\mathcal{K}} \equiv \arg\min_\gamma \mathrm{E}\{R(S - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}})^2\}$ as the true value of $\boldsymbol{\gamma}_{\mathcal{K}}$. Define functions $\psi(t) = G''(t)/G'(t)$ and $\eta(t) = \psi'(t) = G'''(t)/G'(t) - \{G''(t)/G'(t)\}^2$ with $G'''(\cdot)$ being the third derivative of $G(\cdot)$. The Taylor's series expansion of $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ at $(\boldsymbol{\zeta}_0, \boldsymbol{\gamma}_{0\mathcal{K}})$ yields

$$\begin{aligned}
&U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) \\
&= n^{-1/2} \sum_{i=1}^{n} \big\{ \Delta_i + \Delta_i \psi_i(\boldsymbol{\zeta}_0) \xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0) \xi_i(\boldsymbol{\zeta}_0) \big\} \big\{ R_i S_i + (1 - R_i) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K}} \big\} \\
&\qquad + \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) + \widehat{\boldsymbol{I}}_{\beta\zeta}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\zeta\zeta}^{-1} \boldsymbol{U}_{\zeta,i} + o_p(1) \qquad (3.4)
\end{aligned}$$

under some regularity conditions, where the expressions of $\widehat{\boldsymbol{I}}_{\zeta\zeta}$, $\widehat{\boldsymbol{I}}_{\beta\zeta}$, $\widehat{\boldsymbol{I}}_{\gamma\gamma}$ and $\widehat{\boldsymbol{I}}_{\beta\gamma}$ are given in Section 3.6.1, $\boldsymbol{U}_{\zeta,i} = (\boldsymbol{U}_{\alpha,i}^{\mathrm{T}}, U_{\lambda_1,i}, \ldots, U_{\lambda_m,i})^{\mathrm{T}}$,

$$\boldsymbol{U}_{\alpha,i} = \big\{ \Delta_i + \Delta_i \psi_i(\boldsymbol{\zeta}_0) \xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0) \xi_i(\boldsymbol{\zeta}_0) \big\} \boldsymbol{X}_i,$$

and $U_{\lambda_k,i}$ $(k = 1, \ldots, m)$ is the derivative of the ith term of log-likelihood function with respect to the jump λ_k:

$$U_{\lambda_k,i} = \frac{I(Y_i = t_k)\Delta_k}{\lambda_{0,k}} + I(Y_i \geq t_k) \big\{ \Delta_i \psi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0) \big\} \exp(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}_i).$$

Based on this expansion, we can estimate the asymptotic variance of $U_\beta(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ by

$\widehat{\sigma}^2(\mathcal{K}) = n^{-1}\sum_{i=1}^n\{\widehat{\sigma}_i(\mathcal{K}) - \overline{\sigma}(\mathcal{K})\}^2$, where

$$\widehat{\sigma}_i(\mathcal{K}) = \big\{\Delta_i + \Delta_i\psi_i(\widehat{\boldsymbol{\zeta}})\xi_i(\widehat{\boldsymbol{\zeta}}) - G_i'(\widehat{\boldsymbol{\zeta}})\xi_i(\widehat{\boldsymbol{\zeta}})\big\}\big\{R_iS_i + (1-R_i)\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}$$
$$+ \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}R_i(S_i - \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}) + \widehat{\boldsymbol{I}}_{\beta\zeta}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\zeta\zeta}^{-1}\widehat{\boldsymbol{U}}_{\zeta,i},$$

$\overline{\sigma}(\mathcal{K}) = n^{-1}\sum_{i=1}^n\widehat{\sigma}_i(\mathcal{K})$, and $\widehat{\boldsymbol{U}}_{\zeta,i}$ is $\boldsymbol{U}_{\zeta,i}$ with true parameters replaced by estimators. Note that the true parameters in $\widehat{\boldsymbol{I}}_{\zeta\zeta}$, $\widehat{\boldsymbol{I}}_{\beta\zeta}$, $\widehat{\boldsymbol{I}}_{\gamma\gamma}$ and $\widehat{\boldsymbol{I}}_{\beta\gamma}$ are replaced by estimators in the definition of $\widehat{\sigma}_i(\mathcal{K})$. For an asymptotic size α test, we reject H_0 if $U_\beta^2(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})/\widehat{\sigma}^2(\mathcal{K}) \geq \chi_{1,\alpha}^2$.

The proposed test is robust in the sense that it preserves the type I error under the null hypothesis without requirement of fully correct specifications of the model of S and the model of Y, because the proposed empirical variance estimator is derived using sum of squares, instead of a model-based estimator. By contrast, full-likelihood based methods that rely on the second derivative of log-likelihood function to estimate the variance of score statistic need extra assumptions to assure the test validity. For example, misspecification of either the outcome model or the missing covariate model can affect the type I error of the likelihood-based score test (Lawless, 2018).

3.1.2 Supremum Test

The above proposed score test is based on knowledge of the true transformation function G. In practice, however, the function may not be known, and misspecification of G can result in power loss. We propose a supremum test that combines the results from multiple choices of G to improve power. Let $\{G^{(j)}, j = 1, \ldots, q\}$ be a set of monotonically increasing transformation functions with $G^{(j)}(0) = 0$. For a particular choice of $G = G^{(j)}$, the density function of T given $(\boldsymbol{X}, S)$ is in the form of

$$f^{(j)}(t \mid \boldsymbol{X}, S; \boldsymbol{\alpha}, \beta, \Lambda)$$
$$= \exp\big[-G^{(j)}\{\Lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)\}\big]G^{(j)\prime}\{\Lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)\}\lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S),$$

where $\lambda(t) = \mathrm{d}\Lambda(t)/\mathrm{d}t$. Define $\boldsymbol{\alpha}_0^{(j)}$, $\beta_0^{(j)}$ and $\Lambda_0^{(j)}$ to be the values that solve the following equations simultaneously with probability one:

$$\mathrm{E}\Big\{\frac{\partial \log f^{(j)}(T \mid \boldsymbol{X}, S; \boldsymbol{\alpha}, \beta, \Lambda)}{\partial(\boldsymbol{\alpha}, \beta)} \mid \boldsymbol{X}, S\Big\} = \mathbf{0},$$
$$\mathrm{E}\Big[\frac{\partial \log f^{(j)}\{T \mid \boldsymbol{X}, S; \boldsymbol{\alpha}, \beta, \Lambda + \epsilon \int h(s)\,\mathrm{d}\Lambda(s)\}}{\partial \epsilon}\Big|_{\epsilon=0} \mid \boldsymbol{X}, S\Big] = 0 \text{ for } \|h\|_V \leq M,$$

where $\|h\|_V$ is the total variation of $h(t)$ in $[0, \tau]$, and M is some positive constant. In the supremum test, we extend the null hypothesis to $H_0' : \beta_0^{(j)} = 0$ for $j = 1, \ldots, q$. One sufficient condition for H_0' to hold is when S is independent of Y given $\boldsymbol{X}$. Let $\widehat{\boldsymbol{\zeta}}^{(j)}$ denote the nonparametric maximum likelihood estimator under $\beta^{(j)} = 0$ and transformation function $G^{(j)}$. Let $U_\beta^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ and $\widehat{\sigma}^{(j)}(\mathcal{K})$ denote the corresponding score statistic and estimated standard deviation, respectively. Define

$$Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = \max_{1 \leq j \leq q} |Z^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})|, \tag{3.5}$$

where $Z^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) = U_\beta^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})/\widehat{\sigma}^{(j)}(\mathcal{K})$. For notational convenience, we abbreviate $Z^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ and $U_\beta^{(j)}(\widehat{\boldsymbol{\zeta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ as $\widehat{Z}^{(j)}$ and $\widehat{U}_\beta^{(j)}$, respectively. The advantage of using the supremum test statistic $Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ lies in that it takes the uncertainty of the link between the outcome Y and the covariates under consideration. Intuitively, the supremum test statistic reflects how large a test statistic can be across several outcome models under the null hypothesis H_0'. The test statistic would tend to be away from zero if $\beta^{(j)}$ under at least one of the transformation functions is nonzero.

We approximate the distribution of $\big(\widehat{Z}^{(1)}, \ldots, \widehat{Z}^{(q)}\big)$ by a multivariate normal with mean $\mathbf{0}$ and variance $\widehat{\boldsymbol{V}}(\mathcal{K})$ and then approximate the distribution of $Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ by the absolute maximum of the multivariate normal random vector. The variance matrix $\widehat{\boldsymbol{V}}(\mathcal{K})$

is a $(q \times q)$-matrix with diagonal elements 1 and the (j,k)th element in the form of

$$\widehat{V}_{jk}(\mathcal{K}) = \frac{1}{n\widehat{\sigma}^{(j)}(\mathcal{K})\widehat{\sigma}^{(k)}(\mathcal{K})} \sum_{i=1}^{n} \Big(\widehat{U}_{\beta,i}^{(j)} - \frac{1}{n}\sum_{i'=1}^{n} \widehat{U}_{\beta,i'}^{(j)}\Big)\Big(\widehat{U}_{\beta,i}^{(k)} - \frac{1}{n}\sum_{i'=1}^{n} \widehat{U}_{\beta,i'}^{(k)}\Big), \quad j,k = 1,\ldots,q,$$

where $\widehat{U}_{\beta,i}^{(j)}$ is the ith term in the summation of $\widehat{U}_{\beta}^{(j)}$ for $i = 1,\ldots,n$ and $j = 1,\ldots,q$.

To obtain an asymptotic size α test, we use the Monte Carlo method to construct the empirical critical value of the test. The algorithm is as follows:

1. Generate M i.i.d. random samples $(z_m^{(1)},\ldots,z_m^{(q)}), m = 1,\ldots,M$ from a multivariate normal distribution with mean $\mathbf{0}$ and variance $\widehat{\boldsymbol{V}}(\mathcal{K})$;

2. Compute the test statistic T_m from the mth sample: $T_m = \max_{1\le j\le q} |z_m^{(j)}|$;

3. Reject H_0' if $Z^{\max}(\widehat{\boldsymbol{\zeta}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ is larger than the $(1-\alpha)$th quantile of $(T_1,\ldots,T_M)$.

3.2 Asymptotic Theory

In this section, we consider the asymptotic property of the score statistic under multiple choices of the transformation function G, with the single G as a special case. Let $\boldsymbol{I}_{\zeta\zeta}$, $\boldsymbol{I}_{\beta\zeta}$, $\boldsymbol{I}_{\gamma\gamma}$ and $\boldsymbol{I}_{\beta\gamma}$ denote the expectations of $\widehat{\boldsymbol{I}}_{\zeta\zeta}$, $\widehat{\boldsymbol{I}}_{\beta\zeta}$, $\widehat{\boldsymbol{I}}_{\gamma\gamma}$ and $\widehat{\boldsymbol{I}}_{\beta\gamma}$, respectively. Suppose that the transformation functions $G^{(j)}, j = 1,\ldots,q$ are continuously differentiable up to the forth order. This holds for both the proportional hazards model $G(x) = x$ and the proportional odds model $G(x) = \log(1+x)$. Define

$$\begin{aligned} U_{\beta}^{(j)}(\mathcal{K}) = &\big\{\Delta + \Delta\psi^{(j)}(\boldsymbol{\zeta}_0^{(j)})\xi(\boldsymbol{\zeta}_0^{(j)}) - G^{(j)'}(\boldsymbol{\zeta}_0^{(j)})\xi(\boldsymbol{\zeta}_0^{(j)})\big\} \\ &\times \big\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\} + (\boldsymbol{I}_{\beta\gamma}^{(j)})^{\mathrm{T}}(\boldsymbol{I}_{\gamma\gamma}^{(j)})^{-1}\boldsymbol{W}_{\mathcal{K}}R(S - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}) \\ &+ (\boldsymbol{I}_{\beta\zeta}^{(j)})^{\mathrm{T}}(\boldsymbol{I}_{\zeta\zeta}^{(j)})^{-1}\boldsymbol{U}_{\zeta}^{(j)} \end{aligned}$$

for $j = 1, \ldots, q$, where $\psi^{(j)}, \boldsymbol{I}^{(j)}_{\beta\gamma}, \boldsymbol{I}^{(j)}_{\gamma\gamma}, \boldsymbol{I}^{(j)}_{\beta\zeta}, \boldsymbol{I}^{(j)}_{\zeta\zeta}$ and $\boldsymbol{U}^{(j)}_{\zeta}$ are $\psi, \boldsymbol{I}_{\beta\gamma}, \boldsymbol{I}_{\gamma\gamma}, \boldsymbol{I}_{\beta\zeta}, \boldsymbol{I}_{\zeta\zeta}$ and $\boldsymbol{U}_{\zeta}$, respectively, under $G^{(j)}$. Let $\widehat{\boldsymbol{U}}_{\beta}(\mathcal{K}) = \left(\widehat{U}^{(1)}_{\beta}, \ldots, \widehat{U}^{(q)}_{\beta}\right)^{\mathrm{T}}$; note that the right-hand side implicitly depends on $\mathcal{K}$.

Here, we show that under some regularity conditions, $\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})\widehat{\boldsymbol{U}}_{\beta}(\mathcal{K})$ converges to a multivariate normal distribution under H_0' even when $\mathcal{K}$ is chosen randomly, where $\boldsymbol{\Sigma}(\mathcal{K})$ is defined in the proof of Theorem 3.1. To precisely state the theoretical result, let $\mathcal{K}^*$ be a general model selection operator, such that for an m-vector of outcome variables $\mathcal{Y}$ and an $(m \times p)$-matrix of covariates $\mathcal{Z}$, $\mathcal{K}^*(\mathcal{Y}, \mathcal{Z}) : \mathbb{R}^m \times \mathbb{R}^{m \times p} \to \mathcal{C}_p$, where $\mathcal{C}_p$ is the collection of subsets of $\{1, \ldots, p\}$. Suppose that the model for S is selected based on the residual $S - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{A}$, where $\widehat{\boldsymbol{\gamma}}_X \equiv (\sum_{i=1}^n R_i \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}})^{-1} \sum_{i=1}^n R_i \boldsymbol{X}_i S_i$ is the least-squares estimator of S on $\boldsymbol{X}$ using the subjects with $R = 1$. The selected components of $\boldsymbol{A}$ are $\mathcal{K}^*(\mathcal{S} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}_X, \boldsymbol{A})$, where $\mathcal{S}$ is a vector that consists of $\{S_i : R_i = 1\}$, and $\boldsymbol{X}$ and $\boldsymbol{A}$ are matrices that consist of rows of $\{\boldsymbol{X}_i : R_i = 1\}$ and $\{\boldsymbol{A}_i : R_i = 1\}$, respectively. For simplicity of presentation, we write $\mathcal{K}^* = \mathcal{K}^*(\mathcal{S} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}_X, \boldsymbol{A})$. Therefore, $\mathcal{K}$ can be viewed as the observed value of $\mathcal{K}^*$.

Let $\|\cdot\|_{\psi_\xi}$ be an Orlicz norm, such that $\|X\|_{\psi_\xi} = \inf\{\eta > 0 : \mathrm{E}(e^{|X|^\xi/\eta^\xi}) \leq 2\}$, and $\|\cdot\|$ be the Euclidean norm. Define the set

$$\begin{aligned}\mathcal{H} = \{ &(\boldsymbol{h}_\alpha, h_\Lambda) : \boldsymbol{h}_\alpha \in \mathbb{R}^{|\boldsymbol{X}|}, h_\Lambda \text{ is a function with bounded variation on } [0, \tau]; \\ &\|\boldsymbol{h}_\alpha\| \leq 1, \|h_\Lambda\|_V \leq 1\},\end{aligned}$$

where $|\boldsymbol{X}|$ is the dimension of covariate $\boldsymbol{X}$. Also, we define a neighborhood of the true parameter $(\boldsymbol{\alpha}_0, \Lambda_0)$, denoted by $\mathbb{L}$, as

$$\mathbb{L} = \left\{(\boldsymbol{\alpha}, \Lambda) : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| + \sup_{t \in [0,\tau]} |\Lambda(t) - \Lambda_0(t)| < \epsilon_0\right\}$$

for a very small constant ϵ_0. We establish the asymptotic property of $\widehat{\boldsymbol{U}}_{\beta}(\mathcal{K}^*)$ under the

following conditions. Some conditions involve a generic positive constant M.

(C1) For some $\xi \in (0, 2]$, $\|S\|_{\psi_\xi} + \max_j \|A_j\|_{\psi_\xi} < M$. The covariate $\boldsymbol{X}$ is bounded, so that $P(\|\boldsymbol{X}\| < M) = 1$.

(C2) There exists a sequence of collections of models Ω_n, such that $P(\mathcal{K}^* \in \Omega_n) \to 1$, $\sup_{\mathcal{K}\in\Omega_n} |\mathcal{K}| = O(n^\nu)$, and $\log|\Omega_n| = O(n^\kappa)$, where ν and κ are constants that satisfy $\nu < 3\xi/(4\xi+8)$, $\kappa < 1/2$, and $4\nu/3+8\kappa/(3\xi) < 1$, and $|\mathcal{C}|$ denotes the cardinality of the set $\mathcal{C}$. Also, $\inf_{\mathcal{K}\in\Omega_n} \lambda_{\min}\{\mathrm{E}(R\boldsymbol{W}_\mathcal{K}\boldsymbol{W}_\mathcal{K}^{\mathrm{T}})\} > M^{-1}$, $\sup_{\mathcal{K}\in\Omega_n} \mathrm{E}\{(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_\mathcal{K})^4\} < M$, where $\lambda_{\min}(\boldsymbol{C})$ denotes the minimum eigenvalue of the matrix $\boldsymbol{C}$. In addition, $\inf_{\mathcal{K}\in\Omega_n} \lambda_{\min}\{\boldsymbol{\Sigma}(\mathcal{K})\} > M^{-1}$.

(C3) The probability $P(R = 1 \mid Y, \boldsymbol{X}) > M^{-1}$ almost surely.

(C4) Under $\beta = 0$, the residual $(S - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X})$ and $\boldsymbol{X}$ are independent, and $\boldsymbol{A}$ is independent of $(Y, \boldsymbol{X})$.

(C5) The models selected based on the estimated residuals $(S_i - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ and the actual residuals $(S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i)_{i:R_i=1}$ are such that

$$P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) \neq \mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A})\Big\} = o(1)$$

and

$$\sup_{\mathcal{K}\in\Omega_n} \frac{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\widehat{\boldsymbol{\gamma}}_X, \mathcal{A}) = \mathcal{K}\Big\}}{P\Big\{\mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A}) = \mathcal{K}\Big\}} < M.$$

(C6) For a random sample of size m, let $\widetilde{\mathcal{S}} = (S_1, \ldots, S_m)^{\mathrm{T}}$, $\widetilde{\mathcal{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)^{\mathrm{T}}$, and $\widetilde{\mathcal{A}} = (\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)^{\mathrm{T}}$. The random variable

$$\sup_{\mathcal{K}\in\Omega_m} \left| \frac{P\{\mathcal{K}^*(\widetilde{\mathcal{S}} - \widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X}, \widetilde{\mathcal{A}}) = \mathcal{K} \mid \widetilde{\mathcal{A}}\}}{P\{\mathcal{K}^*(\widetilde{\mathcal{S}} - \widetilde{\mathcal{X}}\boldsymbol{\gamma}_{0X}, \widetilde{\mathcal{A}}) = \mathcal{K}\}} - 1 \right|$$

converges to 0 in mean as $m \to \infty$.

