# MISSING COVARIATES AND HIGH-DIMENSIONAL VARIABLE SELECTION IN ADDITIVE HAZARDS REGRESSION

by

Wei Lin

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(APPLIED MATHEMATICS)

August 2011

Copyright 2011 Wei Lin

# Dedication

*To my mom, dad, and sister*

# Acknowledgments

It was almost impossible to imagine, when I started my Ph.D. study here at USC, that I would ever complete a dissertation in statistics. As Columbus set out for the East, and discovered a new continent, I intended to study some lively topics in applied mathematics, and ran into the field of statistics. My discoveries in this new field, if any, are of course very humble, but I am fortunate that I have had a great advisor, Professor Larry Goldstein, without whom I would not have learned these interesting topics. I am so grateful to him for his encouragement, patience, and insight that have brought me so far.

I owe many thanks to Professor Jinchi Lv; the second part of this dissertation has largely grown out of our discussions, and I have been deeply influenced by his enthusiasm for the subject. Special acknowledgment goes to Professor Bryan Langholz, who provided invaluable advice at an earlier stage of my research which has helped to shape the first part of this dissertation. I would like to thank Professor Kenneth Alexander and Professor Fengzhu Sun for serving in my Dissertation Committee, and thank Professor Jay Bartroff for serving in my Guidance Committee. I also wish to express my gratitude to Professor Igor Kukavica, Dr. Florence Lin, Professor Remigijus Mikulevicius, and Professor Alan Schumitzky for their generous help.

I am definitely in debt to my mom, dad, and sister, whose tireless love and support over the years have allowed me the greatest freedom to pursue my dreams. Thanks are also due to my classmates and friends, who have made my study and life enjoyable; I apologize for not including a list here to avoid any unintentional omission.

# Table of Contents

# List of Tables

# List of Figures

# Abstract

This dissertation addresses two challenging problems arising in inference with censored failure time data. The additive hazards model provides a unified framework for these problems to be discussed, not only because it is a useful alternative to the well-known Cox model and has significant practical implications, but also because its simple yet elegant structure allows one to explore some fundamental aspects of these problems.

In the first part of this dissertation, we consider the estimation problem in additive hazards regression with missing covariates. We are interested in both the case where the observation probabilities are known and the case where they are unknown but can be parametrically modeled and estimated. By modifying the pseudoscore function with full data, we introduce some weighted estimators for the regression coefficients and the cumulative baseline hazard function. The proposed estimators are then shown to be consistent and asymptotically normal under mild conditions, with asymptotic variances that can be easily estimated. Our theoretical results and simulation studies indicate that using estimated weights in the simple weighted estimators may yield important efficiency gain and that the augmented weighted estimators are even more efficient. The proposed methods are further illustrated by a mouse leukemia data example.

In the second part, we turn to the variable selection problem in the additive hazards model. Motivated by linking high-throughput genomic data to survival outcomes, we are particularly interested in the high-dimensional setting where the dimension of covariates may grow fast, possibly nonpolynomially, with the sample size. We propose to perform variable selection and estimation simultaneously by using a class of regularized estimators with a general family of concave penalties, including several popular choices such as the lasso, SCAD, MCP, and SICA. In a nonasymptotic framework where the model dimensions are allowed to vary freely, we rigorously investigate the weak oracle properties and oracle properties of the proposed estimators. Our theoretical results are essentially different from those in the existing literature, and provide new insight into the model selection properties of regularized estimators for survival models. We illustrate the proposed method by simulation studies and application to a diffuse large B-cell lymphoma data set.

A common theme underlying the theoretical development in this dissertation is the use of modern empirical process theory. Indeed, we rely on the language of empirical process theory to establish our theoretical results for both problems considered here, and they serve as excellent examples for demonstration of the power and elegance of this mathematical tool, especially in the context of survival analysis.

# CHAPTER 1
# Introduction

## 1.1 Failure Time Regression Models

In the analysis of censored failure time data, the most convenient way of regression modeling is through the *hazard function*, which is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

where $T$ is a failure time of interest. Many failure time regression models have appeared since the seminar work of Cox (1972). Although these models differ in many major ways and often require different inference procedures, the conditional hazard function in most of these models can be described by the very general form

$$\lambda(t \mid \mathbf{Z}) = G(t, \lambda_0, \mathbf{B}^T \mathbf{Z}), \tag{1.1}$$

where $\mathbf{Z}(\cdot)$ is a $p$-vector of possibly time-dependent covariates, $\lambda_0(\cdot)$ is an unspecified baseline hazard function in the sense that $G(t, \lambda_0, \mathbf{0}) \equiv \lambda_0(t)$, $\mathbf{B}$ is a $q \times p$ matrix of regression coefficients, and $G(\cdot, \cdot, \cdot)$ is a known functional. Such models are intrinsically semiparametric in that $\lambda_0(\cdot)$ is an unknown infinite-dimensional parameter. The rationale behind such semiparametric models is that it is often believed that the hazard function depends not only on the covariates, but also on many other factors that are often unknown and not of main scientific interest. Introducing an unspecified baseline hazard function $\lambda_0(\cdot)$, which is viewed as a nuisance parameter, helps avoid too restrictive parametric assumptions and increases the flexibility of the regression models, while retaining the parsimony and interpretability of parametric models.

Model (1.1) realizes dimension reduction by extracting $q$ (often $q = 1$ or 2) linear combinations of the covariates, and allows a very general coupling structure between $\lambda_0(\cdot)$ and $\mathbf{B}^T\mathbf{Z}(\cdot)$. In the simplest form of this structure, the conditional hazard function depends only on the values of $\lambda_0(\cdot)$ and $\mathbf{B}^T\mathbf{Z}(\cdot)$ at time $t$, i.e.,

$$\lambda(t \mid \mathbf{Z}) = G(\lambda_0(t), \mathbf{B}^T\mathbf{Z}(t)). \tag{1.2}$$

This includes the following examples.

**Example 1.1 (The Cox model).** Cox (1972) introduced the model, which postulates a multiplicative effect of the covariates,

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{Z}(t)},$$

where $\boldsymbol{\beta}$ is a $p$-vector of regression coefficients. Cox's estimator for $\boldsymbol{\beta}$ is obtained by maximizing a partial likelihood (Cox, 1975), which is a special case of a semiparametric profile likelihood (Murphy and van der Vaart, 2000). Asymptotic properties of the partial likelihood estimator were derived by Andersen and Gill (1982) using counting process martingale theory, and it has been shown that it is fully efficient in the sense that it achieves the semiparametric information bound (Begun et al., 1983).

**Example 1.2 (Additive hazards model).** In contrast to the multiplicative structure of the Cox model, the additive hazards model (Cox and Oakes, 1984, p. 74; Breslow and Day, 1987, p. 182) assumes the additive structure

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}^T \mathbf{Z}(t). \tag{1.3}$$

Semiparametric estimation in this model was studied by Lin and Ying (1994). The idea of using an additive structure appeared earlier in Aalen (1980), where the regression coefficients are allowed to depend on time, resulting in a fully nonparametric model. An intermediate form between Aalen's model and (1.3) was considered by McKeague and Sasieni (1994). Inference methods for these variants, however, are quite different. Model (1.3) has the most parsimonious form among the three, and serves as the additive counterpart of the Cox model.