(C7) The parameter value $\boldsymbol{\alpha}_0^{(j)}$ $(j = 1, \ldots, q)$ lies in the interior of a known compact set, and $\Lambda_0^{(j)}$ $(j = 1, \ldots, q)$ is continuously differentiable with positive derivatives in $[0, \tau]$. Also, with probability one, $P(C \geq \tau \mid \boldsymbol{X}, S) > M^{-1}$ and $P(T \geq \tau \mid \boldsymbol{X}, S) > M^{-1}$.

(C8) The class of functions $\{U^{(j)}(\boldsymbol{\alpha}, \Lambda)[\boldsymbol{h}_\alpha, h_\Lambda] : (\boldsymbol{\alpha}, \Lambda) \in \mathbb{L}, (\boldsymbol{h}_\alpha, h_\Lambda) \in \mathcal{H}\}$, with $U^{(j)}(\boldsymbol{\alpha}, \Lambda)[\boldsymbol{h}_\alpha, h_\Lambda]$ defined in Section 3.6.2, is Donsker for $j = 1, \ldots, q$. Also, the operator $(\boldsymbol{W}_\alpha^{(j)}(\boldsymbol{h}_\alpha, h_\Lambda), W_\Lambda^{(j)}(\boldsymbol{h}_\alpha, h_\Lambda))$ defined in Section 3.6.2 is invertible for $j = 1, \ldots, q$.

Remark 3.1. Conditions (C1)–(C6) are adopted from Chapter 2. Condition (C7) is standard for semiparametric survival models. Condition (C8) essentially consists of assumptions for the Z-estimator master theorem (Theorem 3.3.1 of van der Vaart and Wellner (1996)), which guarantees the asymptotic normality of $(\widehat{\boldsymbol{\alpha}}^{(j)}, \widehat{\Lambda}^{(j)})$ for $j = 1, \ldots, q$. We directly assume the required conditions instead of proving the conditions based on properties of the true model, because we do not assume the form of the true model. If we assume that one of the transformation model is true, then we may prove the desired results from conditions on the true model along the lines of, for example, Zeng et al. (2008).

The following theorem establishes the asymptotic normality of the score statistic under a random model selection event.

Theorem 3.1. *Under conditions (C1)–(C8) and H_0', $\boldsymbol{\Sigma}^{-1/2}(\mathcal{K}^*)\widehat{\boldsymbol{U}}_\beta(\mathcal{K}^*)$ converges weakly to the standard multivariate normal distribution.*

3.3 Simulation Studies

Let $\boldsymbol{X} = (X_1, \ldots, X_5)^{\mathrm{T}}$, where (X_1, X_2, X_3) are mean-zero multivariate normal variables with $\mathrm{Cov}(X_j, X_k) = 0.5^{|j-k|}$ $(j, k = 1, 2, 3)$, $X_4 \sim$ Bernoulli(0.25), $X_5 \sim$ Bernoulli(0.35), and X_4 and X_5 are independent of each other and (X_1, X_2, X_3). Let $\boldsymbol{A}$ be a p-vector of independent standard normal random variables. We set $S = \boldsymbol{\gamma}_X^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\gamma}_A^{\mathrm{T}}\boldsymbol{A} + \delta$, where δ is standard normal, and $\boldsymbol{\gamma}_X = (0.1, \ldots, 0.1)^{\mathrm{T}}$. We set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components and 0 at the remaining components.

We considered three failure time models:

Model 1: $\Lambda(t \mid \boldsymbol{X}, S) = \Lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)$;

Model 2: $\Lambda(t \mid \boldsymbol{X}, S) = \log\{1 + \Lambda(t)\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} + \beta S)\}$;

Model 3: $T = \exp(-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X} - \beta S) + \epsilon$, where $\epsilon \sim \mathrm{Exp}(1)$.

We set $\boldsymbol{\alpha} = (0.2, -0.2, 0.2, -0.2, 0.2)^{\mathrm{T}}$ and $\Lambda(t) = 0.01t$. Model 1 and Model 2 are the proportional hazards model and proportional odds model, respectively. The censoring time C was generated from an exponential distribution with the mean chosen to yield a censoring rate of about $50\% - 60\%$. We considered two missing-data mechanisms. The first mechanism is missing completely at random (MCAR), where the missing-data status is independent of other variables. The second mechanism is missing at random (MAR), where we first randomly select 20% of the subjects into a subgroup, who will have observed S. For subjects outside the subgroup, we select a fraction of subjects with censored event time to have missing S to attain the desired missing proportion. If the missing proportion is not attained by setting all censored subjects to have missing S, then a subset of subjects with observed event time will also be selected. We considered sample sizes of $n = 500$ and 1000, and numbers of auxiliary variables of $p = 200$, 500, 1000 and 1500. For the alternative hypothesis, we set $\beta = 3n^{-1/2}$ for Models 1 and 2, and $\beta = 1.5n^{-1/2}$ for Model

3. For each setting, we simulated 50,000 and 10,000 replicates for $\beta = 0$ and $\beta \neq 0$, respectively.

In the first study, we consider the performance of the proposed test under a given transformation model. We compare the performance of six tests: (1) the standard score test using complete data only; (2) the standard score test with missing values imputed under a working linear model of S on $\boldsymbol{X}$ and components of $\boldsymbol{A}$ selected using marginal screening, where a component of $\boldsymbol{A}$ is selected if its absolute empirical correlation with $S - \widehat{\boldsymbol{\gamma}}_X^{\mathrm{T}} \boldsymbol{X}$ among the subjects with complete data is larger than a certain threshold; (3) Lawless (2018)'s score test based on the full likelihood with a working linear model of S against $\boldsymbol{X}$ only; (4) Lawless (2018)'s score test based on the full likelihood with the same model of S as (2); (5) the proposed test, where the working model of S is selected in the same way as (2); and (6) the score test based on the full likelihood with a linear model of S against $\boldsymbol{X}$ and the components of $\boldsymbol{A}$ that are associated with S. We refer to methods (1)–(6) as the complete-case analysis, the simple imputation method, the covariate-only analysis, the Lawless method, the proposed method, and the true model method. For methods (2), (4) and (5), the threshold for screening is selected using BIC. For the true model method, the variance of the score statistic is estimated using the proposed empirical variance estimator. For all methods, we fit the correct failure time model under Models 1 and 2, and under Model 3, we fit both the proportional hazards and proportional odds models.

The results under a missing proportion of 60% are plotted in Figures 3.1 and 3.2, and the results under a missing proportion of 30% are presented in Section 3.6.4; for methods that inflate the type I error, their performance under the alternative hypothesis is not presented. In the figures, we use PH and PO to represent the proportional hazards model and the proportional odds model, respectively. The significance level is set to be 0.05. Under Models 1 and 2 with sample size 1000, all methods preserve the type I error.

Theoretically, the complete-case analysis and the simple imputation method would inflate the type I error under MAR, but the empirical results do not exhibit such a pattern under the current setting. Under Model 3, the complete-case analysis, the simple imputation method, the covariate-only analysis, and the Lawless method generally inflate the type I error, because these methods estimate the variance based on the second derivative of the log-likelihood, which is misspecified in this setting. Under the alternative hypothesis, the Lawless method and the simple imputation method under Models 1 and 2 have relatively high power since both methods underestimate the variance of test statistic. As expected, the proposed method utilizes information about missing data contained in the auxiliary variables tends to yield higher power than the covariate-only method.

In the second study, we investigate the supremum test under the same simulation settings. The supremum test is performed with $q = 2$, $G^{(1)}(x) = x$, and $G^{(2)}(x) = \log(1 + x)$. For comparison, we also present the results of the proposed single-model score test with the proportional hazards model and the proportional odds model. The results under a missing proportion of 60% are plotted in Figures 3.3 and 3.4, and the results under a missing proportion of 30% are presented in Section 3.6.4. All three tests preserve the type I error under all three models. Under Models 1 and 2, the supremum test does not lose much power compared with the single-model test with the correct model specification. Under Model 3, the power of the supremum test is substantially larger than that of the single-model test with the proportional hazards model. The power of the single-model test with the proportional odds model is slightly larger than that of the supremum test, probably because Model 3 is close to a proportional odds model. However, the supremum test is never more than 15% less efficient than the single-model tests, whereas the single-model tests may have substantial efficiency loss due to model misspecification. This illustrates that even when the outcome model is unknown or misspecified, we can perform the supremum test to achieve a relatively high power.

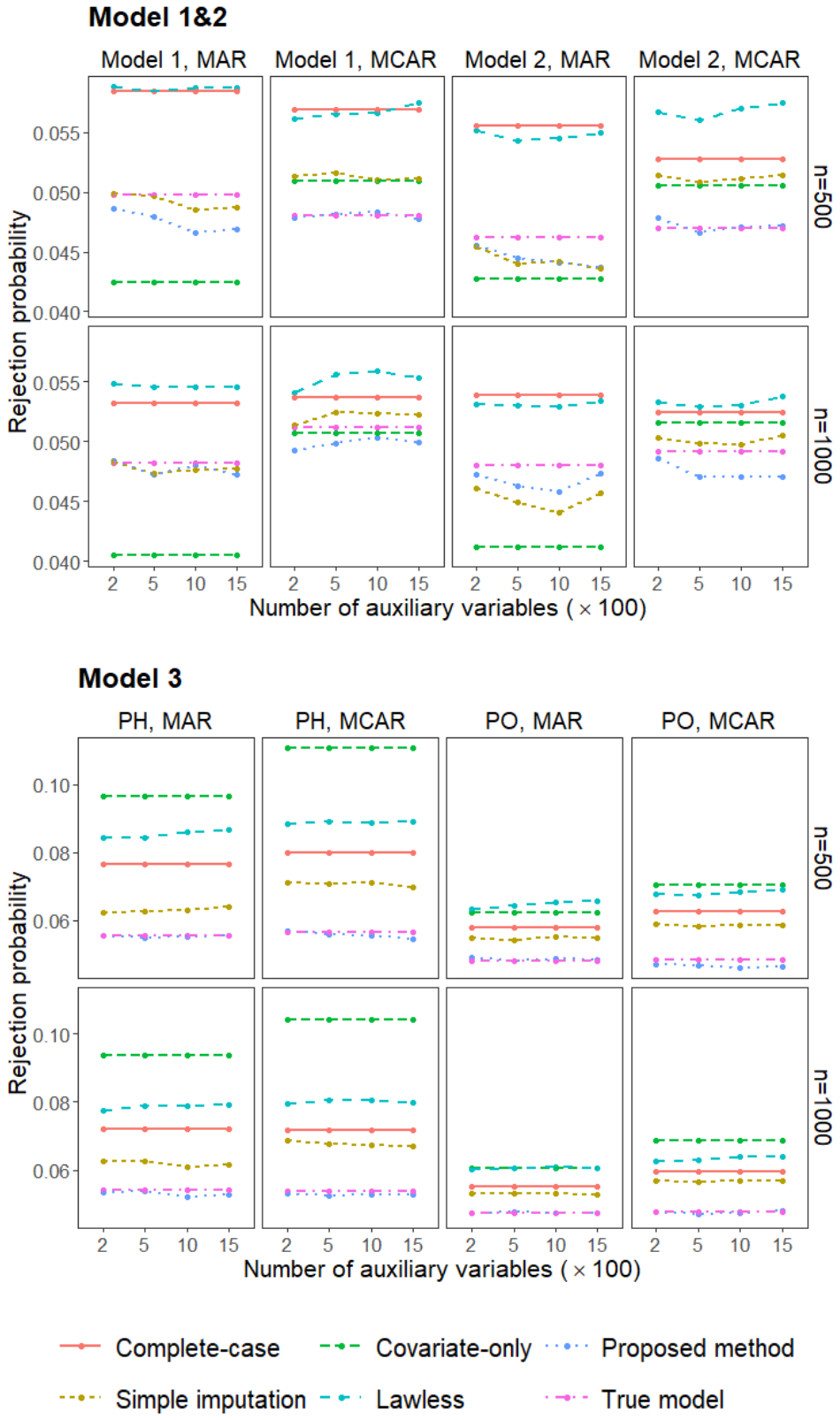


Figure 3.1: Study 1 - Rejection probabilities under a missing proportion of 60% and the null hypothesis.

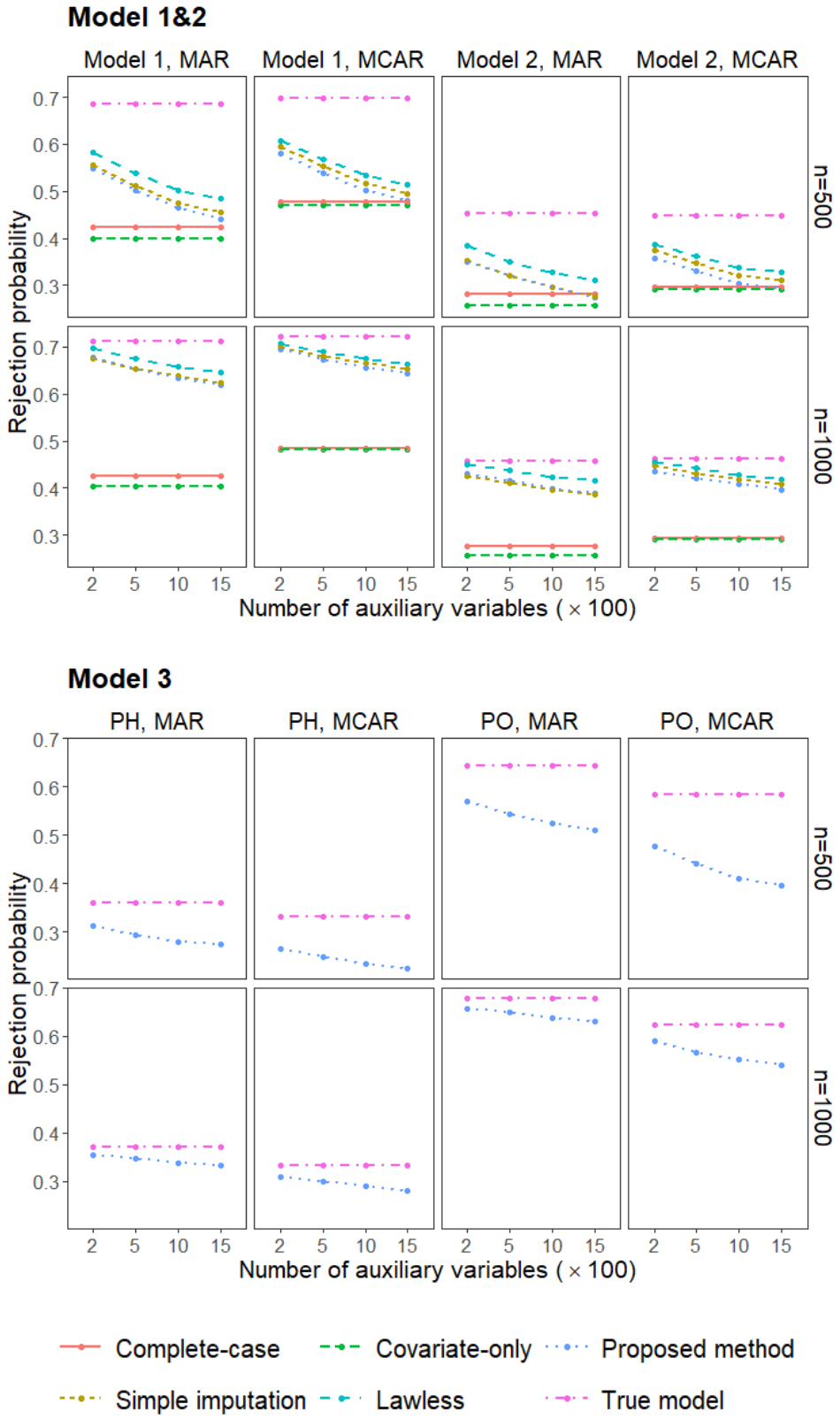


Figure 3.2: Study 1 - Rejection probabilities under a missing proportion of 60% and the alternative hypothesis.

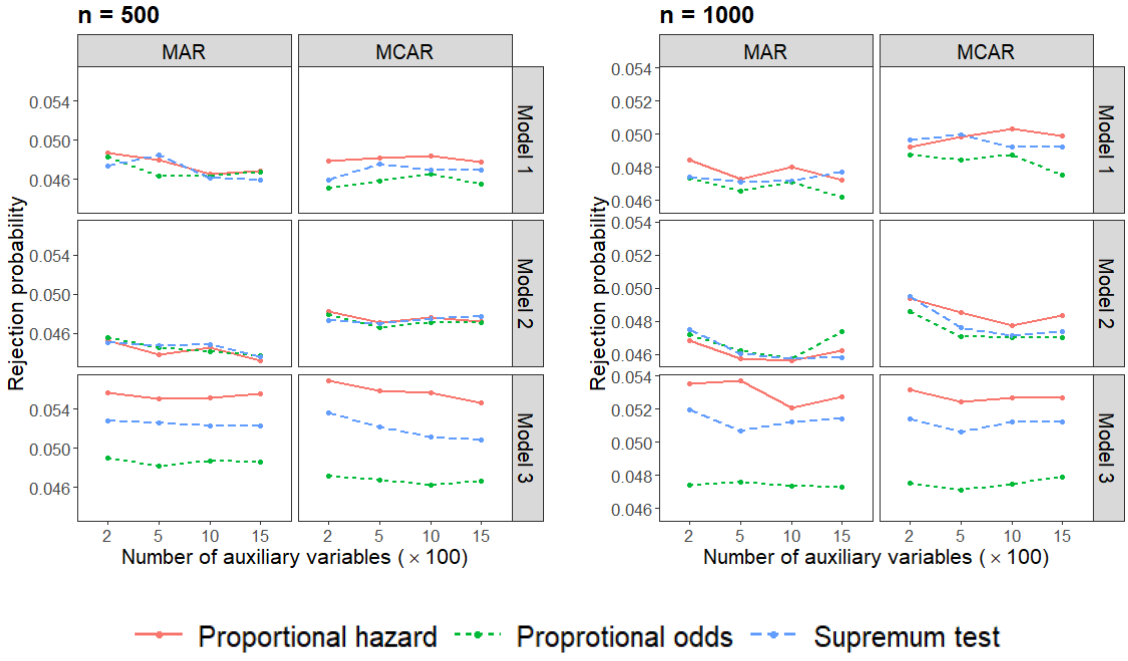


Figure 3.3: Study 2 - Rejection probabilities under a missing proportion of 60% and the null hypothesis.

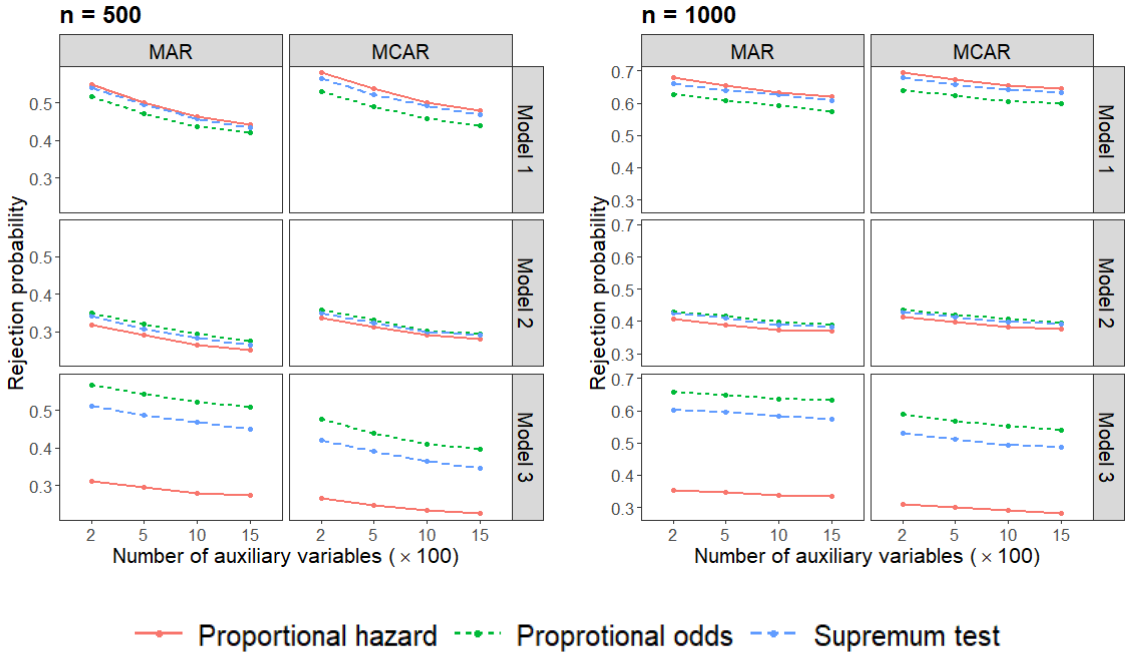


Figure 3.4: Study 2 - Rejection probabilities under a missing proportion of 60% and the alternative hypothesis.

We conducted an additional simulation study with a mixture of strong and weak signals of the auxiliary variables. We set $\boldsymbol{\gamma}_A$ to be 0.25 at the first 20 components, 0.02 at the subsequent 80 components, and 0 at the remaining components. The results are summarized in Section 3.6.4. By results not presented, similar to the results of study 1, the proposed method presents good performance under this setting, too. Under the alternative hypothesis, the true model method tends to have high power. Nevertheless, the proposed method is more powerful than the true model method under some scenarios. This is because the true model contains many auxiliary variables with weak signals, and the extra information contained in the variables does not compensate the variability involved in the estimation of their effects.

3.4 Real Data Analysis

3.4.1 TCGA: Bladder Urothelial Carcinoma

We analyze a dataset of patients with bladder urothelial carcinoma (BLCA) from TCGA (The Cancer Genome Atlas Network, 2014). In the study, most subjects had available clinical variables, including sex, age at diagnosis, time to tumor progression and time to death since initial diagnosis. The expressions of 18224 genes, generated by RNA sequencing, are measured for most subjects. The expressions of 208 proteins or phosphoproteins are available for 82% of the subjects. After removing subjects with missing clinical data, the sample size is 348. The median follow-up time was about 1.3 years, and about 49% of the patients were lost to follow-up before tumor progression or death.

We aim to identity protein expressions that are associated with the time to tumor progression or death, whichever occurs first. The covariates in $\boldsymbol{X}$ include age at diagnosis, sex and stage N. In the sample, 26.44% patients are female. Stage N is classified into N0 (64.08%), N1(12.93%), N2(21.26%) and N3(1.72%) and is represented by a single variable

Table 3.1: Rejection probabilities and references of significant proteins in the TCGA bladder urothelial carcinoma analysis.

Protein expression	**Proposed method**	**Complete-case**		**Covariate-only**		**Reference**
		PH	PO	PH	PO	
`GATA3`	3.40E−05	1.12E−04	5.00E−05	1.04E−04	4.40E−05	Higgins et al. (2007)
`Src`	1.48E−04	8.20E−04	4.86E−04	8.33E−04	4.62E−04	Xu et al. (2021)
`TAZ`	1.80E−04	1.38E−03	5.51E−04	1.09E−03	4.32E−04	Gao et al. (2014)

with values 0, 1, 2, and 3, respectively. In a single analysis, we set the covariate of interest S to be the expression of a protein or phospho-protein. We set the gene expressions as auxiliary variables. About 6% of the gene expression values are missing, and we impute them using k-nearest neighbor imputation with $k = 10$.

We perform the supremum test with $q = 2$ and the two transformation functions corresponding to the proportional hazards and proportional odds models. The working model of S is selected in two steps: first, we select 1000 gene expressions by the correlation-based marginal screening procedure, and then we perform lasso on the selected gene expressions; the tuning parameter in lasso is selected by BIC. For comparison, we also perform the complete-case analysis and the covariate-only method described in the simulation studies under the proportional hazards and proportional odds models.

Under a (family-wise) significance level of 0.05 and the Bonferroni correction, i.e., an individual significance level of $0.05/208 = 0.00024$, 3 proteins are identified to be significantly associated with progression-free survival time under at least one of the five tests. All of the 3 protein expressions are more significant under the proposed method than under other methods with either outcome model. Also, the 3 protein expressions have been identified to be related to the progression of bladder urothelial carcinoma in previous studies. The p-values under all methods of the significant protein expressions and some relevant references are given in Table 3.1.

3.4.2 METABRIC

We also apply the proposed method to analyze the data from the METABRIC study (Curtis et al., 2012) to investigate the association between gene expressions and the time to tumor progression or death of breast cancer patients. The data are available through the cBioPortal for Cancer Genomics (https://www.cbioportal.org/study/summary?id=brca_metabric). The study contains data of clinical variables, gene expressions and copy number alterations (CNAs). For the analysis, we select patients with subtypes Luminal A and Luminal B as study subjects. Also, we select the 1500 genes with largest variances as the study variables. After removing subjects with missing clinical data, the sample size is 1119. The median follow-up time was about 119 months, and 35% of the patients were lost to follow-up before tumor progression or death. We artificially introduce 50% of missingness with the MAR mechanism described in the simulation studies for the gene expressions to demonstrate the proposed method.

The covariates in $\boldsymbol{X}$ include age at diagnosis, Her2 status, indicator of chemotherapy, indicator of hormone therapy, and indicator of radiotherapy. Her2 status is classified into loss (6.08%), neutral (77.57%) and gain (16.35%) and is represented by a single variable with values 0, 1 and 2. In a single analysis, we set the covariate of interest S to be a single gene expression. We set the CNAs as auxiliary variables. For each CNA, if there exists another CNA such that they have more than 95% same values, then we delete it from the analysis. After deletion, the dimension of CNA is 385.

We perform the supremum test with $q = 2$ and the two transformation functions corresponding to the proportional hazards and proportional odds models. The working model of S is selected by lasso, and the tuning parameter is selected using BIC. For comparison, we include the results under the complete-case analysis and the covariate-only method described in the simulation studies with the proportional hazards and proportional odds models. Also, we perform score test using all available gene expressions under the pro-

Table 3.2: Rejection probabilities of significant gene expressions in the METABRIC data analysis.

Gene expression	**Proposed method**	**Complete-data**		**Complete-case**		**Covariate-only**	
		PH	PO	PH	PO	PH	PO
CDCA5	2.92E−03	7.60E−06	7.57E−08	3.30E−02	1.24E−02	1.51E−02	1.37E−02
FAM164A	4.44E−03	2.30E−05	2.16E−05	1.44E−02	3.68E−02	3.17E−02	3.74E−02
S100P	4.39E−03	4.47E−05	3.87E−06	3.15E−03	6.02E−03	9.66E−03	9.31E−03
NFKBIZ	6.50E−03	2.73E−04	1.99E−05	1.35E−02	7.01E−03	3.08E−02	9.39E−03
PTTG1	4.08E−03	7.41E−05	2.22E−05	6.85E−02	1.73E−02	6.33E−02	2.06E−02
CCNB2	9.80E−05	1.50E−04	9.71E−06	8.40E−04	1.38E−04	1.07E−03	2.06E−04
AURKA	1.24E−02	7.94E−04	2.04E−05	1.07E−01	3.56E−03	5.85E−02	1.10E−02

portional hazards and proportional odds models, and we refer it as the complete-data analysis. The results of complete-data analysis can be viewed as the gold standard since it contains no missing values of S and thus no variability caused by the imputation process is introduced.

There are 7 gene expressions identified to be significantly associated with progression-free survival time at the (Bonferroni-corrected) significance level of $0.05/1500 = 3.33 \times 10^{-5}$ under the complete-data analysis with either outcome model. Among these gene expressions, all of them are most significant under the complete-data analysis with the proportional odds model, 5 are more significant under the proposed method than under the complete-case analysis and the covariate-only method with either outcome model. This suggests that the proposed method is more powerful than the other two methods. The p-values under all methods of the significant gene expressions are given in Table 3.2.

3.5 Discussion

In this chapter, we develop a score test to detect the presence of association between a potentially right-censored survival outcome and an incomplete covariate, where the missing values of the incomplete covariate can be imputed using high-dimensional auxiliary

variables. We propose to select a subset of auxiliary variables before performing the score test. The transformation model of survival outcome is considered to relax distributional assumptions on the outcome variable. We propose a supremum test that considers multiple outcome models to improve power. We show that the proposed score statistic is asymptotically normal. Our theoretical development only requires that S is linearly associated with covariates $\boldsymbol{X}$, and the validity of the score test does not depend on the correctness of the working model.

In the proposed method, the model of Y can be misspecified without compromising the validity of the test. The proposed score test preserves the type I error with or without correct specification of the outcome model since the variance of score statistic is derived using sum of squares of the individual score statistics instead of the second derivative of the log-likelihood. As expected, when the outcome model is misspecified, the power is adversely affected. The loss in power is relatively small according to the second simulation study.

Our work can be extended in the following directions. First, the survival data considered in this chapter is right-censored. It is of interest to consider other types of censoring, such as interval censoring, where the event of interest is known only to occur within a time interval. For example, in HIV/AIDS studies, blood samples are taken from study subjects periodically to look for evidence of HIV sero-conversion. Then one subject's event time is only known to fall between two blood drawings. Second, in the current study, the time-to-event outcome is univariate. In genomic studies, we may encounter multivariate survival data, where each subject may experience more than one event. In that case, the interested events may be correlated with each other. We may consider modelling a joint survival function and performing a score test for multiple parameters.

3.6 Technical Details and Additional Results

3.6.1 The Derivative Terms in the Score Statistic

Define $\widehat{\boldsymbol{I}}_{\zeta\zeta} = \begin{bmatrix} \widehat{\boldsymbol{I}}_{\alpha\alpha} & \widehat{\boldsymbol{I}}_{\alpha\lambda} \\ \widehat{\boldsymbol{I}}_{\alpha\lambda}^{\mathrm{T}} & \widehat{\boldsymbol{I}}_{\lambda\lambda} \end{bmatrix}$, where

$$\widehat{\boldsymbol{I}}_{\alpha\alpha} = -\frac{1}{n}\sum_{i=1}^{n}\big\{\Delta\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 + \Delta_i\psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)\big\}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}},$$

$$\begin{aligned}(\widehat{\boldsymbol{I}}_{\alpha\lambda})_k = &-\frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t_k)\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) + \Delta_i\psi_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\big\} \\ &\times \exp(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\boldsymbol{X}_i, \quad k = 1, \ldots, m.\end{aligned}$$

Note that $\widehat{\boldsymbol{I}}_{\lambda\lambda}$ is a $(m \times m)$-matrix with the (j,k)th element as

$$(\widehat{\boldsymbol{I}}_{\lambda\lambda})_{jk} = \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}\Big[-\frac{\Delta_j}{\lambda_j^2} + I(Y_i \geq t_j)\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\big\}\exp(2\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\Big] & \text{if } k = j, \\ -\frac{1}{n}\sum_{i=1}^{n} I\{Y_i \geq \max(t_k, t_j)\}\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\big\}\exp(2\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i) & \text{if } k \neq j. \end{cases}$$

Define $\widehat{\boldsymbol{I}}_{\beta\zeta} = (\widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}}, \widehat{\boldsymbol{I}}_{\beta\lambda}^{\mathrm{T}})^{\mathrm{T}}$, where

$$\begin{aligned}\widehat{\boldsymbol{I}}_{\beta\alpha} = &\frac{1}{n}\sum_{i=1}^{n}\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 + \Delta_i\psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) \\ &- G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)^2 - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)\big\}\big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\big\}\boldsymbol{X}_i,\end{aligned}$$

$$\begin{aligned}(\widehat{\boldsymbol{I}}_{\beta\lambda})_k = &\frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq t_k)\big\{\Delta_i\eta_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) + \Delta_i\psi_i(\boldsymbol{\zeta}_0) - G_i''(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\big\}\exp(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i) \\ &\times \big\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}, \quad k = 1, \ldots, m.\end{aligned}$$

Also, we define

$$\widehat{\boldsymbol{I}}_{\gamma\gamma} = \frac{1}{n}\sum_{i=1}^{n} R_i \boldsymbol{W}_{\mathcal{K},i}\boldsymbol{W}_{\mathcal{K},i}^{\mathrm{T}},$$
$$\widehat{\boldsymbol{I}}_{\beta\gamma} = \frac{1}{n}\sum_{i=1}^{n} \big\{\Delta_i + \Delta_i\psi_i(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0) - G_i'(\boldsymbol{\zeta}_0)\xi_i(\boldsymbol{\zeta}_0)\big\}(1-R_i)\boldsymbol{W}_{\mathcal{K},i}.$$

3.6.2 Proof of Theorem 3.1

For simplicity, we suppress the transformation function index j in the following arguments. Let $G_i(\boldsymbol{\alpha},\Lambda) = G\{\xi_i(\boldsymbol{\alpha},\Lambda)\}$ with $\xi_i(\boldsymbol{\alpha},\Lambda) = \int_0^{Y_i}\exp(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i)\,\mathrm{d}\Lambda(s)$. The (scaled) score statistic for β can be written as

$$\begin{aligned} U_\beta(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda},\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) &= n^{-1/2}\sum_{i=1}^{n}\Big\{\Delta_i + \Delta_i\frac{G_i''(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda})}{G_i'(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda})}\xi_i(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda}) - G_i'(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda})\xi_i(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda})\Big\} \\ &\qquad \times \big\{R_iS_i + (1-R_i)\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\big\} \\ &\equiv n^{-1/2}\sum_{i=1}^{n}\mu_{1,i}(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda})\{R_iS_i + (1-R_i)\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\}. \end{aligned}$$

Define

$$I_{\beta\Lambda}(s) = \mathrm{E}\big[\mu_1'(\boldsymbol{\alpha}_0,\Lambda_0)\exp(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X})\big\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}I(s\le Y)\big],$$

where μ_1' is the first derivative of μ_1. Note that we can also write

$$\begin{aligned} \boldsymbol{I}_{\beta\alpha} &= \mathrm{E}\big[\mu_1'(\boldsymbol{\alpha}_0,\Lambda_0)\xi(\boldsymbol{\alpha}_0,\Lambda_0)\boldsymbol{X}\big\{RS + (1-R)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K}}\big\}\big], \\ \boldsymbol{I}_{\beta\gamma} &= \mathrm{E}\big\{\mu_1(\boldsymbol{\alpha}_0,\Lambda_0)(1-R)\boldsymbol{W}_{\mathcal{K}}\big\}. \end{aligned}$$

Let $\widehat{I}_{\beta\Lambda}(s)$ be the empirical counterpart of $I_{\beta\Lambda}(s)$, with the expectations replaced by empirical means. We suppress the argument s in the notation when there are no ambiguities. For notational convenience, we denote $\mu_1(\boldsymbol{\alpha}_0,\Lambda_0)$ and $\mu_1(\widehat{\boldsymbol{\alpha}},\widehat{\Lambda})$ by μ_1 and $\widehat{\mu}_1$, respectively.

To prove Theorem 3.1, we need to derive the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0, \widehat{\Lambda} - \Lambda_0)$. Let $\ell(\boldsymbol{\alpha}, \Lambda)$ be the log-likelihood for the survival model under H_0 for a generic subject, that is,

$$\ell(\boldsymbol{\alpha}, \Lambda) = \Delta\Big\{ \log \lambda(Y) + \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X} + \log G'(\boldsymbol{\alpha}, \Lambda) \Big\} - G(\boldsymbol{\alpha}, \Lambda).$$

We define the differentiation of $\ell(\boldsymbol{\alpha}, \Lambda)$ along $\big(\boldsymbol{h}_\alpha, \int h_\Lambda(s)\, \mathrm{d}\Lambda(s)\big)$ as a map from $\mathbb{L}$ to $\ell^\infty(\mathcal{H})$:

$$\begin{aligned} U(\boldsymbol{\alpha}, \Lambda)[\boldsymbol{h}_\alpha, h_\Lambda] &:= \frac{\mathrm{d}}{\mathrm{d}\epsilon} \ell\Big(\boldsymbol{\alpha} + \epsilon \boldsymbol{h}_\alpha, \Lambda + \epsilon \int h_\Lambda(s)\, \mathrm{d}\Lambda(s)\Big)\Big|_{\epsilon=0} \\ &= \Delta\{\boldsymbol{h}_\alpha^{\mathrm{T}} \boldsymbol{X} + h_\Lambda(Y)\} + \Delta\psi(\boldsymbol{\alpha}, \Lambda) \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}) \int \{\boldsymbol{h}_\alpha^{\mathrm{T}} \boldsymbol{X} + h_\Lambda(s)\}\, \mathrm{d}\Lambda(s) \\ &\quad - G'(\boldsymbol{\alpha}, \Lambda) \exp(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{X}) \int \{\boldsymbol{h}_\alpha^{\mathrm{T}} \boldsymbol{X} + h_\Lambda(s)\}\, \mathrm{d}\Lambda(s). \end{aligned}$$

Clearly, $\mathcal{P}\big\{U(\boldsymbol{\alpha}_0, \Lambda_0)[\boldsymbol{h}_\alpha, h_\Lambda]\big\} = 0$, where $\mathcal{P}\{g(x)\} = \mathrm{E}\{g(x)\}$ for any measurable function $g(x)$. By Taylor's series expansion, we have

$$\begin{aligned} &\mathcal{P}\big\{U(\widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})(\boldsymbol{h}_\alpha, h_\Lambda) - U(\boldsymbol{\alpha}_0, \Lambda_0)[\boldsymbol{h}_\alpha, h_\Lambda]\big\} \\ &= \mathcal{P}\Bigg((\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^{\mathrm{T}} \boldsymbol{X} \Bigg[\Delta\Big\{ \psi(\boldsymbol{\alpha}_0, \Lambda_0) \exp(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) + \eta(\boldsymbol{\alpha}_0, \Lambda_0) \exp(2\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \Lambda_0(Y) \Big\} \\ &\quad - \Big\{ G'(\boldsymbol{\alpha}_0, \Lambda_0) \exp(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) + G''(\boldsymbol{\alpha}_0, \Lambda_0) \exp(2\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \Lambda_0(Y) \Big\} \Bigg] \int_0^Y \{\boldsymbol{h}_\alpha^{\mathrm{T}} \boldsymbol{X} + h_\Lambda(s)\}\, \mathrm{d}\Lambda_0(s) \\ &\quad + \int_0^\tau I\{s \le Y\} \Bigg[\Big\{ \Delta\psi(\boldsymbol{\alpha}_0, \Lambda_0) - G'(\boldsymbol{\alpha}_0, \Lambda_0) \Big\} \exp(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \{\boldsymbol{h}_\alpha^{\mathrm{T}} \boldsymbol{X} + h_\Lambda(s)\} \\ &\quad + \Big\{ \Delta\eta(\boldsymbol{\alpha}_0, \Lambda_0) - G''(\boldsymbol{\alpha}_0, \Lambda_0) \Big\} \exp(2\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}) \int_0^Y \{\boldsymbol{h}_\alpha^{\mathrm{T}} \boldsymbol{X} + h_\Lambda(t)\}\, \mathrm{d}\Lambda_0(t) \Bigg] \mathrm{d}(\widehat{\Lambda} - \Lambda_0)(s) \Bigg) + o(1) \\ &\equiv (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)^{\mathrm{T}} \boldsymbol{W}_\alpha(\boldsymbol{h}_\alpha, h_\Lambda) + \int_0^\tau W_\Lambda(\boldsymbol{h}_\alpha, h_\Lambda)(s)\, \mathrm{d}(\widehat{\Lambda} - \Lambda_0)(s) + o(1). \end{aligned}$$

With the above arguments, we can use Theorem 3.3.1 of van der Vaart and Wellner (1996) to prove the weak convergence of $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0, \widehat{\Lambda} - \Lambda_0)$, following the arguments of Zeng

et al. (2008). The result is stated in Lemma 3.1 in Section 3.6.3.