**Example 1.3 (Additive-multiplicative hazards model).** In order to accommodate multiplicative and additive effects of the covariates simultaneously, Lin and Ying (1995) introduced the model

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) e^{\boldsymbol{\beta}_1^T \mathbf{Z}_1(t)} + \boldsymbol{\beta}_2^T \mathbf{Z}_2(t),$$

which is in the form of (1.2) with

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\beta}_2 \end{pmatrix} \qquad \text{and} \qquad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix}.$$

The proposed estimators are based on a class of estimating functions, and adaptive procedures are required in order to achieve the semiparametric information bounds.

**Example 1.4 (Transformed hazards model).** Another effort to bridge the Cox model and the additive hazards model was made by Zeng, Yin and Ibrahim (2005), who introduced a class of transformed hazards models

$$\lambda(t \mid \mathbf{Z}) = G(\mu(t) + \boldsymbol{\beta}^T \mathbf{Z}(t)), \tag{1.4}$$

where $G(\cdot)$ is a known, increasing function and $\mu(t) \equiv G^{-1}(\lambda_0(t))$, with the Cox model and the additive hazards model corresponding to $G(x) = e^x$ and $G(x) = x$, respectively. The proposed sieve maximum likelihood estimator is appealing in that it achieves the semiparametric information bound, and thus, in the special case of the additive hazards model, provides a fully efficient, though computationally intensive, alternative to the pseudoscore estimator of Lin and Ying (1994).

Some failure time regression models are not covered by the form of (1.2), especially when the model is not initially derived based on the hazard function. A noteworthy example is the accelerated failure time model, which is very useful but will not be the focus of this dissertation.

In spite of the various forms of failure time regression models, the Cox model remains the most popular one and is overwhelmingly used in practice, partly because (1) Cox's partial likelihood estimator is both easily computable and theoretically optimal, whereas usually no such ideal solutions exist for other models; and (2) the Cox model has been firmly established in the research and practice of survival analysis, with well-developed theories and widely available implementations. There is no particular reason, however, to believe that the proportional hazards assumption in the Cox model is valid for real data, and the merits of some competitive models should not be overlooked. In particular, the additive hazards model enjoys many remarkable features that may not be shared by other failure time regression models. Here are some reasons why the additive hazards model is so useful; more can be found in Aalen, Borgan and Gjessing (2008, pp. 155–157).

- The linear form of (1.3) results from using a first-order Taylor expansion in $\lambda_0(t)$ and $\mathbf{Z}(t)$ to approximate the hazard function, and thus should be able, in many cases, to provide a nice approximation to the effects of covariates. As noted by Lin and Ying (1994), the test statistics for individual covariate effects are often very comparable between the additive hazards model and the Cox model, but the estimated survival functions can be quite different; also, when interaction effects are of interest, there can easily be a difference. Therefore, the additive hazards model can be used to describe some covariate effects that are not reflected by the Cox model as well as other models.

- In contrast to the Cox model which concerns the risk ratio, or *relative risk*, the additive hazards model pertains to the risk difference, or *excess risk*. This latter measure is particularly relevant and informative in public health and clinical studies, because it translates directly into the number of cases that would be avoided by eliminating an exposure or introducing a treatment. The additive hazards model, thus, can provide a more transparent measure in accessing the importance of covariates and is more convenient for use in comparison between groups with different baseline hazards.

- The linearity of the additive hazards model sometimes simplifies theoretical analysis and computations. For example, the pseudoscore estimator for model (1.3) processes a closed-form representation; and Aalen's additive hazards model, though fully nonparametric, does not require use of smoothing techniques such as kernels and splines as in nonparametric versions of the Cox model.

## 1.2 Overview of the Problems

As discussed in the previous section, the additive hazards model is appealing in many aspects and can be a very useful and practical alternative to the Cox model. In this dissertation we will address two problems related to the additive hazards model, which have not been fully explored in the literature. The first problem—missing covariates—have existed for quite a while in survival analysis, and several promising approaches have been available; most of the previous work, however, was confined to the Cox model. We will exhibit that some of these approaches can be extended to the additive hazards model and address some nontrivial issues that arise in these extensions. The second problem—high-dimensional variable selection—only emerged in the past decade, and has seen an explosion of research during the past few years. Due to the complexity of censored data and semiparametric regression models, most of the previous work, however, is empirical or computational, and lacks rigorous theoretical justification. We thus hope to provide a framework unifying and extending some of the existing work and establish theories toward a better understanding of the methodology. Below we give a brief overview of these two problems.

### 1.2.1 Missing Covariates

Missing data are commonly encountered in practice, and have attracted considerable interest since the seminal work of Rubin (1976) and the introduction of the EM algorithm (Dempster, Laird and Rubin, 1977). Broadly speaking, censored data can be viewed as a special case of missing data. In survival analysis, however, missing data usually refer to missingness in covariates or failure indicators. We will focus on the former, while the latter was also considered by a number of authors (e.g., Gijbels, Lin and Ying, 2007). Here are some reasons why missing covariates are prevalent in survival analysis.

- The study cohort typically involves hundreds to thousands of subjects, and it is often very expensive and time-consuming to collect all covariates for all subjects in the cohort. This situation is more severe if some of the covariates are difficult to obtain. Thus, sampling designs are routinely used in practice, so that some of the covariates are measured only on a random sample of subjects from the cohort.

- Many long-term studies last a few years to several decades, during which time the study plan and the means of data collection may change dramatically. For instance, if a new medical examination program is introduced midway through the study, it is usually difficult to collect data retrospectively, and the entire data set is deemed incomplete.

- In many studies the set of covariates is a mixture of complex types of data, some of which are more subject to missingness. For example, if a covariate is collected from questionnaires, it is quite possible that the answers provided by the subjects are incomplete; and if a covariate measurement involves much manual processing, missing data are more likely to occur due to human errors.

Covariates missing by design received the earliest attention, and methods were developed for widely used sampling designs such as the case-cohort design (Prentice, 1986; Self and Prentice, 1988) and the nested case-control design (Thomas, 1977; Goldstein and Langholz, 1992). The general problem of missing covariates in Cox regression has been approached from several directions since Lin and Ying (1993). Among the most popular methods are the nonparametric maximum likelihood (NPML) approach (Chen and Little, 1999; Chen, 2002) and the weighted estimating equations (WEE) approach (Pugh et al., 1993; Wang and Chen, 2001; Qi, Wang and Prentice, 2005; Xu et al., 2009; Luo, Tsai and Xu, 2009). The NPML approach is attractive in that it generally leads to a semiparametric efficient estimator, although its implementation via an EM-type algorithm may be computationally intensive. The WEE approach requires less computational effort and may be more robust against misspecification of auxiliary distributions than the NPML approach, but the resulting estimator is in general not fully efficient. Other approaches include the imputation methods of Paik and Tsai (1997) and Paik (1997), the likelihood-based methods of Lipsitz and Ibrahim (1998) and Herring and Ibrahim (2001), and the efficient score method of Nan (2004).