We now turn to the proof of Theorem 3.1.

Proof of Theorem 3.1. First we consider the score statistic $U_\beta^{(j)}(\widehat{\boldsymbol{\alpha}}^{(j)}, \widehat{\Lambda}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ for $j = 1, \ldots, q$. For simplicity, we suppress the index j when the content is clear. Note that for any fixed $\mathcal{K}$, the Taylor's series expansion of $U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})$ at $(\boldsymbol{\alpha}_0, \Lambda_0, \boldsymbol{\gamma}_{0\mathcal{K}})$ is

$$
\begin{aligned}
&U_\beta(\widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}) \\
&= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[\mu_{1,i} \big\{ R_i S_i + (1 - R_i) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} \big\} + \widehat{\boldsymbol{I}}_{\beta\alpha}^{\mathrm{T}} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \\
&\quad + \int_0^\tau \widehat{I}_{\beta\Lambda}(s) \, \mathrm{d}(\widehat{\Lambda} - \Lambda_0)(s) + \widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \Big] + o_p(1) \\
&= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[\mu_{1,i} \{ R_i S_i + (1 - R_i) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} \} + \boldsymbol{I}_{\beta\alpha}^{\mathrm{T}} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \\
&\quad + \int_0^\tau I_{\beta\Lambda}(s) \, \mathrm{d}(\widehat{\Lambda} - \Lambda_0)(s) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \Big] + o_p(1) \\
&= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Big[\mu_{1,i} \big\{ R_i S_i + (1 - R_i) \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i} + \widetilde{\boldsymbol{q}}_\alpha^{\mathrm{T}} \boldsymbol{X}_i \big\} + \mu_{2,i} + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1} \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i}) \Big] \\
&\quad + o_p(1), \qquad (3.6)
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_{2,i} = \Delta_i \widetilde{q}_\Lambda(Y) &+ \Delta_i \frac{G_i''(\boldsymbol{\alpha}_0, \Lambda_0)}{G_i'(\boldsymbol{\alpha}_0, \Lambda_0)} \exp(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}_i) \int_0^\tau I(s \le Y_i) \widetilde{q}_\Lambda(s) \, \mathrm{d}\Lambda_0(s) \\
&- G_i'(\boldsymbol{\alpha}_0, \Lambda_0) \exp(\boldsymbol{\alpha}_0^{\mathrm{T}} \boldsymbol{X}_i) \int_0^\tau I(s \le Y_i) \widetilde{q}_\Lambda(s) \, \mathrm{d}\Lambda_0(s),
\end{aligned}
$$

and $(\widetilde{\boldsymbol{q}}_\alpha, \widetilde{q}_\Lambda) = (\boldsymbol{W}_\alpha, W_\Lambda)^{-1} (\boldsymbol{I}_{\beta\alpha}, I_{\beta\Lambda})$; the existence of the inverse is guaranteed by condition (C8). The second equality follows from the convergence of $\widehat{\boldsymbol{\gamma}}_{\mathcal{K}}$, $\widehat{\boldsymbol{I}}_{\beta\alpha}$ and $\widehat{I}_{\beta\Lambda}$ to the true values (by Lemma 3.2 presented below) and the convergence of $n^{-1/2} \sum_{i=1}^n (\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}} \widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1} - \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}} \boldsymbol{I}_{\gamma\gamma}^{-1}) \boldsymbol{W}_{\mathcal{K},i} R_i (S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}} \boldsymbol{W}_{\mathcal{K},i})$ to zero (by Lemma 3.4 presented below). The third equality follows from the convergence of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\Lambda}$ to the true values (by Lemma 3.1 presented

below). The first term on the right-hand side of (3.6) can be written as

$$
\begin{aligned}
&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big[\big\{\mu_{1,i}-\mathrm{E}(\mu_1\mid R_i,\boldsymbol{X}_i)\big\}\big\{R_iS_i+(1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}+\widetilde{\boldsymbol{q}}_\alpha^{\mathrm{T}}\boldsymbol{X}_i\big\}\\
&+\big\{\mu_{2,i}-\mathrm{E}(\mu_2\mid R_i,\boldsymbol{X}_i)\big\}\Big]\\
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big[(\boldsymbol{\gamma}_{0X}^{\mathrm{T}}+\widetilde{\boldsymbol{q}}_\alpha^{\mathrm{T}})\big\{\mathrm{E}(\mu_1\mid R_i,\boldsymbol{X}_i)\boldsymbol{X}_i-\mathrm{E}(\mu_1\boldsymbol{X}\mid R_i)\big\}\\
&+\big\{\mathrm{E}(\mu_1\mid R_i,\boldsymbol{X}_i)-\mathrm{E}(\mu_1\mid R_i)\big\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\\
&+\big\{\mathrm{E}(\mu_1\mid R_i,\boldsymbol{X}_i)+\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}\big\}R_i(S_i-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})+\big\{\mathrm{E}(\mu_2\mid R_i,\boldsymbol{X}_i)-\mathrm{E}(\mu_2\mid R_i)\big\}\Big]\\
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\Big\{(\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}+\widetilde{\boldsymbol{q}}_\alpha^{\mathrm{T}})\mathrm{E}(\mu_1\boldsymbol{X}\mid R_i)+\mathrm{E}(\mu_1\mid R_i)\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}+\mathrm{E}(\mu_2\mid R_i)\Big\}\\
\equiv&\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{1i}+\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{2i}+\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{3i},
\end{aligned}
$$

where $\boldsymbol{\gamma}_{0X}$ and $\boldsymbol{\gamma}_{0A,\mathcal{K}}$ are the subvectors of $\boldsymbol{\gamma}_{0\mathcal{K}}$ that correspond to $\boldsymbol{X}$ and the selected components of $\boldsymbol{A}$, respectively. Note that U_{1i}, U_{2i} and U_{3i} generally depend on the selected model $\mathcal{K}$. Now we reintroduce the index $j=1,\ldots,q$ for the transformation function. Let $U_{1i}^{(j)}$, $U_{2i}^{(j)}$ and $U_{3i}^{(j)}$ be U_{1i}, U_{2i} and U_{3i} computed under $G^{(j)}$, respectively. The score statistic under $\beta^{(j)}=0$ and $G^{(j)}$ can be written as

$$
U_\beta^{(j)}(\widehat{\boldsymbol{\alpha}}^{(j)},\widehat{\Lambda}^{(j)},\widehat{\boldsymbol{\gamma}}_{\mathcal{K}})=\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{1i}^{(j)}+\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{2i}^{(j)}+\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{3i}^{(j)}+o_p(1).
$$

Let $\boldsymbol{U}_{ki}=\big(U_{ki}^{(1)},\ldots,U_{ki}^{(q)}\big)^{\mathrm{T}}$ for $k=1,2$ and 3. For $j,l=1,\ldots,q$ and $k=1,2$ and 3, define $\sigma_k^{2(jl)}(\mathcal{K})=\mathrm{Cov}\big(U_{ki}^{(j)},U_{ki}^{(l)}\big)$. Let $\boldsymbol{\Sigma}_k(\mathcal{K})=\big(\sigma_k^{2(jl)}(\mathcal{K})\big)_{j,l=1,\ldots,q}$ for $k=1,2$ and 3, and $\boldsymbol{\Sigma}(\mathcal{K})=\sum_{k=1}^{3}\boldsymbol{\Sigma}_k(\mathcal{K})$.

By the Cramer-Wold Device, we only need to show that $\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K}^*)\widehat{\boldsymbol{U}}_\beta(\mathcal{K}^*)$ converges to a standard normal distribution for any vector $\boldsymbol{t}\in\mathbb{R}^q$. By a version of the portmanteau

theorem (Pollard, 2002, p.177), it suffices to show that for any $g \in \mathcal{C}_{\mathrm{B}}^3$,

$$\mathrm{E}\Big[g\Big\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K}^*)\widehat{\boldsymbol{U}}_\beta(\mathcal{K}^*)\Big\}\Big] \to \mathrm{E}\{g(Z)\}, \tag{3.7}$$

where Z is a standard normal random variable. Based on the above results and the mean-value theorem,

$$\begin{aligned}
&\mathrm{E}\Big[g\Big\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K}^*)\widehat{\boldsymbol{U}}_\beta(\mathcal{K}^*)\Big\}\Big] \\
&= \int \mathrm{E}\Big[g\Big\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})\widehat{\boldsymbol{U}}_\beta(\mathcal{K})\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big]\, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) \\
&= \int_{\mathcal{K}\in\Omega_n} \mathrm{E}\Big[g\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^n \boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})(\boldsymbol{U}_{1i}+\boldsymbol{U}_{2i}+\boldsymbol{U}_{3i})\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big]\, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) + o(1),
\end{aligned} \tag{3.8}$$

where $\mathcal{P}_{\mathcal{K}^*}$ is the probability measure of $\mathcal{K}^*$.

For $i = 1, \ldots, n$, let

$$\widetilde{\boldsymbol{U}}_{1i} = \mathrm{Var}\big(\boldsymbol{U}_1 \mid R_i, \boldsymbol{X}_i, S_i, \boldsymbol{A}_i\big)^{1/2}\boldsymbol{Z}_{1i},$$

where $\boldsymbol{Z}_{11}, \ldots, \boldsymbol{Z}_{1n}$ are i.i.d. standard multivariate normal variables that are independent of the observed data. Let $\boldsymbol{V}_{1i} = \widetilde{\boldsymbol{U}}_{11} + \cdots + \widetilde{\boldsymbol{U}}_{1,i-1} + \boldsymbol{U}_{1,i+1} + \cdots + \boldsymbol{U}_{1n}$ for $i = 1, \ldots, n$. Note that

$$\begin{aligned}
&\mathrm{E}\Big[g\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^n \boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})(\boldsymbol{U}_{1i}+\boldsymbol{U}_{2i}+\boldsymbol{U}_{3i})\Big\} \\
&\quad - g\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^n \boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})(\widetilde{\boldsymbol{U}}_{1i}+\boldsymbol{U}_{2i}+\boldsymbol{U}_{3i})\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big] \\
&= \sum_{i=1}^n \mathrm{E}\Big[g\Big\{\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\Big(\boldsymbol{V}_{1i} + \sum_j \boldsymbol{U}_{2j} + \sum_j \boldsymbol{U}_{3j} + \boldsymbol{U}_{1i}\Big)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&\quad - g\Big\{\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\Big(\boldsymbol{V}_{1i}+\sum_j \boldsymbol{U}_{2j}+\sum_j \boldsymbol{U}_{3j}+\widetilde{\boldsymbol{U}}_{1i}\Big)\Big\} \mid \mathcal{K}^*=\mathcal{K}\Big] \\
&= \frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\sum_{i=1}^{n}\mathrm{E}\Big[g'\Big\{\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\Big(\boldsymbol{V}_{1i}+\sum_j \boldsymbol{U}_{2j}+\sum_j \boldsymbol{U}_{3j}\Big)\Big\}(\boldsymbol{U}_{1i}-\widetilde{\boldsymbol{U}}_{1i}) \mid \mathcal{K}^*=\mathcal{K}\Big] \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}\mathrm{E}\Big(g''\Big\{\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\Big(\boldsymbol{V}_{1i}+\sum_j \boldsymbol{U}_{2j}+\sum_j \boldsymbol{U}_{3j}\Big)\Big\} \\
&\qquad\qquad \times\big[\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})\boldsymbol{U}_{1i}\}^2-\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})\widetilde{\boldsymbol{U}}_{1i}\}^2\big] \mid \mathcal{K}^*=\mathcal{K}\Big) \\
&\quad + \frac{1}{6n^{3/2}}\sum_{i=1}^{n}\mathrm{E}\Big[g'''(a)\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})\boldsymbol{U}_{1i}\}^3-g'''(\widetilde{a})\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})\widetilde{\boldsymbol{U}}_{1i}\}^3 \mid \mathcal{K}^*=\mathcal{K}\Big] \qquad (3.9)
\end{aligned}
$$

for some variables a and $\widetilde{a}$. By construction, $\boldsymbol{U}_{1i}$ and $\widetilde{\boldsymbol{U}}_{1i}$ are independent of $\boldsymbol{V}_{1i}$ and $\sum_j(\boldsymbol{U}_{2j}+\boldsymbol{U}_{3j})$ given $\mathcal{O}_1 \equiv (R_i, S_i, \boldsymbol{X}_i, \boldsymbol{A}_i)_{i=1,\ldots,n}$. The expectation in the first term on the right-hand side of (3.9) is

$$
\begin{aligned}
&\mathrm{E}\Big(\mathrm{E}\Big[g'\Big\{\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\Big(\boldsymbol{V}_{1i}+\sum_j \boldsymbol{U}_{2j}+\sum_j \boldsymbol{U}_{3j}\Big)\Big\} \mid \mathcal{O}_1, \mathcal{K}^*=\mathcal{K}\Big] \\
&\qquad \times \mathrm{E}(\boldsymbol{U}_{1i}-\widetilde{\boldsymbol{U}}_{1i} \mid \mathcal{O}_1, \mathcal{K}^*=\mathcal{K}) \mid \mathcal{K}^*=\mathcal{K}\Big)=\boldsymbol{0},
\end{aligned}
$$

because $\mathrm{E}(\boldsymbol{U}_{1i}-\widetilde{\boldsymbol{U}}_{1i} \mid \mathcal{O}_1, \mathcal{K}^*=\mathcal{K}) = \mathrm{E}(\boldsymbol{U}_{1i}-\widetilde{\boldsymbol{U}}_{1i} \mid \mathcal{O}_1)=\boldsymbol{0}$. Likewise, the second term on the right-hand side of (3.9) is 0, because the conditional second moments of $\boldsymbol{U}_{1i}$ and $\widetilde{\boldsymbol{U}}_{1i}$ given $\mathcal{O}_1$ match ($i=1,\ldots,n$). For $\mathcal{K}\in\Omega_n$, the right-hand side of (3.9) is bounded above by

$$
\sum_{j=1}^{q}\zeta_{1n}^{(j)} = \sum_{j=1}^{q} M n^{-3/2}\sum_{i=1}^{n}\mathrm{E}\Big\{\sup_{\mathcal{K}\in\Omega_n}\Big(|U_{1i}^{(j)}|^3+|\widetilde{U}_{1i}^{(j)}|^3\Big) \mid \mathcal{K}^*=\mathcal{K}\Big\}
$$

for some positive constant M. By Lemma 3.5 presented below, $\int_{\Omega_n}\zeta_{1n}^{(j)}\,\mathrm{d}\mathcal{P}_{\mathcal{K}^*}\to 0$ for $j=1,\ldots,q$.

Next, we show that $\boldsymbol{U}_{2i}$'s in (3.8) can be similarly replaced by normal random vari-

ables. Define

$$\widetilde{\boldsymbol{U}}_{2i} = \mathrm{Var}(\boldsymbol{U}_2 \mid R_i, \boldsymbol{A}_i, S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}} \boldsymbol{X}_i)^{1/2} \boldsymbol{Z}_{2i}$$

for $i = 1, \ldots, n$, where $\boldsymbol{Z}_{21}, \ldots, \boldsymbol{Z}_{2n}$ are i.i.d. standard multivariate normal random variables that are independent of the observed data and $\boldsymbol{Z}_{11}, \ldots, \boldsymbol{Z}_{1n}$. Note that the above conditional variance is taken with respect to $\boldsymbol{X}$. We wish to show that

$$\begin{aligned}\int_{\Omega_n} \mathrm{E}\Big[g\Big\{ \sum_{i=1}^{n} \frac{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}} (\widetilde{\boldsymbol{U}}_{1i} + \boldsymbol{U}_{2i} + \boldsymbol{U}_{3i}) \Big\} \\ - g\Big\{ \sum_{i=1}^{n} \frac{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}} (\widetilde{\boldsymbol{U}}_{1i} + \widetilde{\boldsymbol{U}}_{2i} + \boldsymbol{U}_{3i}) \Big\} \mid \mathcal{K}^* = \mathcal{K} \Big] \, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) = o(1). \end{aligned} \tag{3.10}$$

Note that $n^{-1/2} \sum_{i=1}^{n} \widetilde{\boldsymbol{U}}_{1i}$ can be written as $\boldsymbol{\Sigma}_1^{1/2}(\mathcal{K}) \boldsymbol{Z}_1$, where $\boldsymbol{Z}_1$ is a standard multivariate normal random variable independent of the observed data. Let $\mathcal{K}_0^* = \mathcal{K}^*(\mathcal{S} - \mathcal{X}\boldsymbol{\gamma}_{0X}, \mathcal{A})$. By the proof of Theorem 2.1, the event $\{\mathcal{K}^* = \mathcal{K}\}$ in the conditional expectation in (3.10) can be replaced by $\{\mathcal{K}_0^* = \mathcal{K}\}$. Let $\boldsymbol{V}_{2i} = \widetilde{\boldsymbol{U}}_{21} + \cdots + \widetilde{\boldsymbol{U}}_{2,i-1} + \boldsymbol{U}_{2,i+1} + \cdots + \boldsymbol{U}_{2n}$ for $i = 1, \ldots, n$. The term inside the integration of the left-hand side of (3.10) is up to a vanishing term equal to

$$\begin{aligned}&\sum_{i=1}^{n} \mathrm{E}\Big[g\Big\{ \frac{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}} \Big(\boldsymbol{V}_{2i} + n^{1/2} \boldsymbol{\Sigma}_1^{1/2}(\mathcal{K}) \boldsymbol{Z}_1 + \boldsymbol{U}_{2i} + \sum_j \boldsymbol{U}_{3j} \Big) \Big\} \\ &\qquad - g\Big\{ \frac{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}} \Big(\boldsymbol{V}_{2i} + n^{1/2} \boldsymbol{\Sigma}_1^{1/2}(\mathcal{K}) \boldsymbol{Z}_1 + \widetilde{\boldsymbol{U}}_{2i} + \sum_j \boldsymbol{U}_{3j} \Big) \Big\} \mid \mathcal{K}_0^* = \mathcal{K} \Big] \\ &= \frac{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}} \sum_{i=1}^{n} \mathrm{E}\Big[g'\Big\{ \frac{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}} \Big(\boldsymbol{V}_{2i} + n^{1/2} \boldsymbol{\Sigma}_1^{1/2}(\mathcal{K}) \boldsymbol{Z}_1 + \sum_j \boldsymbol{U}_{3j} \Big) \Big\} (\boldsymbol{U}_{2i} - \widetilde{\boldsymbol{U}}_{2i}) \mid \mathcal{K}_0^* = \mathcal{K} \Big] \\ &\quad + \frac{1}{2n} \sum_{i=1}^{n} \mathrm{E}\Big(g''\Big\{ \frac{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}} \Big(\boldsymbol{V}_{2i} + n^{1/2} \boldsymbol{\Sigma}_1^{1/2}(\mathcal{K}) \boldsymbol{Z}_1 + \sum_j \boldsymbol{U}_{3j} \Big) \Big\} \\ &\qquad\qquad \times \big[\{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K}) \boldsymbol{U}_{2i}\}^2 - \{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K}) \widetilde{\boldsymbol{U}}_{2i}\}^2 \big] \mid \mathcal{K}_0^* = \mathcal{K} \Big) \\ &\quad + \frac{1}{6n^{3/2}} \sum_{i=1}^{n} \mathrm{E}\Big[g'''(b) \{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K}) \boldsymbol{U}_{2i}\}^3 - g'''(\widetilde{b}) \{\boldsymbol{t}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1/2}(\mathcal{K}) \widetilde{\boldsymbol{U}}_{2i}\}^3 \mid \mathcal{K}_0^* = \mathcal{K} \Big] \end{aligned} \tag{3.11}$$

for some variables b and $\widetilde{b}$. Let $\mathcal{O}_2 = (R_i, \boldsymbol{A}_i, S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}} \boldsymbol{X}_i)_{i=1,\ldots,n}$. Since the event $\{\mathcal{K}_0^* = \mathcal{K}\}$ is implied by $\mathcal{O}_2$, we have

$$\begin{aligned}
&\mathrm{E}\Big[g'\Big\{\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\Big(\boldsymbol{V}_{2i} + n^{1/2}\boldsymbol{\Sigma}_1^{1/2}(\mathcal{K})\boldsymbol{Z}_1 + \sum_j \boldsymbol{U}_{3j}\Big)\Big\}(\boldsymbol{U}_{2i} - \widetilde{\boldsymbol{U}}_{2i}) \mid \mathcal{K}_0^* = \mathcal{K}\Big] \\
&= \mathrm{E}\Big(\mathrm{E}\Big[g'\Big\{\frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}\Big(\boldsymbol{V}_{2i} + n^{1/2}\boldsymbol{\Sigma}_1^{1/2}(\mathcal{K})\boldsymbol{Z}_1 + \sum_j \boldsymbol{U}_{3j}\Big)\Big\} \mid \mathcal{O}_2, \mathcal{K}_0^* = \mathcal{K}\Big] \\
&\qquad \times \mathrm{E}(\boldsymbol{U}_{2i} - \widetilde{\boldsymbol{U}}_{2i} \mid \mathcal{O}_2) \mid \mathcal{K}_0^* = \mathcal{K}\Big) = \boldsymbol{0}.
\end{aligned}$$

Likewise, the second term on the right-hand side of (3.11) is zero because the conditional second moments of $\boldsymbol{U}_{2i}$ and $\widetilde{\boldsymbol{U}}_{2i}$ match. By Lemma 3.5 presented below, the third term on the right-hand side of (3.11) is bounded by $\sum_{j=1}^q \zeta_{2n}^{(j)}$ such that $\int_{\Omega_n} \zeta_{2n}^{(j)} \, \mathrm{d}\mathcal{P}_{\mathcal{K}^*} \to 0$, and (3.10) holds.

Let $\widetilde{\boldsymbol{U}}_{3i} = \mathrm{Var}(\boldsymbol{U}_3 \mid \boldsymbol{A}_{\mathcal{K},i})^{1/2}\boldsymbol{Z}_{3i}$ for $i = 1, \ldots, n$, where $\boldsymbol{Z}_{31}, \ldots, \boldsymbol{Z}_{3n}$ are i.i.d. standard multivariate normal variables that are independent of the observed data and $(\boldsymbol{Z}_{1i}, \boldsymbol{Z}_{2i})_{i=1,\ldots,n}$. By arguments similar to the proof of Theorem 2.1, we can show that

$$\begin{aligned}
\int_{\Omega_n} &\mathrm{E}\Big[g\Big\{\sum_{i=1}^n \frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}(\widetilde{\boldsymbol{U}}_{1i} + \widetilde{\boldsymbol{U}}_{2i} + \boldsymbol{U}_{3i})\Big\} \\
&- g\Big\{\sum_{i=1}^n \frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}(\widetilde{\boldsymbol{U}}_{1i} + \widetilde{\boldsymbol{U}}_{2i} + \widetilde{\boldsymbol{U}}_{3i})\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big] \, \mathrm{d}\mathcal{P}_{\mathcal{K}^*}(\mathcal{K}) = o(1). \qquad (3.12)
\end{aligned}$$

Combining the above results, we have

$$\begin{aligned}
&\mathrm{E}\Big[g\Big\{\sum_{i=1}^n \frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}(\boldsymbol{U}_{1i} + \boldsymbol{U}_{2i} + \boldsymbol{U}_{3i})\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big] \\
&= \mathrm{E}\Big[g\Big\{\sum_{i=1}^n \frac{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})}{n^{1/2}}(\widetilde{\boldsymbol{U}}_{1i} + \widetilde{\boldsymbol{U}}_{2i} + \widetilde{\boldsymbol{U}}_{3i})\Big\} \mid \mathcal{K}^* = \mathcal{K}\Big] + o(1) \\
&= \mathrm{E}\Big[g\Big\{\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}(\mathcal{K})(\boldsymbol{\Sigma}_1^{1/2}(\mathcal{K})\boldsymbol{Z}_1 + \boldsymbol{\Sigma}_2^{1/2}(\mathcal{K})\boldsymbol{Z}_2 + \boldsymbol{\Sigma}_3^{1/2}(\mathcal{K})\boldsymbol{Z}_3)\Big\}\Big] + o(1)
\end{aligned}$$

uniformly over $\mathcal{K} \in \Omega_n$, where $\boldsymbol{Z}_2$ is a standard multivariate normal random variable independent of the observed data and $\boldsymbol{Z}_1$, and $\boldsymbol{Z}_3$ is a standard normal random variable independent of $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, and the observed data. Because $\boldsymbol{\Sigma}_1^{1/2}(\mathcal{K})\boldsymbol{Z}_1 + \boldsymbol{\Sigma}_2^{1/2}(\mathcal{K})\boldsymbol{Z}_2 + \boldsymbol{\Sigma}_3^{1/2}(\mathcal{K})\boldsymbol{Z}_3$ is multivariate normal with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}(\mathcal{K})$, the desired convergence (3.7) follows. □

3.6.3 Additional Theoretical Results

For simplicity, Lemma 3.1, Lemma 3.2, Lemma 3.4 and Lemma 3.5 are stated under the null hypothesis H_0 and a known transformation function G. The arguments for $G^{(j)},\ j = 1, \ldots, q$ under H_0' are essentially the same. Lemma 3.3 is stated under the null hypothesis H_0'.

Lemma 3.1. *Suppose that there exists* $(\boldsymbol{h}_\alpha, h_\Lambda) \in \mathcal{H}$ *such that* $(\boldsymbol{W}_\alpha(\boldsymbol{h}_\alpha, h_\Lambda), W_\Lambda(\boldsymbol{h}_\alpha, h_\Lambda)) = (\boldsymbol{q}_\alpha, q_\Lambda)$*, then under conditions (C1), (C7) and (C8), we have*

$$\sqrt{n}\Big\{\boldsymbol{q}_\alpha^{\mathrm{T}}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \int_0^\tau q_\Lambda(s)\,\mathrm{d}(\widehat{\Lambda} - \Lambda_0)(s)\Big\} = -\sqrt{n}(\mathcal{P}_n - \mathcal{P})U(\boldsymbol{\alpha}_0, \Lambda_0)[\widetilde{\boldsymbol{q}}_\alpha, \widetilde{q}_\Lambda] + o_p(1),$$

where $\boldsymbol{q}_\alpha = \boldsymbol{I}_{\beta\alpha}$*,* $q_\Lambda = I_{\beta\Lambda}$ *and* $(\widetilde{\boldsymbol{q}}_\alpha, \widetilde{q}_\Lambda) = (\boldsymbol{W}_\alpha, W_\Lambda)^{-1}(\boldsymbol{q}_\alpha, q_\Lambda)$*.*

Lemma 3.2. *Under conditions (C1)–(C3), the inequalities*

$$\sup_{\mathcal{K}\in\Omega_n} \Big\|\widehat{\boldsymbol{\gamma}}_{\mathcal{K}} - \boldsymbol{\gamma}_{0\mathcal{K}}\Big\| > C_1\bigg\{\Big(\frac{t + \log r_n + q_n}{n}\Big)^{1/2} + \frac{q_n(\log n)^{2/\xi}(t + \log r_n + q_n)^{2/\xi}}{n}\bigg\},$$

$$\sup_{\mathcal{K}\in\Omega_n} \Big\|\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}\widehat{\boldsymbol{I}}_{\beta\gamma} - \boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{I}_{\beta\gamma}\Big\| > C_1\bigg\{\Big(\frac{t + \log r_n + q_n}{n}\Big)^{1/2} + \frac{q_n(\log n)^{2/\xi}(t + \log r_n + q_n)^{2/\xi}}{n}\bigg\},$$

$$\sup_{\mathcal{K}\in\Omega_n} \Big\|\widehat{\boldsymbol{I}}_{\beta\alpha} - \boldsymbol{I}_{\beta\alpha}\Big\| > C_1\bigg\{\Big(\frac{t + \log r_n}{n}\Big)^{1/2} + \frac{q_n^{1/2}(\log n)^{1/\xi}(t + \log r_n)^{1/\min(1,\xi)}}{n}\bigg\},\ \textit{and}$$

$$\sup_{\mathcal{K}\in\Omega_n}\sup_{0\le s\le\tau} \Big|\widehat{I}_{\beta\Lambda}(s) - I_{\beta\Lambda}(s)\Big|$$

$$> C_1\bigg[\frac{(tq_n)^{1/2}}{n^{1/4}} + \Big\{\frac{t + \log n + \log(2r_n)}{n}\Big\}^{1/2} + \frac{q_n^{1/2}\{\log(2n)\}^{1/\xi}\{t + \log n + \log(2r_n)\}^{1/\min(1,\xi)}}{n}\bigg]$$

hold with probability at most $C_2 e^{-t}$ for large enough n and t, where C_1 and C_2 are positive constants.

For $j, l = 1, \ldots, q$, we have

$$\begin{aligned}
&\operatorname{Cov}\{U_\beta^{(j)}(\widehat{\boldsymbol{\alpha}}^{(j)}, \widehat{\Lambda}^{(j)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}}), U_\beta^{(l)}(\widehat{\boldsymbol{\alpha}}^{(l)}, \widehat{\Lambda}^{(l)}, \widehat{\boldsymbol{\gamma}}_{\mathcal{K}})\} \\
&= \operatorname{Cov}\Big\{\frac{1}{n^{1/2}}\sum_{i=1}^{n}(U_{1i}^{(j)} + U_{2i}^{(j)} + U_{3i}^{(j)}), \frac{1}{n^{1/2}}\sum_{i=1}^{n}(U_{1i}^{(l)} + U_{2i}^{(l)} + U_{3i}^{(l)})\Big\} \\
&= \sum_{k=1}^{3}\operatorname{Cov}\Big(\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{ki}^{(j)}, \frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{ki}^{(l)}\Big),
\end{aligned}$$

where the last equality follows because the cross terms are zero. To see this, consider $k = 1$ and $k' = 2$. We have

$$\begin{aligned}
&\operatorname{Cov}\Big(\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{1i}^{(j)}, \frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{2i}^{(l)}\Big) \\
&= \mathrm{E}\Big\{\operatorname{Cov}\Big(\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{1i}^{(j)}, \frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{2i}^{(l)} \mid \mathcal{O}_1\Big)\Big\} \\
&\qquad + \operatorname{Cov}\Big\{\mathrm{E}\Big(\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{1i}^{(j)} \mid \mathcal{O}_1\Big), \mathrm{E}\Big(\frac{1}{n^{1/2}}\sum_{i=1}^{n}U_{2i}^{(l)} \mid \mathcal{O}_1\Big)\Big\} = 0,
\end{aligned}$$

because $\mathrm{E}(U_{1i}^{(j)} \mid \mathcal{O}_1) = 0$, and $U_{2i}^{(l)}$ is constant given $\mathcal{O}_1$. Analogously, the other cross terms are zero.

For $j, l = 1, \ldots, q$, define

$$\begin{aligned}
\widehat{\sigma}_1^{2(jl)}(\mathcal{K}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\big(U_{1i}^{(j)}U_{1i}^{(l)} \mid R_i, \boldsymbol{X}_i, S_i, \boldsymbol{A}_i\big) \\
\widehat{\sigma}_2^{2(jl)}(\mathcal{K}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\big(U_{2i}^{(j)}U_{2i}^{(l)} \mid R_i, \boldsymbol{A}_i, S_i - \boldsymbol{\gamma}_{0X}^{\mathrm{T}}\boldsymbol{X}_i\big) \\
\widehat{\sigma}_3^{2(jl)}(\mathcal{K}) &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\big(U_{3i}^{(j)}U_{3i}^{(l)} \mid \boldsymbol{A}_{\mathcal{K},i}\big).
\end{aligned}$$

Lemma 3.3. *Under conditions (C1)–(C4), for $j, l = 1, \ldots, q$, and large enough n and t,*

$$P\left[\sup_{\mathcal{K}\in\Omega_n}\sum_{k=1}^{3}\left|\widehat{\sigma}_k^{2(jl)}(\mathcal{K})-\sigma_k^{(2jl)}(\mathcal{K})\right| > C_1\left\{\left(\frac{t+\log r_n}{n}\right)^{1/2}+\frac{q_n(\log n)^{2/\xi}(t+\log r_n)^{2/\xi}}{n}\right\}\right]$$

$$\leq C_2 e^{-t},$$

where C_1 and C_2 are positive constants.

Lemma 3.4. *Under conditions (C1)–(C4),*

$$\mathrm{E}\left\{\sup_{\mathcal{K}\in\Omega_n}\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(\widehat{\boldsymbol{I}}_{\beta\gamma}^{\mathrm{T}}\widehat{\boldsymbol{I}}_{\gamma\gamma}^{-1}-\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1})\boldsymbol{W}_{\mathcal{K},i}R_i(S_i-\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i})\right|\right\}=o(1).$$

Lemma 3.5. *Assume that conditions (C1)–(C3) hold. For U_{ki} and $\widetilde{U}_{ki}$ ($k = 1, 2, 3$; $i = 1, \ldots, n$) defined in the proof of Theorem 3.1,*

$$P\left[\sum_{k=1}^{3}\sup_{\mathcal{K}\in\Omega_n}\frac{1}{n^{3/2}}\sum_{i=1}^{n}\left(|U_{ki}|^3+|\widetilde{U}_{ki}|^3\right) > C_1\left\{\frac{(t+\log r_n)^{1/2}}{n}+\frac{q_n^{3/2}(\log n)^{3/\xi}(t+\log r_n)^{3/\xi}}{n^{3/2}}\right\}\right]$$

is smaller than $C_2 e^{-t}$ for large enough n and t, where C_1 and C_2 are positive constants.

The proofs of Lemmas 3.3 and 3.4 are analogous to the arguments in Section 2.6.4, and we omit the proofs here.

Proof of Lemma 3.1. The result follows from Theorem 3.3.1 of van der Vaart and Wellner (1996). □

Proof of Lemma 3.2. We refer the proofs of the first to the third results to Section 2.6.4.