Much less is known about missing covariates beyond the Cox model. Under the additive hazards model, Kulich and Lin (2000a) studied the case-cohort design, where covariates are measured only on the failures and a random sample from the entire cohort, while Kulich and Lin (2000b) and Jiang and Zhou (2007) considered the measurement error problem, where some of the covariates are measured only on a random validation sample, with surrogate covariates available for all subjects. All of the aforementioned work, however, concerns special cases where covariates are missing by design, and thus requires the observation probabilities to be exactly known. Also, it is unclear whether and how the efficiency of the existing estimators can be further improved.

In Chapter 2 we will consider a general missing-data problem in the additive hazards model and developed some estimation procedures based on the method of weighted estimating equations. A more detailed introduction to our methods and results will be given in Section 2.1.

### 1.2.2 High-Dimensional Variable Selection

Advances in experimental technologies in molecular biology during the past decade have brought in a wealth of biomedical data. For example, DNA microarrays now can be used to measure the expression of tens of thousands of genes in a sample of cells or to identify hundreds of thousands of single nucleotide polymorphisms (SNPs) for an individual at the same time. Data of this kind pose a tremendous challenge to effective statistical inference, since while the number of features, $p$, is easily in the thousands or tens of thousands, the number of observations, $n$, is typically in the tens or hundreds. Many classical methods of inference can easily fail in such high-dimensional settings. Variable selection, therefore, is a fundamental task in high-dimensional regression problems, which aims to select only a small set of important predictors from a large number of features, in the hope of alleviating the overfitting problem in high dimensions and improving the interpretability of the resulting model. Fan and Lv (2010) gave an overview of statistical challenges and some recent developments on the problem of high-dimensional variable selection.

Since gene expression and SNPs play key roles in many biological processes and are potentially associated with human diseases such as cancer, it would be informative to link high-dimensional genomic data to survival outcomes. A number of efforts have recently been made in this direction. Regularization methods, which can yield sparse models and thus perform variable selection and estimation simultaneously, are particularly useful and have gained wide popularity. Several regularization methods originally developed for linear regression have been adapted to survival models. For example, Tibshirani (1997), Fan and Li (2002), and Zhang and Lu (2007) extended, respectively, the lasso (Tibshirani, 1996), nonconcave penalized likelihood (Fan and Li, 2001), and the adaptive lasso (Zou, 2006) to the Cox model, and the last two methods are shown to enjoy the oracle property (Fan and Li, 2001). These authors, however, considered only the classical setting where the dimension of covariates, $p$, is fixed. Cai et al. (2005) was the first to study the properties of penalized likelihood methods for survival models in an asymptotic framework with $p$ growing with $n$; they demonstrated the oracle property of nonconcave penalized pseudo-partial likelihood for multivariate failure time data and allowed $p = o(n^{1/4})$. The Dantzig selector (Candes and Tao, 2007), a method specifically developed for high-dimensional estimation problems, was adapted to the Cox model by Antoniadis, Fryzlewicz and Letué (2010); however, they did not address the question of model selection consistency.

Variable selection techniques have also been extended beyond the Cox model. For example, Leng and Ma (2007) proposed a modified lasso approach, and Martinussen and Scheike (2009) considered several regularization methods for the additive hazards model. Extensions of variable selection methods to the accelerated failure time model were studied by Huang, Ma and Xie (2006), Johnson (2008), Wang et al. (2008), and Cai, Huang and Tian (2009), among others. A nonparametric approach based on random survival forests was introduced by Ishwaran et al. (2010). Some other high-dimensional inference methods for survival data, including those that do not lead to sparsity, were surveyed by Witten and Tibshirani (2010).

Despite the aforementioned developments on high-dimensional variable selection for survival data, we note that, however, there is still a gap in the literature between the theory and practice of such methods; the properties of the existing methods in high-dimensional settings and under what conditions model selection consistency can be guaranteed remain largely unknown. The need for the development of a general, rigorous theory for regularization-based variable selection techniques in the survival setting is especially urgent, in view of the recent breakthroughs in establishing such theories for linear regression models (Zhao and Yu, 2006; Wainwright, 2009; Lv and Fan, 2009).

In Chapter 3 we will propose to perform variable selection and estimation simultaneously for survival data, by using a general class of regularized estimators which combine the nonconcave penalized likelihood approach and the pseudoscore method for the additive hazards model. The additive hazards model assumption here should not be considered too restrictive, for at least two reasons. First, in principle, by incorporating time-dependent covariates, many survival models can fit most real data reasonably well. Second, the test statistics for individual covariate effects are often very comparable between the additive hazards model and the Cox model. We will develop a general high-dimensional theory for the proposed estimators, which sheds light on the model selection properties of regularization methods for survival data. We will continue our discussion in Section 3.1.

## 1.3 Empirical Process Techniques

Empirical process techniques play a crucial role in establishing the theoretical results that will be needed in Chapters 2 and 3, and hence we now give a brief introduction. The laws of large numbers and central limit theorems are essential tools for the asymptotic analysis of many statistical procedures. The theory of empirical processes is a natural extension of these ideas to a "uniform" or "functional" sense, giving rise to results that are uniform in classes of functions. Such results are intrinsically needed in survival analysis as well as many other fields where semiparametric inference plays an important role. For instance, in the analysis of semiparametric hazards regression models, we typically need to obtain results that are uniform in time $t$ and sometimes also in the regression parameter $\boldsymbol{\beta}$. The seminal paper by Andersen and Gill (1982), which is often attributed to the power and elegance of martingale techniques, provided in its Appendix III an extension of the laws of large numbers to the Skorohod space $D[0, 1]$, which can viewed as a classical empirical process argument. The recent paper by Zeng and Lin (2007) strongly advocated the use of modern empirical process theory in survival analysis, especially in applications related to nonparametric maximum likelihood estimation.

### 1.3.1 Glivenko–Cantelli and Donsker Classes

We introduce some notation and definitions. Let $X_1, \ldots, X_n$ be a random sample from a distribution $P$ on a measurable space $(\mathcal{X}, \mathcal{A})$. Let $\mathbb{P}_n$ denote the *empirical measure*, i.e., the discrete uniform measure on the observations $X_1, \ldots, X_n$. Given a measurable function

$f\colon \mathcal{X} \to \mathbb{R}$, denote by $\mathbb{P}_n f$ the expectation of $f$ under $P_n$, and by $Pf$ the expectation of $f$ under $P$, i.e.,

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \qquad \text{and} \qquad Pf = \int f\, dP.$$

The *empirical process* $\mathbb{G}_n$ is defined as $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. Let $\|\cdot\|_{\mathcal{F}}$ denote the uniform norm over $\mathcal{F}$. A class $\mathcal{F}$ of measurable functions $f$ is called *$P$-Glivenko–Cantelli* if $\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf| \to 0$ almost surely, and is called *$P$-Donsker* if the sequence of processes $\{\mathbb{G}_n f\colon f \in \mathcal{F}\}$ converges weakly to a tight limiting process in $\ell^\infty(\mathcal{F})$, where, for any index set $T$, $\ell^\infty(T)$ is the space of bounded functions on $T$ equipped with the uniform norm.