For the forth result, let $\zeta_m = \{0, \frac{\tau}{m}, \frac{2\tau}{m}, \ldots, \tau\}$ for $m = n^{1/2}$, so

$$\begin{aligned}
&\sup_{\mathcal{K}\in\Omega_n} \sup_{0\le s\le\tau} \Big|\widehat{I}_{\beta\Lambda}(s) - I_{\beta\Lambda}(s)\Big| \\
&\le \sup_{\mathcal{K}\in\Omega_n} \sup_{s\in\zeta_m} \Big|\widehat{I}_{\beta\Lambda}(s) - I_{\beta\Lambda}(s)\Big| + \sup_{\mathcal{K}\in\Omega_n} \sup_{s,s':|s-s'|\le\tau/m} \Big|\widehat{I}_{\beta\Lambda}(s) - \widehat{I}_{\beta\Lambda}(s')\Big| \\
&\quad + \sup_{\mathcal{K}\in\Omega_n} \sup_{s,s':|s-s'|\le\tau/m} \Big|I_{\beta\Lambda}(s) - I_{\beta\Lambda}(s')\Big|. \qquad (3.13)
\end{aligned}$$

The second term on the right-hand side of (3.13) can be written as

$$\begin{aligned}
&\sup_{\mathcal{K}\in\Omega_n} \sup_{s,s':|s-s'|\le\tau/m} \Big|\frac{1}{n}\sum_{i=1}^n \widehat{I}_{\beta\Lambda,i}\big\{I(s\le Y_i) - I(s'\le Y_i)\big\}\Big| \\
&\le \sup_{\mathcal{K}\in\Omega_n} \sup_{s,s':|s-s'|\le\tau/m} \Big|\Big(\frac{1}{n}\sum_{i=1}^n \widehat{I}^2_{\beta\Lambda,i}\Big)^{1/2}\Big[\frac{1}{n}\sum_{i=1}^n \big\{I(s\le Y_i) - I(s'\le Y_i)\big\}^2\Big]^{1/2}\Big| \\
&= \sup_{\mathcal{K}\in\Omega_n} \Big(\frac{1}{n}\sum_{i=1}^n \widehat{I}^2_{\beta\Lambda,i}\Big)^{1/2} \times \sup_{s,s':|s-s'|\le\tau/m} \Big[\frac{1}{n}\sum_{i=1}^n \big\{I(s\le Y_i) - I(s'\le Y_i)\big\}^2\Big]^{1/2},
\end{aligned}$$

where $\widehat{I}_{\beta\Lambda,i} = \mu'_{1i}\exp(\boldsymbol{\alpha}_0^{\mathrm{T}}\boldsymbol{X}_i)\{R_iS_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\}$, and the inequality follows from the Cauchy-Schwarz inequality. Note that

$$\sup_{\mathcal{K}\in\Omega_n} \frac{1}{n}\sum_{i=1}^n \widehat{I}^2_{\beta\Lambda,i} > M_1\Big[q_n + \Big\{\frac{t+\log(2r_n)}{n}\Big\}^{1/2} + \frac{q_n\{\log(2n)\}^{1/\xi}\{t+\log(2r_n)\}^{2/\xi}}{n}\Big]$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_1, following the arguments in Section 2.6.5. Let I_k denote the interval $\big[\frac{(k-2)\tau}{m}, \frac{k\tau}{m}\big)$ for $k = 2, \ldots, m$. Since $I(s\le Y_i) - I(s'\le Y_i)$ takes value 1 if Y_i is between s and s', and takes value 0 otherwise, we have

$$\begin{aligned}
&P\Big(\sup_{s,s':|s-s'|\le\tau/m} \Big[\frac{1}{n}\sum_{i=1}^n \big\{I(s\le Y_i) - I(s'\le Y_i)\big\}^2\Big]^{1/2} > \Big(\frac{t}{m}\Big)^{1/2}\Big) \\
&= P\Big[\sup_{s,s':|s-s'|\le\tau/m} \frac{1}{n}\sum_{i=1}^n I\{\min(s,s') \le Y_i < \max(s,s')\} > \frac{t}{m}\Big]
\end{aligned}$$

$$\le P\Big\{ \sup_{k=1,\dots,m} \frac{1}{n}\sum_{i=1}^{n} I(Y_i \in I_k) > \frac{t}{m} \Big\}$$
$$\le m \sup_{k=1,\dots,m} P\Big\{ \frac{1}{n}\sum_{i=1}^{n} I(Y_i \in I_k) > \frac{t}{m} \Big\},$$

for any positive t. Let $p = P(Y_i \in I_k)$. By Bernstein's inequality, we have

$$P\Big\{ \frac{1}{n}\sum_{i=1}^{n} I(Y_i \in I_k) > \frac{t}{m} \Big\} \le \exp\Big\{ - \frac{n^2(\frac{t}{m} - p)^2/2}{np + n(\frac{t}{m} - p)/3} \Big\}.$$

Note that $p \le \sup_{s\in[0,\tau]} f_Y(s)/m$, where f_Y is the density of Y; the supremum is finite under condition (C7). Then

$$P\Big(\sup_{s,s':|s-s'|\le\tau/m} \Big[\frac{1}{n}\sum_{i=1}^{n} \big\{ I(s \le Y_i) - I(s' \le Y_i) \big\}^2 \Big]^{1/2} > \Big(\frac{t}{m} \Big)^{1/2} \Big)$$
$$\le m \exp\Big\{ - \frac{n}{2} \frac{(t - \sup f_Y)^2}{m(t/2 + 2\sup f_Y/3)} \Big\} \lesssim e^{-t}$$

for large enough t. Thus

$$\sup_{\mathcal{K}\in\Omega_n} \sup_{s,s':|s-s'|\le\tau/m} \Big| \frac{1}{n}\sum_{i=1}^{n} \widehat{I}_{\beta\Lambda,i} \big\{ I(s \le Y_i) - I(s' \le Y_i) \big\} \Big|$$
$$> M_2 t^{1/2} \Big[\frac{q_n}{n^{1/2}} + \Big\{ \frac{t + \log(2r_n)}{n^2} \Big\}^{1/2} + \frac{q_n \{\log(2n)\}^{1/\xi} \{t + \log(2r_n)\}^{2/\xi}}{n^{3/2}} \Big]^{1/2}$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_2. Clearly, because the derivative of $I_{\beta\Lambda}$ is bounded, $\sup_{\mathcal{K}\in\Omega_n} \sup_{s,s':|s-s'|\le\tau/m} \Big| I_{\beta\Lambda}(s) - I_{\beta\Lambda}(s') \Big| = O(m^{-1})$. Using Lemma 2.5 again, we have

$$\sup_{\mathcal{K}\in\Omega_n} \sup_{s\in\zeta_m} \Big| \widehat{I}_{\beta\Lambda}(s) - I_{\beta\Lambda}(s) \Big|$$
$$> M_3 \Big[\Big\{ \frac{t + \log n + \log(2r_n)}{n} \Big\}^{1/2} + \frac{q_n^{1/2} \{\log(2n)\}^{1/\xi} \{t + \log n + \log(2r_n)\}^{1/\min(1,\xi)}}{n} \Big]$$

with probability at most $3e^{-t}$. Combining the above results yield the desired result. $\square$

Proof of Lemma 3.5. Recall that $U_{1i} = \{\mu_{1,i} - \mathrm{E}(\mu_1 \mid R_i, \boldsymbol{X}_i)\}\{R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \widetilde{\boldsymbol{q}}_\alpha^{\mathrm{T}}\boldsymbol{X}_i\} + \{\mu_{2,i} - \mathrm{E}(\mu_2 \mid R_i, \boldsymbol{X}_i)\}$. Note that

$$\big\|U_{1i}^3\big\|_{\psi_{\xi/3}} = O\big(1 + \big\|R_i S_i + (1-R_i)\boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i} + \widetilde{\boldsymbol{q}}_\alpha^{\mathrm{T}}\boldsymbol{X}_i\big\|_{\psi_\xi}^3\big) \lesssim q_n^{3/2}.$$

By Lemma 2.5,

$$\begin{aligned}&\sup_{\mathcal{K}\in\Omega_n}\Big\{\frac{1}{n}\sum_{i=1}^n |U_{1i}|^3 - \mathrm{E}(|U_{11}|^3)\Big\}\\ &> M_1\Big[\Big\{\frac{t+\log(2r_n)}{n}\Big\}^{1/2} + \frac{q_n^{3/2}\{\log(2n)\}^{3/\xi}\{t+\log(2r_n)\}^{3/\xi}}{n}\Big]\end{aligned}$$

with probability at most $3e^{-t}$ for any $t > 0$ and some positive constant M_1. Because $\mathrm{E}(|U_{11}|^3)$ is uniformly bounded over $\mathcal{K} \in \Omega_n$,

$$\frac{1}{n^{3/2}}\sup_{\mathcal{K}\in\Omega_n}\sum_{i=1}^n |U_{1i}|^3 > M_2\Big[\frac{\{t+\log(2r_n)\}^{1/2}}{n} + \frac{q_n^{3/2}\{\log(2n)\}^{3/\xi}\{t+\log(2r_n)\}^{3/\xi}}{n^{3/2}}\Big]$$

with probability at most $M_3 e^{-t}$ for any $t > 0$ and some positive constants M_2 and M_3. Recall that U_{2i} is equal to

$$\begin{aligned}&(\boldsymbol{\gamma}_{0X}^{\mathrm{T}} + \widetilde{\boldsymbol{q}}_\alpha^{\mathrm{T}})\{\mathrm{E}(\mu_1 \mid R_i, \boldsymbol{X}_i)\boldsymbol{X}_i - \mathrm{E}(\mu_1\boldsymbol{X} \mid R_i)\} + \{\mathrm{E}(\mu_1 \mid R_i, \boldsymbol{X}_i) - \mathrm{E}(\mu_1 \mid R_i)\}\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\\ &+ \{\mathrm{E}(\mu_1 \mid R_i, \boldsymbol{X}_i) + \boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K},i}\}R_i(S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}) + \{\mathrm{E}(\mu_2 \mid R_i, \boldsymbol{X}_i) - \mathrm{E}(\mu_2 \mid R_i)\}\end{aligned}$$

and note that $\|\boldsymbol{I}_{\beta\gamma}^{\mathrm{T}}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{W}_{\mathcal{K}}\|$ is bounded, so

$$\big\|U_{2i}^3\big\|_{\psi_{\xi/3}} = O\big(1 + \big\|\boldsymbol{\gamma}_{0A,\mathcal{K}}^{\mathrm{T}}\boldsymbol{A}_{\mathcal{K},i}\big\|_{\psi_\xi}^3 + \big\|S_i - \boldsymbol{\gamma}_{0\mathcal{K}}^{\mathrm{T}}\boldsymbol{W}_{\mathcal{K},i}\big\|_{\psi_\xi}^3\big) \lesssim q_n^{3/2}.$$

By Lemma 2.5,

$$\sup_{\mathcal{K}\in\Omega_n}\Big\{\frac{1}{n}\sum_{i=1}^{n}|U_{2i}|^3-\mathrm{E}(|U_{21}|^3)\Big\}$$
$$>M_4\Big[\Big\{\frac{t+\log(2r_n)}{n}\Big\}^{1/2}+\frac{q_n^{3/2}\{\log(2n)\}^{3/\xi}\{t+\log(2r_n)\}^{3/\xi}}{n}\Big]$$

with probability at most $3e^{-t}$ for any $t>0$ and some positive constant M_4. Because $\mathrm{E}(|U_{21}|^3)$ is uniformly bounded over $\mathcal{K}\in\Omega_n$,

$$\frac{1}{n^{3/2}}\sup_{\mathcal{K}\in\Omega_n}\sum_{i=1}^{n}|U_{2i}|^3>M_5\Big[\frac{\{t+\log(2r_n)\}^{1/2}}{n}+\frac{q_n^{3/2}\{\log(2n)\}^{3/\xi}\{t+\log(2r_n)\}^{3/\xi}}{n^{3/2}}\Big]$$

with probability at most M_6e^{-t} for any $t>0$ and some positive constants M_5 and M_6. Similar arguments show that the same bound applies to the terms involving $\widetilde{U}_{1i}$, $\widetilde{U}_{2i}$, U_{3i}, and $\widetilde{U}_{3i}$. □

3.6.4 Additional Numerical Results

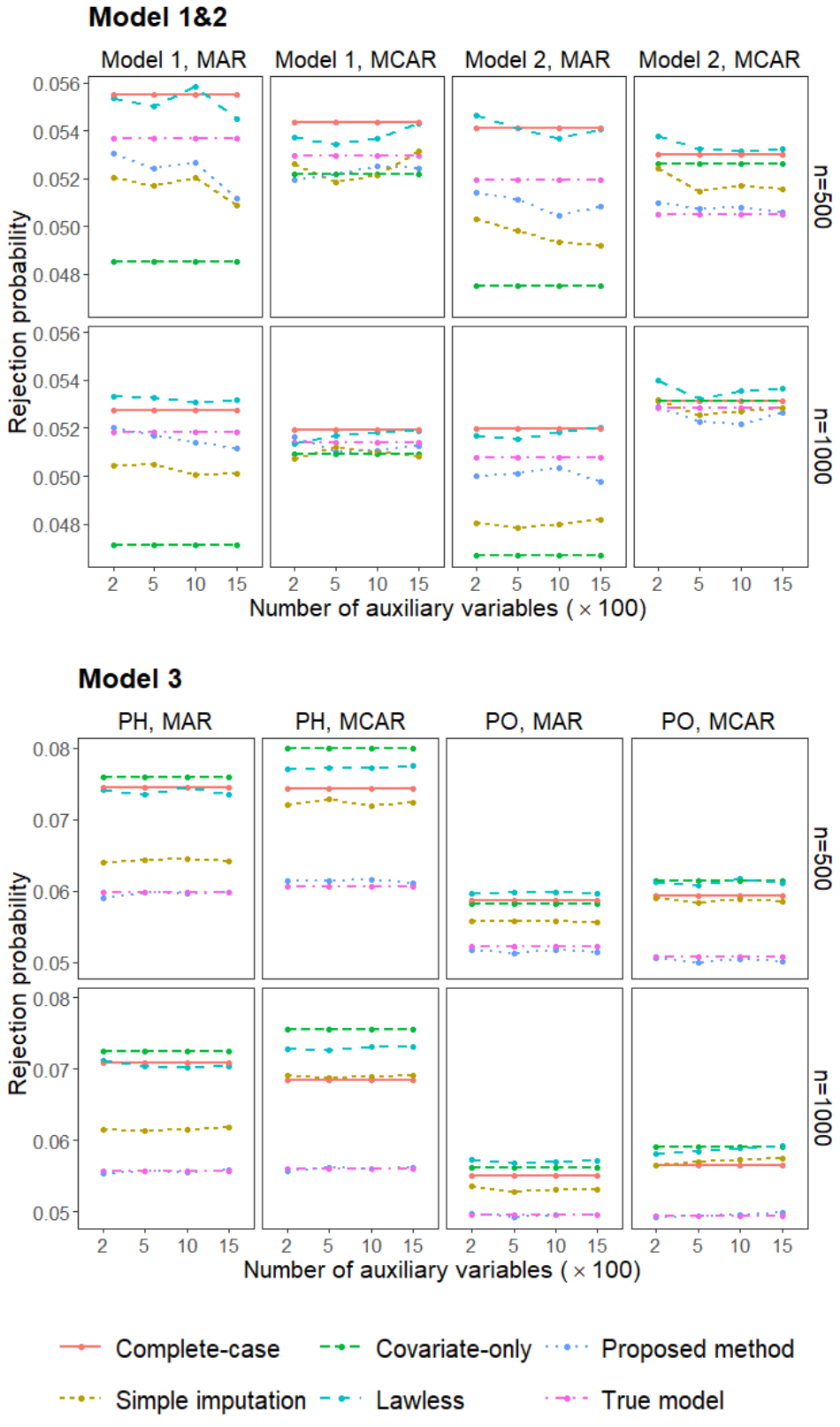


Figure 3.5: Study 1 - Rejection probabilities under a missing proportion of 30% and the null hypothesis.

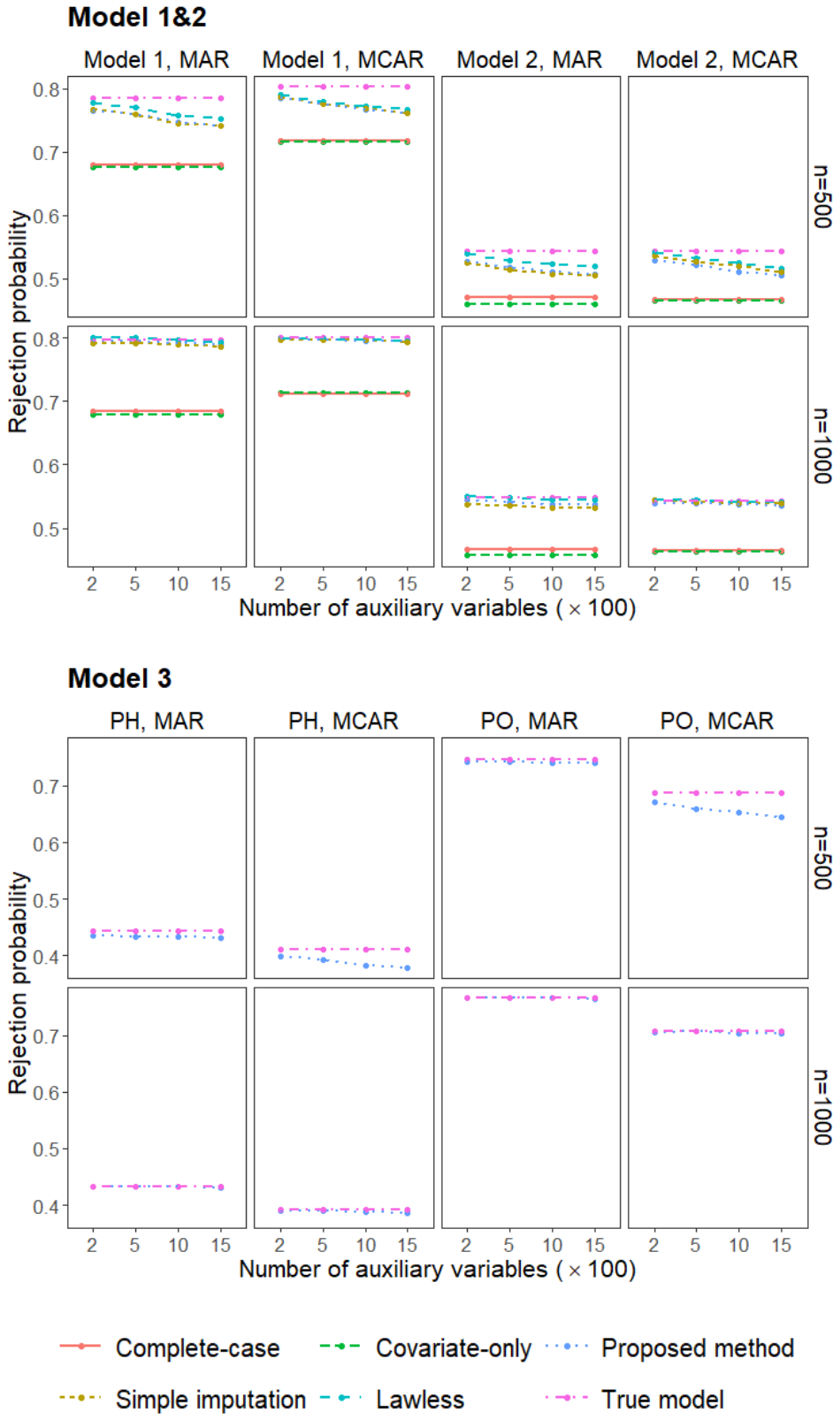


Figure 3.6: Study 1 - Rejection probabilities under a missing proportion of 30% and the alternative hypothesis.