Let $\|\cdot\|_{P,r}$ denote the usual $L_r(P)$-norm, and we call $F$ an *envelope* of the class $\mathcal{F}$ if $|f| \le F$ for all $f \in \mathcal{F}$. The asymptotic behavior of the class $\mathcal{F}$ depends on its “complexity”, which can be characterized in either the *bracketing number* $N_{[\,]}(\varepsilon, \mathcal{F}, L_r(P))$ or the *uniform covering number* $\sup_Q N(\varepsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q))$, where in the latter notation the supremum is taken over all probability measures $Q$ for which $\mathcal{F}$ is not identically zero. The logarithms of the bracketing number and the uniform covering number are called the *bracketing entropy* and the *uniform entropy*, respectively. We do not give here the precise definitions of these two complexity measures, for in most cases we will be able to draw results from the empirical process literature on the entropy of special classes of functions (notably, VC classes), rather than calculate the entropy from scratch.

To be used in the Donsker theorems, we define the *bracketing integral*

$$J_{[\,]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L_2(P))}\, d\varepsilon$$

and the *uniform entropy integral*

$$J(\delta, \mathcal{F}, L_2) = \int_0^\delta \sqrt{\log \sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, d\varepsilon.$$

We have the following Glivenko–Cantelli and Donsker theorems which provide simple conditions for a class of functions to be Glivenko–Cantelli or Donsker. These results can be found in several books on empirical processes; for example, they are Theorems 2.2, 2.3, 8.14, and 8.19 of Kosorok (2008), respectively.

**Theorem 1.1.** *Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$. Then $\mathcal{F}$ is $P$-Glivenko–Cantelli.*

**Theorem 1.2.** *Let $\mathcal{F}$ be a class of measurable functions such that $J_{[]}(1, \mathcal{F}, L_2(P)) < \infty$. Then $\mathcal{F}$ is $P$-Donsker.*

The Glivenko–Cantelli and Donsker theorems with uniform entropy conditions require some additional measurability assumptions.

**Theorem 1.3.** *Let $\mathcal{F}$ be a $P$-measurable class of measurable functions with envelope $F$ such that $\sup_Q N(\varepsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty$ for every $\varepsilon > 0$. If $PF < \infty$, then $\mathcal{F}$ is $P$-Glivenko–Cantelli.*

**Theorem 1.4.** *Let $\mathcal{F}$ be a $P$-measurable class of measurable functions with envelope $F$ such that $J(1, \mathcal{F}, L_2) < \infty$, and the classes $\mathcal{F}_\delta \equiv \{f - g\colon f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$, for all $\delta > 0$, and $\mathcal{F}_\infty^2 \equiv \{(f - g)^2\colon f, g \in \mathcal{F}\}$ are $P$-measurable. If $PF^2 < \infty$, then $\mathcal{F}$ is $P$-Donsker.*

We do not give the definition of $P$-measurability here, but it suffices in all the cases we will encounter to verify a stronger measurability: A class $\mathcal{F}$ of measurable functions is *pointwise measurable* if there exists a countable subset $\mathcal{G} \subset \mathcal{F}$ such that for every $f \in \mathcal{F}$, there exists a sequence of functions $g_m \in \mathcal{G}$ such that $g_m \to f$ pointwise.

More complicated Cantelli–Glivenko and Donsker classes can be built from simpler classes. For this purpose, a variety of Glivenko–Cantelli and Donsker preservation results are available. Also, when applying the Glivenko–Cantelli and Donsker theorems with uniform entropy conditions, the assumptions can be verified by building bounded uniform entropy integral (BUEI) classes and pointwise measurable (PM) classes. We will also use BUEI preservation results separately in Chapter 3. For a summary of such preservation results, we refer to Kosorok (2008, Chapter 9).

### 1.3.2 Random Functions

In Chapter 2 we will need to consider empirical processes indexed by a class of random functions. We now introduce some concepts and a key lemma that will be useful in such applications. Here a *random function* $x \mapsto \widehat{f}_n(x;\omega)$, for every fixed $x$, is a function defined on the same probability space as the observations $X_1(\omega), \ldots, X_n(\omega)$. This concept is useful in cases where the empirical process is indexed by a set of functions that depend on some quantities estimated from the entire sample. The notations $\mathbb{P}_n \widehat{f}_n$ and $P\widehat{f}_n$ are understood as the expectations of $\widehat{f}_n$ under $\mathbb{P}_n$ and $P$, respectively, with $\omega$ fixed.

The following lemma, which is Lemma 19.24 of van der Vaart (1998), will play a key role in establishing the asymptotic results in Chapter 2.

**Lemma 1.1.** *Suppose that $\mathcal{F}$ is a $P$-Donsker class of measurable functions and $\widehat{f}_n$ is a sequence of random functions taking values in $\mathcal{F}$ such that $\int \big(\widehat{f}_n(x) - f_0(x)\big)^2 \, dP(x) = o_p(1)$ for some $f_0 \in L_2(P)$. Then $\mathbb{G}_n(\widehat{f}_n - f_0) = o_p(1)$.*

Using this lemma allows us to bypass construction of an almost sure representation, for instance, by the Skorokhod–Dudley–Wichura theorem (Shorack and Wellner, 1986, p. 47). The latter approach has been taken by a few authors in dealing with problems similar to what we will consider in Chapter 2; its application, however, is often quite involved.

### 1.3.3 Maximal and Concentration Inequalities

In Chapter 3 we will need to bound tail probabilities for the supremum of an empirical process. We now introduce two fundamental tools for such applications. The idea is to first bound tail probabilities for the mean of the supremum by a *maximal inequality*, and then control the deviation of the supremum from its mean by a *concentration inequality*.

The following maximal inequalities for empirical processes lead to bounds in terms of the bracketing integral or the uniform entropy integral, and will be useful in Chapter 3. These are Corollary 19.35 and Lemma 19.38 of van der Vaart (1998), respectively. Here $\lesssim$ stands for "smaller than up to a universal constant."

**Lemma 1.2.** *Let $\mathcal{F}$ be a class of measurable functions with envelope $F$. Then*

$$E_P\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}(\|F\|_{P,2}, \mathcal{F}, L_2(P)).$$

**Lemma 1.3.** *Let $\mathcal{F}$ be a $P$-measurable class of measurable functions with envelope $F$. Then*

$$E_P\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J(1, \mathcal{F}, L_2)\|F\|_{P,2}.$$

In Chapter 3 we will also need the following functional Hoeffding-type inequality, which is Theorem 9 of Massart (2000) and generalizes an earlier result of Ledoux (1995).

**Lemma 1.4.** *Let $\mathcal{F}$ be a class of measurable functions $f\colon \mathcal{X} \to \mathbb{R}$ such that $a \le f \le b$ for every $f \in \mathcal{F}$. Denote $W_n = \|\mathbb{P}_n - P\|_{\mathcal{F}}$. Then, for every $x > 0$,*

$$\Pr(W_n \ge EW_n + x) \le \exp\biggl(-\frac{nx^2}{2(b-a)^2}\biggr).$$

We close this chapter with a notational remark. Since the random elements of our interest, for example, the supremum of an empirical process, may not be Borel-measurable, we need to replace expectation and probability by outer expectation and outer probability, respectively; the concepts of weak convergence, convergence in probability, and almost sure convergence can be accordingly defined. However, to avoid redundancy in notation, we do not add a star, as done by some authors, to stress that the concept in question is in the more general "outer" sense. For such measurability issues, we refer to van der Vaart (1998, Chapter 18) for a short introduction and van der Vaart and Wellner (1996) for a lengthy exposition.