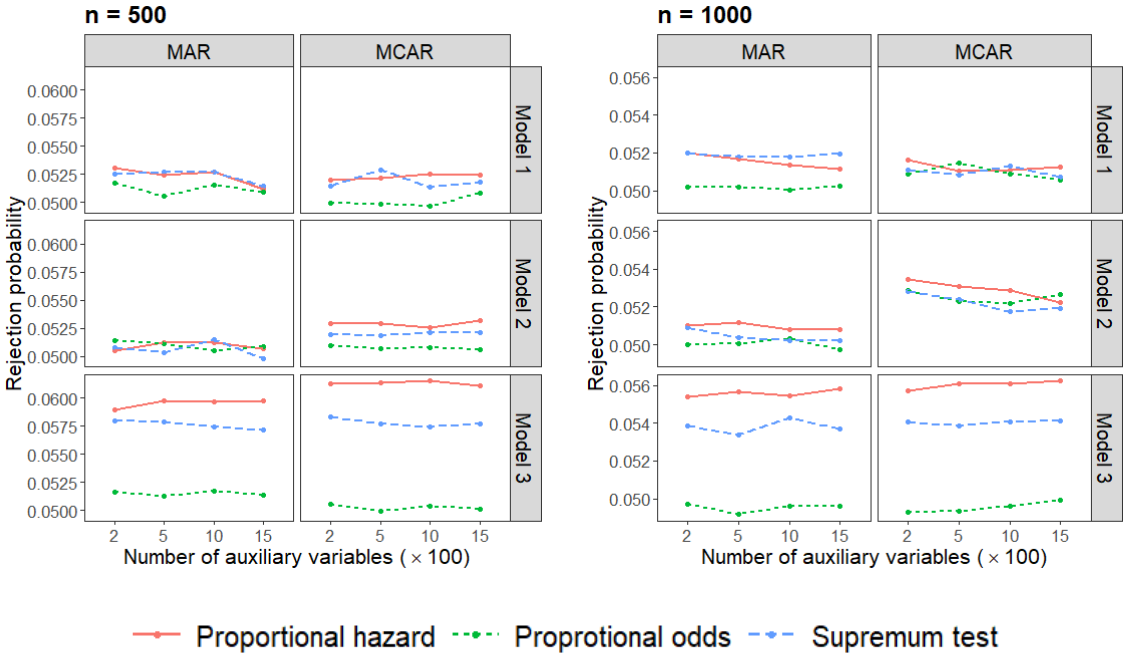


Figure 3.7: Study 2 - Rejection probabilities under a missing proportion of 30% and the null hypothesis.

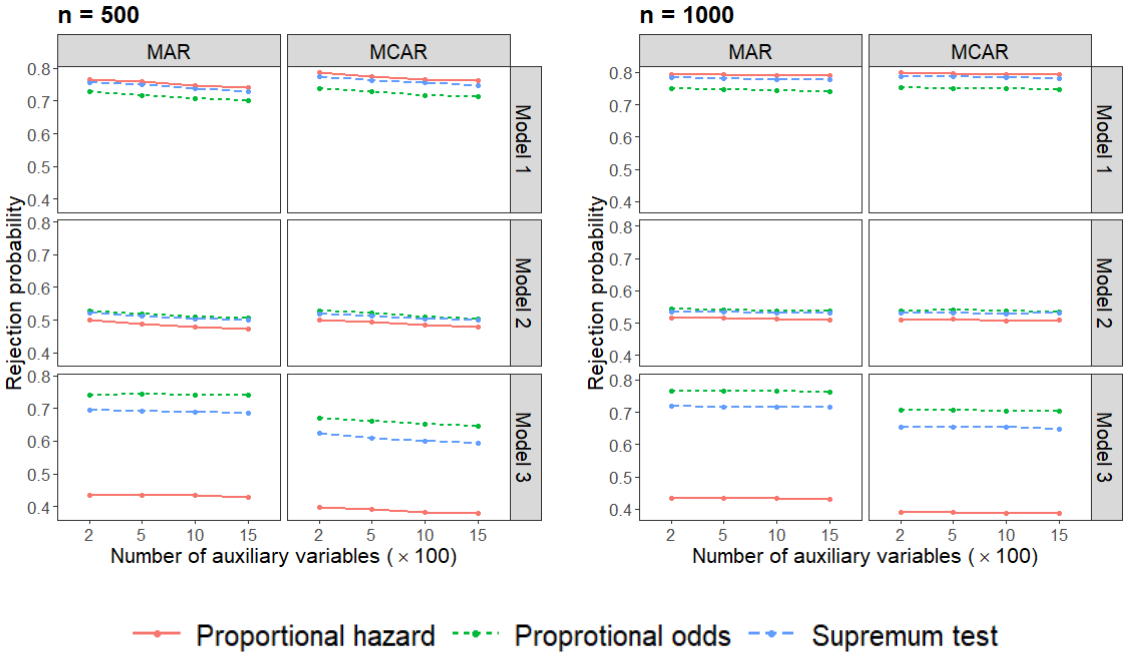


Figure 3.8: Study 2 - Rejection probabilities under a missing proportion of 30% and the alternative hypothesis.

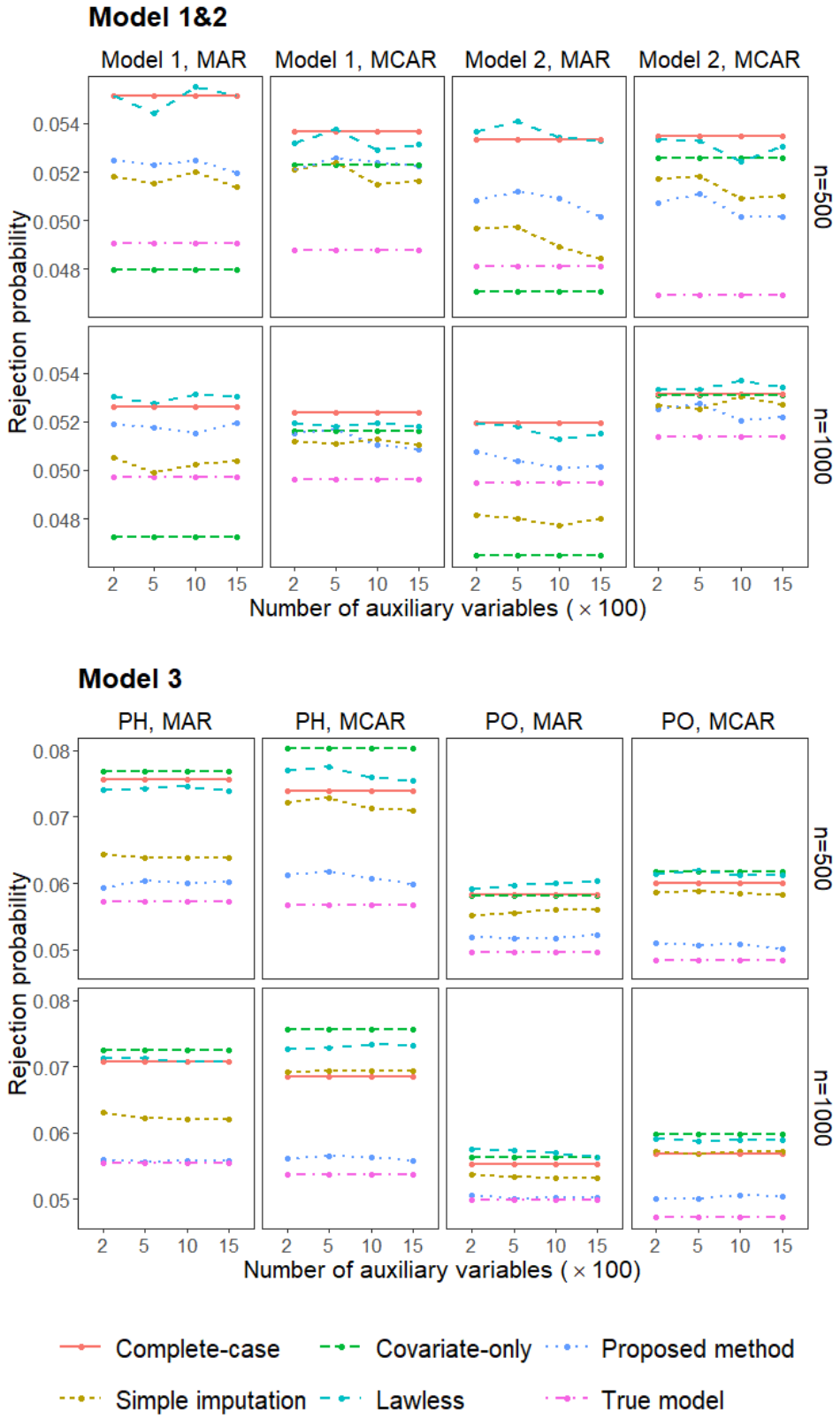


Figure 3.9: Study 3 - Rejection probabilities under a missing proportion of 30% and the null hypothesis.

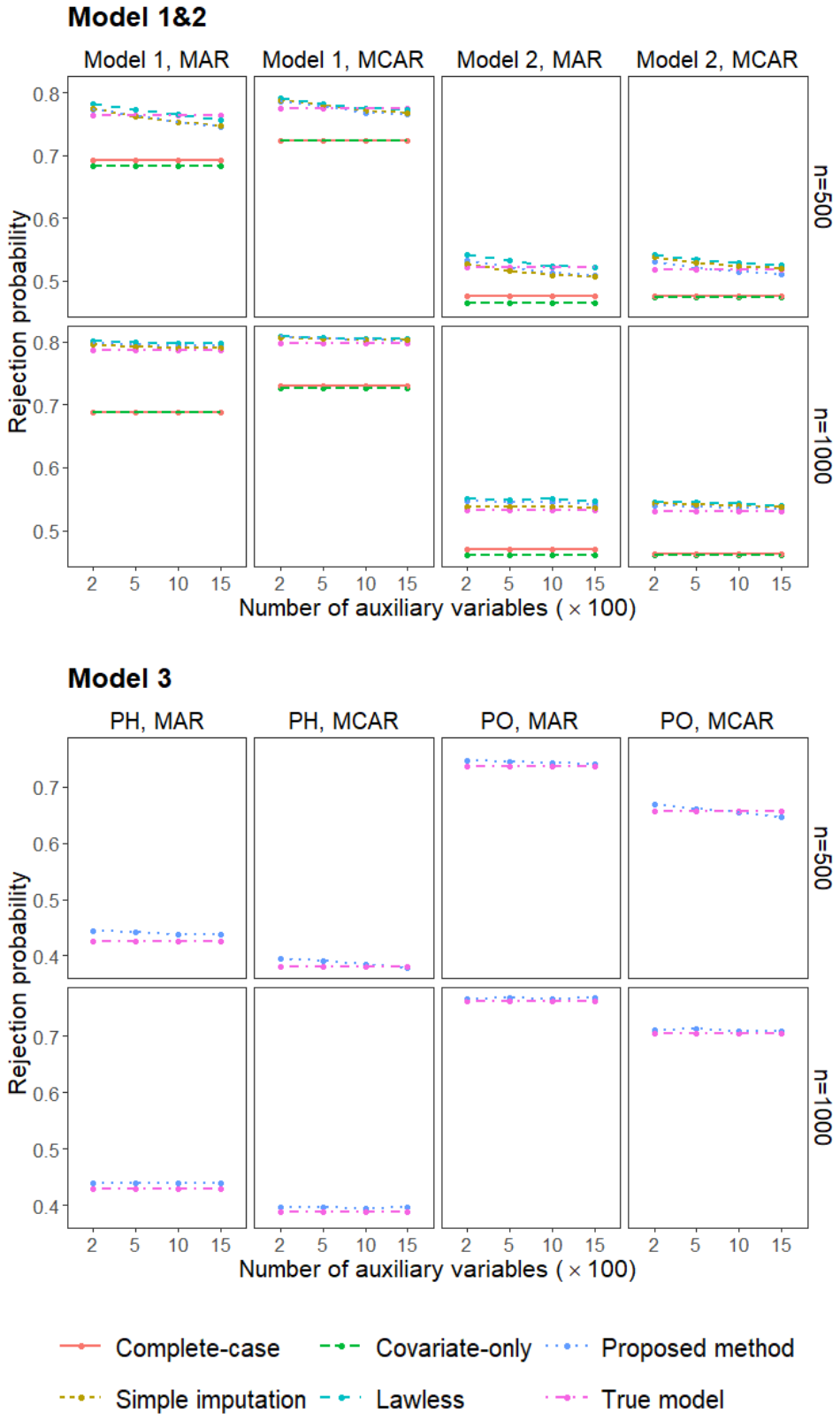


Figure 3.10: Study 3 - Rejection probabilities under a missing proportion of 30% and the alternative hypothesis.

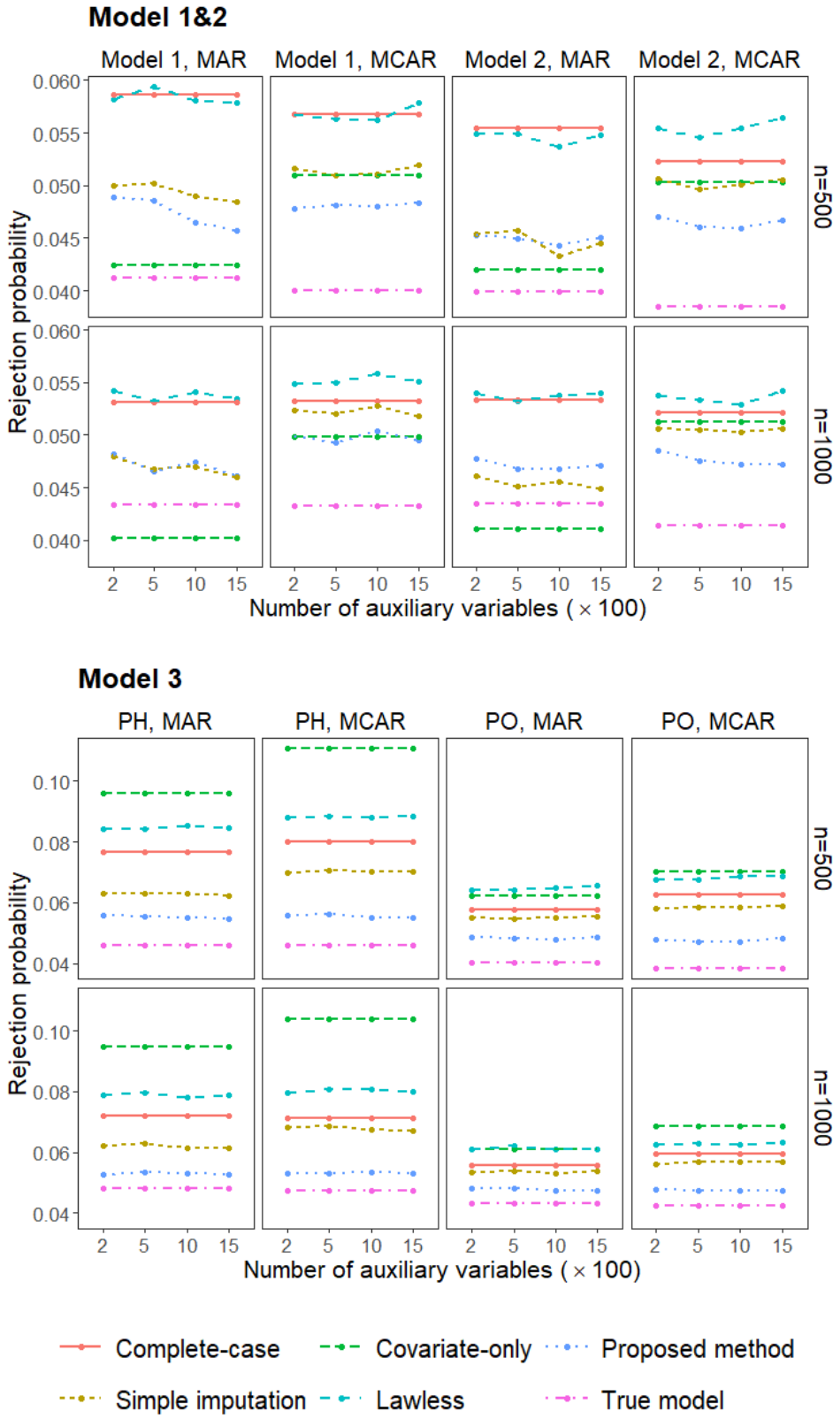


Figure 3.11: Study 3 - Rejection probabilities under a missing proportion of 60% and the null hypothesis.

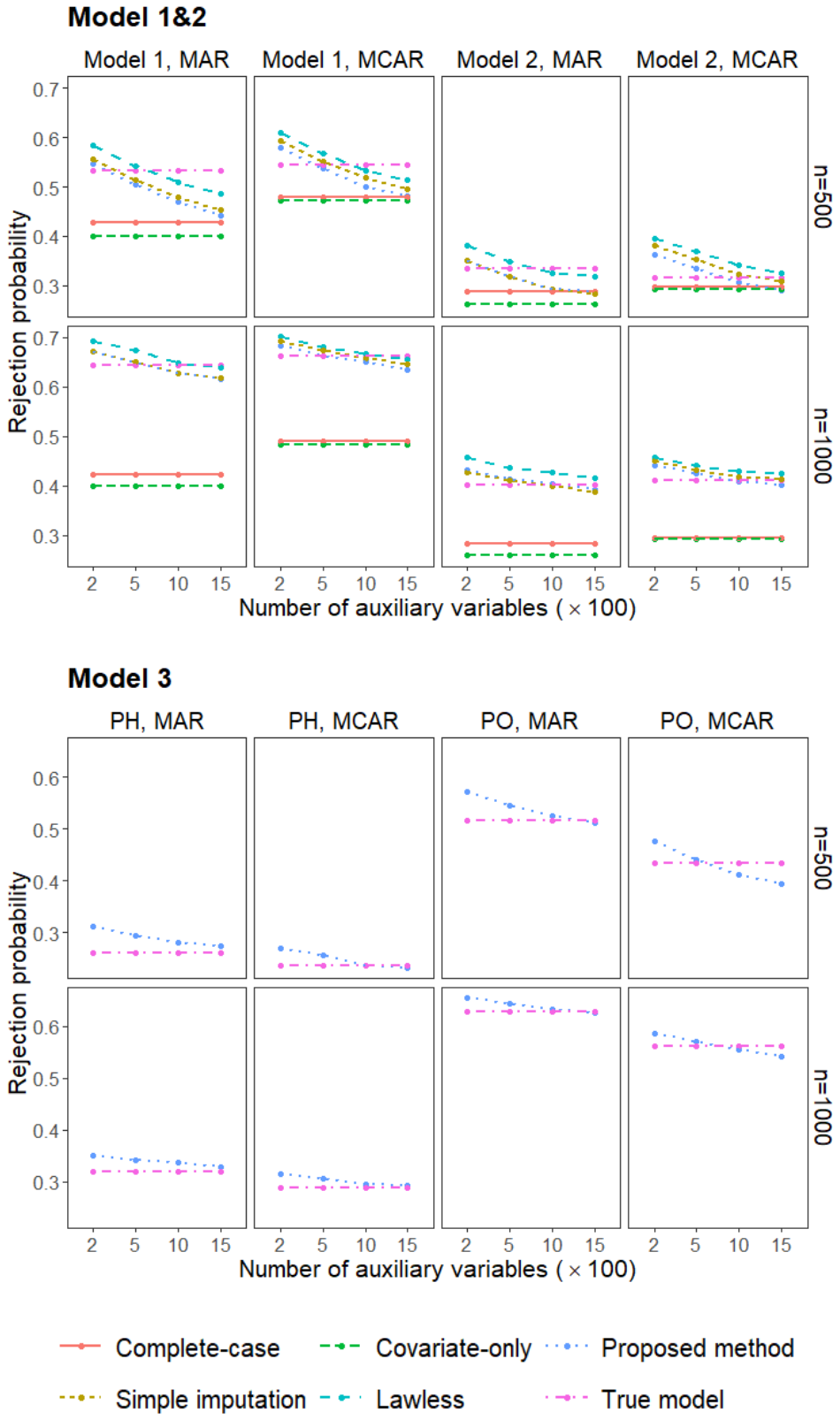


Figure 3.12: Study 3 - Rejection probabilities under a missing proportion of 60% and the alternative hypothesis.