CHAPTER 2

# Additive Hazards Regression with Missing Covariates

## 2.1 Introduction

In this chapter we consider a general missing data problem for the additive hazards model where some of the covariates are *missing at random* (MAR) in the sense of Rubin (1976); that is, the missingness depends on the observed data, which in our case include the censored failure time, failure indicator, and the always observed covariates, but not on the missing covariates. We consider both the case where the observation probabilities are known and the case where they are unknown but can be parametrically modeled and estimated. As discussed in Section 1.2.1, the nonparametric maximum likelihood (NPML) and weighted estimating equations (WEE) approaches are among the most promising methods for dealing with missing covariates in survival models. In the additive hazards model, however, since the inference is typically based on a pseudoscore function which is not derived from a likelihood, it is unclear how the NPML approach should be applied. Therefore, our developments will build on the WEE approach, which seems very natural in this context.

By modifying the pseudoscore function of Lin and Ying (1994) with full data, we introduce in Sections 2.2 and 2.3 two different weighted estimators for the regression coefficients and a weighted estimator for the cumulative baseline hazard function. Both true weights and estimated weights are considered. In establishing the asymptotic theory, since the weights may depend on the outcome variables and hence are not predictable, counting process martingale theory is not applicable. Therefore, we rely on modern empirical process theory to show that the proposed estimators are consistent and asymptotically normal under mild conditions, and the asymptotic variances can be easily estimated. We then conduct simulations in Section 2.4 to demonstrate the good performance of the proposed estimators. Both asymptotical results and simulation studies indicate that using estimated weights in the simple weighted estimators may yield important efficiency gain and that the augmented weighted estimators can be even more efficient. In the special case of the case-cohort design with Bernoulli sampling, the proposed estimators improve on the efficiency of those in Kulich and Lin (2000a). The proposed methods are further illustrated by a real data example in Section 2.5. We offer some discussion in Section 2.6, and the regularity conditions and proofs are deferred to Section 2.7.

## 2.2 Simple Weighted Estimators

### 2.2.1 Estimation of the Regression Coefficients

We begin with the problem setup. Let $T$ denote the failure time and $C$ the censoring time. We observe the censored failure time $X = T \wedge C$ and the failure indicator $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Let $\mathbf{Z}(\cdot)$ be a $p$-vector of predictable covariate processes and assume that $T$ and $C$ are conditionally independent given $\mathbf{Z}(\cdot)$. Let $\mathbf{Z}(\cdot)$ be partitioned as $\mathbf{Z}(\cdot) = (\mathbf{Z}_m^T(\cdot), \mathbf{Z}_o^T(\cdot))^T$, where $\mathbf{Z}_o(\cdot)$ is always observed and $\mathbf{Z}_m(\cdot)$ is possibly missing. Let $R = 1$ if $\mathbf{Z}_m(\cdot)$ is observed, and $R = 0$ otherwise. Assume that

$\mathbf{Z}_m(\cdot)$ is MAR so that the observation probability $\pi$ satisfies

$$\pi \equiv \Pr\{R = 1 \mid X, \Delta, \mathbf{Z}(\cdot)\} = \Pr\{R = 1 \mid X, \Delta, \mathbf{Z}_o(\cdot)\}.$$

The full data $(X_i, \Delta_i, \mathbf{Z}_i(\cdot), R_i)$, $i = 1, \ldots, n$, are independent copies of $(X, \Delta, \mathbf{Z}(\cdot), R)$, and we denote $\pi_i = \pi(X_i, \Delta_i, \mathbf{Z}_{oi}(\cdot))$. We consider the additive hazards model

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^T \mathbf{Z}(t), \tag{2.1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0$ is a $p$-vector of regression coefficients.

Adopting the usual counting process notation, we denote the counting process for the observed failure by $N_i(t) = I(X_i \le t, \Delta_i = 1)$, and the at-risk indicator by $Y_i(t) = I(X_i \ge t)$. For notational convenience, write $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$, and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$, for any vector $\mathbf{v}$. The definition of the simple weighted estimators involves a two-stage weighting procedure. First, in the pseudoscore function of Lin and Ying (1994) with full data, replace

$$\mathbf{S}^{(k)}(t) = n^{-1} \sum_{j=1}^{n} Y_j(t) \mathbf{Z}_j(t)^{\otimes k}, \qquad k = 0, 1,$$

by their inverse probability weighted counterparts

$$\mathbf{S}_{\mathrm{SW}}^{(k)}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{R_j}{\pi_j} Y_j(t) \mathbf{Z}_j(t)^{\otimes k}, \qquad k = 0, 1. \tag{2.2}$$

Second, weight each term in the pseudoscore function by $R_i/\pi_i$. We then obtain the simple weighted pseudoscore function

$$\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_{\mathrm{SW}}(t)\}\{dN_i(t) - Y_i(t) \boldsymbol{\beta}^T \mathbf{Z}_i(t)\, dt\}, \tag{2.3}$$

where

$$\overline{\mathbf{Z}}_{\mathrm{SW}}(t) = \mathbf{S}^{(1)}_{\mathrm{SW}}(t)/S^{(0)}_{\mathrm{SW}}(t)$$

and $\tau$ is the maximum follow-up time.

We first consider the case where the observation probabilities $\pi_i$ are known and the true weights are used in the above weighting procedure. In this case, we can directly define the simple weighted estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$ as the unique solution to the estimating equation $\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}) = 0$, which has the closed form

$$\widehat{\boldsymbol{\beta}}_{\mathrm{SW}} = \left[\sum_{i=1}^{n} \frac{R_i}{\pi_i} \int_0^{\tau} \{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\mathrm{SW}}(t)\}^{\otimes 2} Y_i(t)\, dt\right]^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \int_0^{\tau} \{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\mathrm{SW}}(t)\}\, dN_i(t). \tag{2.4}$$

To present the asymptotic properties of $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$, we need more notation. Define

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\{\lambda_0(s) + \boldsymbol{\beta}_0^T \mathbf{Z}_i(s)\}\, ds,$$

which is the counting process martingale. Moreover, define

$$\begin{aligned}
\mathbf{s}^{(k)}(t) &= E\{Y(t)\mathbf{Z}(t)^{\otimes k}\}, \qquad k = 0, 1,\\
\mathbf{e}(t) &= \mathbf{s}^{(1)}(t)/s^{(0)}(t),\\
\mathbf{M}_{\mathbf{Z}} &= \int_0^{\tau} \{\mathbf{Z}(t) - \mathbf{e}(t)\}\, dM(t),\\
\mathbf{D} &= E\left[\int_0^{\tau} Y(t)\{\mathbf{Z}(t) - \mathbf{e}(t)\}^{\otimes 2}\, dt\right],
\end{aligned}$$

and

$$\boldsymbol{\Sigma} = E(\mathbf{M}_{\mathbf{Z}}^{\otimes 2}) = E\left[\int_0^{\tau} \{\mathbf{Z}(t) - \mathbf{e}(t)\}^{\otimes 2}\, dN(t)\right],$$

where $N(\cdot)$, $Y(\cdot)$, and $M(\cdot)$ are the generic processes of $N_i(\cdot)$, $Y_i(\cdot)$, and $M_i(\cdot)$, respectively.

Note that in deriving the asymptotic properties of $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$, the familiar martingale central limit theorem does not apply, since the observation probabilities $\pi_i$ may depend on the outcome variables $X_i$ and $\Delta_i$, and hence are not predictable. Thus, we will rely on empirical process theory to establish the needed asymptotic results. The asymptotic properties of $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$ are given by the following theorem.

**Theorem 2.1.** *Under Conditions 2.1 and 2.2 in Section 2.7,* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{SW}} - \boldsymbol{\beta}_0)$ *is asymptotically normal with mean zero and covariance matrix* $\mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathrm{SW}}\mathbf{D}^{-1}$*, where*

$$\boldsymbol{\Sigma}_{\mathrm{SW}} = E(\pi^{-1}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}) = \boldsymbol{\Sigma} + E\{(1-\pi)\pi^{-1}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}\}.$$

Theorem 2.1 shows that the missing data add the extra amount of variability $\mathbf{D}^{-1}E\{(1-\pi)\pi^{-1}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}\}\mathbf{D}^{-1}$ to the full-data covariance matrix $\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$. In the special case where data are missing by the case-cohort design with Bernoulli sampling, $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$ reduces to the estimator of Kulich and Lin (2000a), with $\pi_i = \Delta_i + (1-\Delta_i)p_i$, where $p_i$ is the probability that the $i$th subject is selected into the subcohort.

It is clear that $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$, though very simple to implement if $\pi_i$ are known, only uses data of the complete cases and hence can be very inefficient. We will see below that efficiency can be substantially improved by using estimated weights in the weighting procedure, which is an effective way of incorporating incomplete covariate data in the estimation procedure.

We now describe how to achieve this efficiency improvement. We consider the case where the observation probabilities $\pi_i$ can be estimated from a parametric model $\pi = \pi(\mathbf{W}; \boldsymbol{\alpha}_0)$, where $\mathbf{W} \equiv (X, \Delta, \mathbf{Z}_o(\cdot))$ is the observed data and $\boldsymbol{\alpha}_0$ is the true parameter. Let $\mathbf{W}_i$, $i = 1, \ldots, n$, be independent copies of $\mathbf{W}$ and let $\widehat{\boldsymbol{\alpha}}$ be an estimator of $\boldsymbol{\alpha}_0$. Replacing $\pi_i$ by the estimates $\pi_i(\widehat{\boldsymbol{\alpha}}) \equiv \pi(\mathbf{W}_i; \widehat{\boldsymbol{\alpha}})$ in (2.3) and (2.4), we obtain the simple weighted pseudoscore function $\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})$ with estimated weights, and the resulting estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})$. Denote $\pi'(\boldsymbol{\alpha}) = \partial\pi(\mathbf{W};\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. The asymptotic properties of $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})$ are given by the following result.

**Theorem 2.2.** *Under Conditions 2.1–2.3 in Section 2.7, $\sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}}) - \boldsymbol{\beta}_0\}$ is asymptotically normal with mean zero and covariance matrix $\mathbf{D}^{-1}(\boldsymbol{\Sigma}_{\mathrm{SW}} - \mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\mathbf{A}^T)\mathbf{D}^{-1}$, where $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ is the asymptotic variance of $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ and $\mathbf{A} = E\{\pi(\boldsymbol{\alpha}_0)^{-1}\mathbf{M}_{\mathbf{Z}}\pi'(\boldsymbol{\alpha}_0)^T\}$.*

Compared with Theorem 2.1, Theorem 2.2 shows that using estimated weights in the simple weighted estimators yields an efficiency improvement quantified by the term $\mathbf{D}^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\mathbf{A}^T\mathbf{D}^{-1}$. Hence, it is preferable to use estimated weights even if the observation probabilities $\pi_i$ are known. This phenomenon, though somewhat counterintuitive, has often been noted in the missing data literature; see, e.g., Robins, Rotnitzky and Zhao (1994), Lawless, Kalbfleisch and Wild (1999), and Henmi and Eguchi (2004). Note, however, that $\mathbf{A} = \mathbf{0}$ if $\pi$ is predictable; thus, in order to increase efficiency, the model for $\pi$ should include the outcome variables. In this case, a richly parameterized model not only helps avoid model misspecification, but can also lead to greater efficiency improvement.

*Remark* 2.1. The weighting methods discussed here are closely related to those developed for the Cox model under the case-cohort design (Borgan et al., 2000; Kulich and Lin, 2004). However, as noted by Kulich and Lin (2000a), the pseudoscore function differs in fundamental ways from the partial likelihood score function, so that the development of asymptotic theory under model (2.1) requires a new approach. The empirical process arguments used by Kulich and Lin (2000a) rely on the Skorokhod strong embedding theorem and the second Helly's theorem. Here we take a somewhat simpler and more transparent approach that uses the language of Glivenko–Cantelli and Donsker classes and a key lemma of van der Vaart (1998, Lemma 19.24). This approach is in a similar spirit to that of Nan, Kalbfleisch and Yu (2009), who considered the accelerated failure time model with time-independent and bounded covariates.

### 2.2.2 Estimation of the Cumulative Baseline Hazard Function

We now turn to estimation of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$. A similar two-stage weighting procedure can be applied. Specifically, in the estimator of $\Lambda_0(\cdot)$ in Lin and Ying (1994), replace $S^{(0)}(t)$ by $S^{(0)}_{\mathrm{SW}}(t)$ defined in (2.2), and then weight each term by $R_i/\pi_i$. This gives the simple weighted estimator with true weights,

$$\widehat{\Lambda}_0(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\pi_i} \int_0^t \frac{dN_i(u)}{S^{(0)}_{\mathrm{SW}}(u)} - \widehat{\boldsymbol{\beta}}^T_{\mathrm{SW}} \int_0^t \overline{\mathbf{Z}}_{\mathrm{SW}}(u)\,du.$$

If $\pi_i$ are replaced by the estimates $\pi_i(\widehat{\boldsymbol{\alpha}})$, we obtain the simple weighted estimator with estimated weights,

$$\widehat{\Lambda}_0(t,\widehat{\boldsymbol{\alpha}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})} \int_0^t \frac{dN_i(u)}{S^{(0)}_{\mathrm{SW}}(u,\widehat{\boldsymbol{\alpha}})} - \widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})^T \int_0^t \overline{\mathbf{Z}}_{\mathrm{SW}}(u,\widehat{\boldsymbol{\alpha}})\,du.$$

Let $\ell^\infty[0,\tau]$ denote the space of bounded functions on $[0,\tau]$ equipped with the uniform norm, and define

$$M_Y(t) = \int_0^t s^{(0)}(u)^{-1}\,dM(u).$$

The asymptotic properties of $\widehat{\Lambda}_0(\cdot)$ and $\widehat{\Lambda}_0(\cdot,\widehat{\boldsymbol{\alpha}})$ are given by Theorems 2.3 and 2.4, respectively.