Chapter 4

Conclusion

In this dissertation, we have studied the score tests to investigate the association between an outcome variable and an incomplete covariate with different types of outcomes. We propose a novel methodology to conduct association analysis of data with high-dimensional genomic variables, which is applicable to a wide range of cancer genomic studies such as TCGA and METABRIC. We provide theoretical and numerical results for parametric and semiparametric outcome models in Chapter 2 and Chapter 3, respectively.

In Chapter 2, we consider a parametric model for the outcome variable. Since the covariate of interest is subject to missing, we propose to select a subset of auxiliary variables and fit a regression model of the incomplete covariate against the selected variables. We then perform inference on the parameter of interest using the selected model based on the observed data. The proposed method is not restricted to a specific model selection procedure, and no assumptions on the correctness of the selected model are made. In fact, we show that the variability of model selection does not affect the asymptotic distribution of the test statistic. The proposed method presents better statistical performance in terms of efficiency by including the high-dimensional auxiliary variables into analysis. In the simulation studies, we show superior identification performance of the proposed method compared with several other methods. In the real data analysis of TCGA colorec-

tal adenocarcinoma, we find notable number of protein markers that have been reported in earlier studies. We also discover some potential candidates that are worth further investigation.

In Chapter 3, we consider a semiparametric transformation model for a right-censored survival outcome variable. We adopt the two-step test procedure proposed in Chapter 2 to capture the association between a time-to-event outcome and an incomplete covariate. We provide a flexible framework in which multiple transformation functions are taken into consideration. Specifically, we perform a single-model score test under each transformation function, and then combine the results to form a supremum score test to account for the uncertainty of the outcome model. We conduct extensive simulation studies and demonstrate the superiority of the proposed method over some existing approaches. We apply the proposed method to the TCGA data of bladder urothelial carcinoma and the METABRIC dataset, and identify important genomic signatures relevant with the time to tumor progression or death.

For further research, we can consider several directions. First, in both work, the covariate of interest S is one-dimensional. In many situations, however, we are interested in testing whether or not a group of covariates has effect on the outcome variable. One may adopt the variance component test to test for the effect of a covariate set. The advantage of the variance component test is that it takes the correlation among covariates into account.

Second, in our proposed framework, only a low-dimensional subset of auxiliary variables is used to impute the missing data, and the imputation model is fitted using least-squares estimation. It is of interest to consider a general imputation procedure that involves many auxiliary variables based on some regularized estimators, such as lasso, elastic net, and boosting. Such imputation procedures may be more accurate when many auxiliary variables are weakly associated with the incomplete covariate. The theoretical

development would be highly challenging, because the regularized estimators may not have closed-form expressions, and the dimension of the working model could be high.

Third, we have focused on hypothesis testing, and the theoretical results are developed under the null hypothesis. One may consider estimation and inference of the outcome model. In this case, the two-step procedure is invalid, because the missing mechanism would depend on S through its dependence with Y, and estimation of the model of S using only the subjects with observed data would be inconsistent. Also, one generally needs to account for the selection variability of the model of S using the methods of, for example, Taylor and Tibshirani (2018).

References

Agarwal, E., Brattain, M. G., & Chowdhury, S. (2013). Cell survival and metastasis regulation by Akt signaling in colorectal cancer. *Cellular Signalling*, 25, 1711–1719.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10, 1100–1120.

Bachoc, F., Leeb, H., & Pötscher, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics*, 47, 1475–1504.

Bachoc, F., Preinerstorfer, D., & Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48, 440–463.

Bair, E., Tibshirani, R., & Golub, T. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2, 511–522.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

Belaguli, N. S., Aftab, M., Rigi, M., Zhang, M., Albo, D., & Berger, D. H. (2010). GATA6 promotes colon cancer cell invasion by regulating urokinase plasminogen activator gene expression. *Neoplasia*, 12, 856–865.

Bennett, S. (1983a). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2, 273–277.

Bennett, S. (1983b). Log-logistic regression models for survival data. *Journal of the Royal Statistical Society: Series C*, 32, 165–171.

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41, 802–837.

Bertoni, F., Codegoni, A. M., Furlan, D., Tibiletti, M. G., Capella, C., & Broggini, M. (1999). CHK1 frameshift mutations in genetically unstable colorectal and endometrial cancers. *Genes, Chromosomes and Cancer*, 26, 176–180.

Bjørnland, T., Bye, A., Ryeng, E., Wisløff, U., & Langaas, M. (2018). Powerful extreme phenotype sampling designs and score tests for genetic association studies. *Statistics in Medicine*, 37, 4234–4251.

Breslow, N. (1972). Discussion of paper of D. R. Cox. *Journal of the Royal Statistical Society: Series B*, 34, 216–217.

Breslow, N., McNeney, B., & Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics*, 31, 1110–1139.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19, 1212–1242.

Bussey, K. J., Chin, K., Lababidi, S., Reimers, M., Reinhold, W. C., Kuo, W.-L., Gwadry, F., Kouros-Mehr, H., Fridlyand, J., Jain, A., et al. (2006). Integrating data on dna copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Molecular Cancer Therapeutics*, 5, 853–867.

Chatterjee, N., Chen, Y.-H., & Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98, 158–168.

Chen, K., Jin, Z., & Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89, 659–668.

Cheng, S., Wei, L., & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82, 835–845.

Chung, K. L. (2001). *A Course in Probability Theory* (3rd ed.) Academic Press.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34, 187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486, 346–352.

Dabrowska, D. M., & Doksum, K. A. (1988). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*, 83, 744–749.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39, 1–38.

Derkach, A., Lawless, J. F., & Sun, L. (2015). Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, 102, 988–994.

Diab-Assaf, M., Abou-Khouzam, R., Saadallah-Zeidan, N., Habib, K., Bitar, N., Karam, W., Liagre, B., Harakeh, S., & Azar, R. (2015). Expression of eukaryotic initiation factor 4E and 4E binding protein 1 in colorectal carcinogenesis. *International Journal of Clinical and Experimental Pathology*, 8, 404–413.

Ding, C., Luo, J., Yu, W., Gao, S., Yang, L., Chen, C., & Feng, J. (2015). Gab2 is a novel prognostic factor for colorectal cancer patients. *International Journal of Clinical and Experimental Pathology*, 8, 2779–2786.

Erisman, M., Scott, J., Watt, R., & Astrin, S. (1988). The c-Myc protein is constitutively expressed at elevated levels in colorectal carcinoma cell lines. *Oncogene*, 2, 367–378.

Esposito, D. L., Aru, F., Lattanzio, R., Morgano, A., Abbondanza, M., Malekzadeh, R., Bishehsari, F., Valanzano, R., Russo, A., & Piantelli, M. (2012). The insulin receptor substrate 1 (IRS1) in intestinal epithelial differentiation and in colorectal cancer. *PLoS ONE*, 7, 36190–36203.

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

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70, 849–911.

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38, 3567–3604.

Fithian, W., Sun, D., & Taylor, J. (2014). Optimal inference after model selection [Available at arXiv:1410.2597]. *Preprint*,

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109–135.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7, 397–416.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, 77, 270–278.

Gan, W., Zhao, H., Li, T., Liu, K., & Huang, J. (2017). CDK1 interacts with iASPP to regulate colorectal cancer cell proliferation through p53 pathway. *Oncotarget*, 8, 71618–71629.

Gao, Y., Shi, Q., Xu, S., Du, C., Liang, L., Wu, K., Wang, K., Wang, X., Chang, L. S., He, D., et al. (2014). Curcumin promotes KLF5 proteasome degradation through

downregulating YAP/TAZ in bladder cancer cells. *International Journal of Molecular Sciences*, 15, 15173–15187.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, 10, 80–100.

Hague, A., Moorghen, M., Hicks, D., Chapman, M., & Paraskeva, C. (1994). BCL-2 expression in human colorectal adenomas and carcinomas. *Oncogene*, 9, 3367–3370.

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Medical Research Methodology*, 12, 184–196.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361–387.

He, T.-Y., Tsai, L.-H., Huang, C.-C., Chou, M.-C., & Lee, H. (2014). LKB1 loss at transcriptional level promotes tumor malignancy and poor patient outcomes in colorectal cancer. *Annals of Surgical Oncology*, 21, 703–710.

Higgins, J. P., Kaygusuz, G., Wang, L., Montgomery, K., Mason, V., Zhu, S. X., Marinelli, R. J., Presti Jr, J. C., van de Rijn, M., & Brooks, J. D. (2007). Placental S100 (S100P) and GATA3: markers for transitional epithelium and urothelial carcinoma discovered by complementary DNA microarray. *The American Journal of Surgical Pathology*, 31, 673–680.

Hu, Y.-J., Li, Y., Auer, P. L., & Lin, D. Y. (2015). Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations. *Proceedings of the National Academy of Sciences*, 112, 1019–1024.

Huang, J., Ma, S., & Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18, 1603–1618.

Huang, M.-Y., Tsai, H.-L., Lin, C.-H., Huang, C.-W., Ma, C.-J., Huang, C.-M., Chai, C.-Y., & Wang, J.-Y. (2013). Predictive value of ERCC1, ERCC2, and XRCC1 overexpression for stage III colorectal cancer patients receiving FOLFOX-4 adjuvant chemotherapy. *Journal of Surgical Oncology*, 108, 457–464.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85, 765–769.

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics*, 55, 591–596.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, 100, 332–346.

Javanmard, A., & Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15, 2869–2909.

Javanmard, A., & Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory. *IEEE Transactions on Information Theory*, 60, 6522–6554.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 523–539.

Kim, J. K., & Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106, 157–165.

Kim, Y.-W., Koul, D., Kim, S. H., Lucio-Eterovic, A. K., Freire, P. R., Yao, J., Wang, J., Almeida, J. S., Aldape, K., & Yung, W. A. (2013). Identification of prognostic gene

signatures of glioblastoma: a study based on TCGA data analysis. *Neuro-oncology*, 15, 829–839.

Kountourakis, P., Pavlakis, K., Psyrri, A., Rontogianni, D., Xiros, N., Patsouris, E., Pectasides, D., & Economopoulos, T. (2006). Prognostic significance of HER3 and HER4 protein expression in colorectal adenocarcinomas. *BMC Cancer*, 6, 46–54.

Kristensen, V. N., Vaske, C. J., Ursini-Siegel, J., Van Loo, P., Nordgard, S. H., Sachidanandam, R., Sørlie, T., Wärnberg, F., Haakensen, V. D., Helland, Å., et al. (2012). Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling. *Proceedings of the National Academy of Sciences*, 109, 2802–2807.

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I., & Zhao, L. H. (2020). Valid post-selection inference in model-free linear regression. *The Annals of Statistics*, 48, 2953–2981.

Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., & Zhao, L. (2018). A model free perspective for linear regression: uniform-in-model bounds for post selection inference [Available at arXiv:1802.05801]. *Preprint*,

Kuchibhotla, A. K., & Chakrabortty, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: applications in covariance estimation and linear regression [Available at arXiv:1804.02605]. *Preprint*,

Lawless, J. (2018). Two-phase outcome-dependent studies for failure times and testing for effects of expensive covariates. *Lifetime Data Analysis*, 24, 28–44.

Lawless, J., Kalbfleisch, J., & Wild, C. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B*, 61, 413–438.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44, 907–927.

Li, G., Peng, H., Zhang, J., & Zhu, L. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, 40, 1846–1877.

Li, M., Li, R., & Ma, Y. (2021). Inference in high dimensional linear measurement error models. *Journal of Multivariate Analysis*, 184, 104759–104775.

Li, R., Zhong, W., & Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107, 1129–1139.

Lipsitz, S. R., & Ibrahim, J. G. (1996). Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2, 5–14.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.) John Wiley & Sons.

Little, R. J., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72, 497–512.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34, 1436–1462.

Metzeler, K. H., Hummel, M., Bloomfield, C. D., Spiekermann, K., Braess, J., Sauerland, M.-C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S. P., et al. (2008). An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*, 112, 4193–4201.

Murphy, S., Rossini, A., & van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92, 968–976.

Nam, K. T., Lee, H.-J., Smith, J. J., Lapierre, L. A., Kamath, V. P., Chen, X., Aronow, B. J., Yeatman, T. J., Bhartur, S. G., Calhoun, B. C., Condie, B., Manley, N. R., Beauchamp, R. D., Coffey, R. J., & Goldenring, J. R. (2010). Loss of Rab25 promotes the development of intestinal neoplasia in mice and is associated with human colorectal adenocarcinomas. *The Journal of Clinical Investigation*, 120, 840–849.

Ning, Y., & Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45, 158–195.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society: Series B*, 44, 234–243.

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Børresen-Dale, A.-L., & Brown, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99, 12963–12968.

Pollard, D. (1990). *Empirical Processes: Theory and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics* (Vol. 2). Hayward, CA: IMS.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–411.

Pryczynicz, A., Gryko, M., Niewiarowska, K., Cepowicz, D., Ustymowicz, M., Kemona, A., & Guzińska-Ustymowicz, K. (2014). Bax protein may influence the invasion of colorectal cancer. *World Journal of Gastroenterology*, 20, 1305–1310.

Qi, F., Yuan, Y., Zhi, X., Huang, Q., Chen, Y., Zhuang, W., Zhang, D., Teng, B., Kong, X., & Zhang, Y. (2015). Synergistic effects of AKAP95, Cyclin D1, Cyclin E1, and Cx43 in the development of rectal cancer. *International Journal of Clinical and Experimental Pathology*, 8, 1666–1673.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106–121.

Rodrigues, N. R., Rowan, A., Smith, M., Kerr, I. B., Bodmer, W. F., Gannon, J. V., & Lane, D. P. (1990). P53 mutations in colorectal cancer. *Proceedings of the National Academy of Sciences*, 87, 7555–7559.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, 346, 1937–1947.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W. C., Connors, J. M., Campo, E., Gascoyne, R. D., Grogan, T. M., Muller-Hermelink, H. K., Smeland, E. B., et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, 3, 185–197.

Rotnitzky, A., & Robins, J. M. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82, 805–820.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1, 20–34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25, 2906–2912.

Tang, Z., Yuan, S., Hu, Y., Zhang, H., Wu, W., Zeng, Z., Yang, J., Yun, J., Xu, R., & Huang, P. (2012). Over-expression of GAPDH in human colorectal carcinoma as a pre-

ferred target of 3-bromopyruvate propyl ester. *Journal of Bioenergetics and Biomembranes*, 44, 117–125.

Taylor, J., & Tibshirani, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics*, 46, 41–61.

The Cancer Genome Atlas Network. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487, 330–337.

The Cancer Genome Atlas Network. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507, 315–322.

Thyagarajan, A., Jedinak, A., Nguyen, H., Terry, C., Baldridge, L. A., Jiang, J., & Sliva, D. (2010). Triterpenes from ganoderma lucidum induce autophagy in colon cancer through the inhibition of p38 mitogen-activated kinase (p38 MAPK). *Nutrition and Cancer*, 62, 630–640.

Tian, X., Loftus, J. R., & Taylor, J. E. (2018). Selective inference with unknown variance via the square-root lasso. *Biometrika*, 105, 755–768.

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

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111, 600–620.

Tsofack, S. P., Garand, C., Sereduk, C., Chow, D., Aziz, M., Guay, D., Yin, H. H., & Lebel, M. (2011). NONO and RALY proteins are required for YB-1 oxaliplatin induced resistance in colon adenocarcinoma cell lines. *Molecular Cancer*, 10, 145–162.

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42, 1166–1202.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer.

Wang, W., Wang, M., Xu, J., Long, F., & Zhan, X. (2020). Overexpressed GATA3 enhances the sensitivity of colorectal cancer cells to oxaliplatin through regulating MiR-29b. *Cancer Cell International*, 20, 339–354.

Wei, X.-L., Wang, D.-S., Xi, S.-Y., Wu, W.-J., Chen, D.-L., Zeng, Z.-L., Wang, R.-Y., Huang, Y.-X., Jin, Y., Wang, F., Qiu, M.-Z., Luo, H.-Y., Zhang, D.-S., & Xu, R.-H. (2014). Clinicopathologic and prognostic relevance of ARID1A protein loss in colorectal cancer. *World Journal of Gastroenterology*, 20, 18404–18412.

Wong, K. Y., Fan, C., Tanioka, M., Parker, J. S., Nobel, A. B., Zeng, D., Lin, D. Y., & Perou, C. M. (2019a). I-Boost: an integrative boosting approach for predicting survival time with multiple genomics platforms. *Genome Biology*, 20, 52–66.

Wong, K. Y., Zeng, D., & Lin, D. Y. (2019b). Robust score tests with missing data in genomics studies. *Journal of the American Statistical Association*, 114, 1778–1786.

Xu, W., Anwaier, A., Ma, C., Liu, W., Tian, X., Palihati, M., Hu, X., Qu, Y., Zhang, H., & Ye, D. (2021). Multi-omics reveals novel prognostic implication of SRC protein expression in bladder cancer and its correlation with immunotherapy response. *Annals of Medicine*, 53, 596–610.

Yamamoto, H., Soh, J.-W., Monden, T., Klein, M. G., Zhang, L. M., Shirin, H., Arber, N., Tomita, N., Schieren, I., Stein, C., & Weinstein, I. B. (1999). Paradoxical increase in retinoblastoma protein in colorectal carcinomas may protect cells from apoptosis. *Clinical Cancer Research*, 5, 1805–1815.

Yang, L., Ding, C., Tang, W., Yang, T., Liu, M., Wu, H., Wen, K., Yao, X., Feng, J., & Luo, J. (2020). INPP4B exerts a dual function in the stemness of colorectal cancer stem-like cells through regulating Sox2 and Nanog expression. *Carcinogenesis*, 41, 78–90.

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

Zeng, D., & Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B*, 69, 507–564.

Zeng, D., Lin, D. Y., & Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statistica Sinica*, 18, 355–377.

Zeng, D., Mao, L., & Lin, D. Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103, 253–271.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38, 894–942.

Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76, 217–242.

Zhang, Z., & Rockette, H. E. (2005). On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, 134, 206–223.

Zhang, Z., & Rockette, H. E. (2006). Semiparametric maximum likelihood for missing covariates in parametric regression. *Annals of the Institute of Statistical Mathematics*, 58, 687–706.

Zhao, C., Du, S., Dang, X., & Gong, M. (2015). Expression of Paxillin is correlated with clinical prognosis in colorectal cancer patients. *Medical Science Monitor*, 21, 1989–1995.

Zhao, L., & Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*, 11, 769–782.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7, 2541–2563.

Zhao, Y., Lawless, J. F., & McLeish, D. L. (2009). Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal*, 51, 123–136.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.