**Theorem 2.3.** *Under Conditions 2.1 and 2.2 in Section 2.7, $\sqrt{n}\{\widehat{\Lambda}_0(\cdot) - \Lambda_0(\cdot)\}$ converges weakly in $\ell^\infty[0,\tau]$ to a zero-mean Gaussian process with covariance function*

$$C_{\mathrm{SW}}(s,t) = B(s,t) + \mathbf{h}(s)^T\mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathrm{SW}}\mathbf{D}^{-1}\mathbf{h}(t) - \mathbf{h}(s)^T\mathbf{D}^{-1}\mathbf{k}(t) - \mathbf{h}(t)^T\mathbf{D}^{-1}\mathbf{k}(s),$$

*where $B(s,t) = E\{\pi^{-1}M_Y(s)M_Y(t)\}$, $\mathbf{h}(t) = \int_0^t \mathbf{e}(u)\,du$, and $\mathbf{k}(t) = E\{\pi^{-1}M_Y(t)\mathbf{M}_{\mathbf{Z}}\}$.*

Recall the decomposition of $\boldsymbol{\Sigma}_{\mathrm{SW}}$ in Theorem 2.1, and note that $B(s,t)$ and $\mathbf{k}(t)$ defined above can be decomposed in a similar way. Then Theorem 2.3 shows that the missing data add to the full-data covariance function the extra terms

$$E\{(1-\pi)\pi^{-1}M_Y(s)M_Y(t)\} + \mathbf{h}(s)^T\mathbf{D}^{-1}E\{(1-\pi)\pi^{-1}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}\}\mathbf{D}^{-1}\mathbf{h}(t)$$
$$- \mathbf{h}(s)^T\mathbf{D}^{-1}E\{(1-\pi)\pi^{-1}M_Y(t)\mathbf{M}_{\mathbf{Z}}\} - \mathbf{h}(t)^T\mathbf{D}^{-1}E\{(1-\pi)\pi^{-1}M_Y(s)\mathbf{M}_{\mathbf{Z}}\}.$$

**Theorem 2.4.** *Under Conditions 2.1–2.3 in Section 2.7,* $\sqrt{n}\{\widehat{\Lambda}_0(\cdot,\widehat{\boldsymbol{\alpha}}) - \Lambda_0(\cdot)\}$ *converges weakly in* $\ell^\infty[0,\tau]$ *to a zero-mean Gaussian process with covariance function*

$$C_{\mathrm{SW}}(s,t,\widehat{\boldsymbol{\alpha}}) = C_{\mathrm{SW}}(s,t) - \{\mathbf{m}(s)^T - \mathbf{h}(s)^T\mathbf{D}^{-1}\mathbf{A}\}\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\{\mathbf{m}(t) - \mathbf{A}^T\mathbf{D}^{-1}\mathbf{h}(t)\},$$

*where* $\mathbf{m}(t) = E\{\pi(\boldsymbol{\alpha}_0)^{-1}M_Y(t)\pi'(\boldsymbol{\alpha}_0)\}$.

Similar to Theorem 2.2, Theorem 2.4 shows that an efficiency gain in estimating $\Lambda_0(\cdot)$, quantified by the second term in the expression of $C_{\mathrm{SW}}(s,t,\widehat{\boldsymbol{\alpha}})$, can be obtained by using estimated weights in the simple weighted estimators.

### 2.2.3 Variance Estimation

The covariance matrices for the estimators in Theorems 2.1–2.4 can be consistently estimated by replacing the unknown quantities by the inverse probability weighted sample estimates. For example, when estimated weights are used, define

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\,d\widehat{\Lambda}_0(u,\widehat{\boldsymbol{\alpha}}) - \widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})^T \int_0^t Y_i(u)\mathbf{Z}_i(u)\,du.$$

The matrices $\mathbf{D}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{\mathrm{SW}}$, and $\mathbf{A}$ in Theorems 2.1 and 2.2 can be consistently estimated, respectively, by

$$\begin{aligned}
\widehat{\mathbf{D}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^{\tau} Y_i(t)\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\widehat{\boldsymbol{\alpha}})\}^{\otimes 2}\,dt,\\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^{\tau}\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\widehat{\boldsymbol{\alpha}})\}^{\otimes 2}\,dN_i(t),\\
\widehat{\boldsymbol{\Sigma}}_{\mathrm{SW}} &= \widehat{\boldsymbol{\Sigma}} + \frac{1}{n}\sum_{i=1}^{n}\frac{R_i\{1-\pi_i(\widehat{\boldsymbol{\alpha}})\}}{\pi_i(\widehat{\boldsymbol{\alpha}})^2}\left[\int_0^{\tau}\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\widehat{\boldsymbol{\alpha}})\}\,d\widehat{M}_i(t)\right]^{\otimes 2},
\end{aligned}$$

and

$$\widehat{\mathbf{A}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})^2}\int_0^{\tau}\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\widehat{\boldsymbol{\alpha}})\}\,d\widehat{M}_i(t)\,\pi_i'(\widehat{\boldsymbol{\alpha}})^T.$$

Similarly, the quantities $B(s,t)$, $\mathbf{k}(t)$, $\mathbf{h}(t)$, and $\mathbf{m}(t)$ in Theorems 2.3 and 2.4 can be consistently estimated, respectively, by

$$\begin{aligned}
\widehat{B}(s,t) &= \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^{s\wedge t}\frac{dN_i(u)}{S_{\mathrm{SW}}^{(0)}(u,\widehat{\boldsymbol{\alpha}})^2}\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{R_i\{1-\pi_i(\widehat{\boldsymbol{\alpha}})\}}{\pi_i(\widehat{\boldsymbol{\alpha}})^2}\int_0^{s}\frac{d\widehat{M}_i(u)}{S_{\mathrm{SW}}^{(0)}(u,\widehat{\boldsymbol{\alpha}})}\int_0^{t}\frac{d\widehat{M}_i(u)}{S_{\mathrm{SW}}^{(0)}(u,\widehat{\boldsymbol{\alpha}})},\\
\widehat{\mathbf{h}}(t) &= \int_0^{t}\bar{\mathbf{Z}}_{\mathrm{SW}}(u,\widehat{\boldsymbol{\alpha}})\,du,\\
\widehat{\mathbf{k}}(t) &= \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^{t}\{\mathbf{Z}_i(u)-\bar{\mathbf{Z}}_{\mathrm{SW}}(u,\widehat{\boldsymbol{\alpha}})\}\frac{dN_i(u)}{S_{\mathrm{SW}}^{(0)}(u,\widehat{\boldsymbol{\alpha}})}\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{R_i\{1-\pi_i(\widehat{\boldsymbol{\alpha}})\}}{\pi_i(\widehat{\boldsymbol{\alpha}})^2}\int_0^{t}\frac{d\widehat{M}_i(u)}{S_{\mathrm{SW}}^{(0)}(u,\widehat{\boldsymbol{\alpha}})}\int_0^{\tau}\{\mathbf{Z}_i(u)-\bar{\mathbf{Z}}_{\mathrm{SW}}(u,\widehat{\boldsymbol{\alpha}})\}\,d\widehat{M}_i(u),
\end{aligned}$$

and

$$\widehat{\mathbf{m}}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})^2}\int_0^{t}\frac{d\widehat{M}_i(u)}{S_{\mathrm{SW}}^{(0)}(u,\widehat{\boldsymbol{\alpha}})}\pi_i'(\widehat{\boldsymbol{\alpha}}).$$

The consistency of these estimators can be established following the same lines of arguments as in the proofs of Theorems 2.1–2.4.

### 2.2.4 Modeling Observation Probabilities

To model the observation probabilities, recall that under the MAR assumption, $\pi$ depends only on the observed data $\mathbf{W}$. One important scenario is when the cohort can be divided into a finite number of strata by matching on the values of $\mathbf{W}$, and within stratum $j$, each subject is independently selected with probability $\alpha_j$ so that their covariates are completely observed. Let $N_j$ denote the number of subjects in stratum $j$, and $n_j$ the number of subjects that are selected; then the maximum likelihood estimator for $\alpha_j$ is $n_j/N_j$. More generally, let $\mathbf{W}^*$ be a time-independent random vector which may include an intercept, the components of $\mathbf{W}$, their transforms, and interactions. One then can estimate the observation probabilities $\pi_i$ by fitting the logistic regression model

$$\pi = \frac{\exp(\boldsymbol{\alpha}_0^T W^*)}{1 + \exp(\alpha_0^T W^*)}.$$

In general, it is preferable to include as many terms as possible in this model, with twofold benefits—both to prevent model misspecification and to improve efficiency. One should be cautious, however, with small sample sizes, in which case fitting a high-dimensional model for $\pi$ could result in an unstable estimator with inferior performance.

## 2.3 Augmented Weighted Estimators

### 2.3.1 The Estimators

Within a general framework of semiparametric models with missing data, Robins, Rotnitzky and Zhao (1994) introduced a class of estimators by adding an augmentation term to the inverse probability weighted estimating function, where the augmentation term is a function of the observed data with conditional mean zero given the full data. More specifically, the estimators $\widehat{\boldsymbol{\beta}}(\mathbf{f}, \mathbf{g})$ they proposed solve the estimating equation

$$\sum_{i=1}^{n} \mathbf{H}_i(\boldsymbol{\beta}, \mathbf{f}, \mathbf{g}) = \mathbf{0},$$

where $\mathbf{H}_i$ are independent terms of the form

$$\mathbf{H} = \frac{R}{\pi}\mathbf{f} - \frac{R - \pi}{\pi}\mathbf{g}.$$

Here $\mathbf{f}$ is a mean-zero estimating function based on the full data, and $\mathbf{g}$ is a square integrable function of the observed data. These authors showed that a semiparametric efficient estimator can be obtained by first minimizing the variance of $\mathbf{H}$ with respect to $\mathbf{g}$ for a fixed $\mathbf{f}$ and then optimize over $\mathbf{f}$. The second step of this optimization procedure is often difficult, but the first step is easily carried out by taking $\mathbf{g}$ to be the conditional expectation of $\mathbf{f}$ given the observed data, thus resulting in a so-called locally efficient estimator; see also the discussion in van der Vaart (1998, p. 383). The estimators thus derived are in general more efficient than those based on the original estimating function, and more importantly, they enjoy the double robustness property; that is, they are consistent if either the model for the observation probabilities or the model for the joint distribution of the covariates is correctly specified (Robins and Rotnitzky, 2001).

To adapt the general methodology to model (2.1), we need a two-stage augmentation procedure. First, we add an augmentation term to $\mathbf{S}_{\mathrm{SW}}^{(k)}(t)$ defined in (2.2) and obtain

$$\mathbf{S}_{\mathrm{AW}}^{(k)}(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{R_j}{\pi_j} Y_j(t)\mathbf{Z}_j(t)^{\otimes k} - \frac{1}{n}\sum_{j=1}^{n} \frac{R_j - \pi_j}{\pi_j} Y_j(t) E\{\mathbf{Z}_j(t)^{\otimes k} \mid \mathbf{W}_j\}, \quad k = 0, 1.$$

Second, we add an augmentation term to $\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta})$ defined in (2.3) and obtain the augmented weighted pseudoscore function

$$\begin{aligned}\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}) = &\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\int_0^{\tau}\{\mathbf{Z}_i(t)-\overline{\mathbf{Z}}_{\mathrm{AW}}(t)\}\{dN_i(t)-Y_i(t)\boldsymbol{\beta}^T\mathbf{Z}_i(t)\,dt\}\\ &-\frac{1}{n}\sum_{i=1}^{n}\frac{R_i-\pi_i}{\pi_i}\int_0^{\tau}[E\{\mathbf{Z}_i(t)\,dN_i(t)-Y_i(t)\mathbf{Z}_i(t)^{\otimes 2}\boldsymbol{\beta}\,dt\mid\mathbf{W}_i\}\\ &\qquad-\overline{\mathbf{Z}}_{\mathrm{AW}}(t)E\{dN_i(t)-Y_i(t)\boldsymbol{\beta}^T\mathbf{Z}_i(t)\,dt\mid\mathbf{W}_i\}],\end{aligned}$$

where

$$\overline{\mathbf{Z}}_{\mathrm{AW}}(t)=\mathbf{S}_{\mathrm{AW}}^{(1)}(t)/S_{\mathrm{AW}}^{(0)}(t).$$

We first consider the case where the observation probabilities $\pi_i$ are known. Define the augmented weighted estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}$ as the unique solution to the estimating equation $\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta})=\mathbf{0}$, which has the closed form

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{AW}} = &\left[\sum_{i=1}^{n}\frac{R_i}{\pi_i}\int_0^{\tau}Y_i(t)\{\mathbf{Z}_i(t)-\overline{\mathbf{Z}}_{\mathrm{AW}}(t)\}^{\otimes 2}\,dt-\mathbf{F}\right]^{-1}\\ &\times\left[\sum_{i=1}^{n}\frac{R_i}{\pi_i}\int_0^{\tau}\{\mathbf{Z}_i(t)-\overline{\mathbf{Z}}_{\mathrm{AW}}(t)\}\,dN_i(t)-\mathbf{G}\right],\end{aligned}$$

where

$$\mathbf{F}=\sum_{i=1}^{n}\frac{R_i-\pi_i}{\pi_i}\int_0^{\tau}Y_i(t)[E\{\mathbf{Z}_i(t)^{\otimes 2}\mid\mathbf{W}_i\}-\overline{\mathbf{Z}}_{\mathrm{AW}}(t)E\{\mathbf{Z}_i(t)^T\mid\mathbf{W}_i\}]\,dt$$

and

$$\mathbf{G}=\sum_{i=1}^{n}\frac{R_i-\pi_i}{\pi_i}\int_0^{\tau}[E\{\mathbf{Z}_i(t)\mid\mathbf{W}_i\}-\overline{\mathbf{Z}}_{\mathrm{AW}}(t)]\,dN_i(t).$$

Although the estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}$ involves some unknown conditional expectations, it will be clear below that its implementation is quite convenient and flexible; see the discussion in Section 2.3.4.

Now consider the case where the observation probabilities $\pi_i$ and the unknown conditional expectations $E\{\mathbf{Z}_m(\cdot)^{\otimes k} \mid \mathbf{W}\}$ are estimated from the parametric models $\pi = \pi(\mathbf{W}; \boldsymbol{\alpha}_0)$ and $E\{\mathbf{Z}_m(\cdot)^{\otimes k} \mid \mathbf{W}\} = \boldsymbol{\mu}_k(\mathbf{W}; \boldsymbol{\gamma}_k)$, $k = 1, 2$, respectively, where $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0 \equiv (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ are the true parameters. Let $\widehat{\boldsymbol{\alpha}}$ be an estimator of $\boldsymbol{\alpha}_0$, and $\widehat{\boldsymbol{\gamma}}$ an estimator of $\boldsymbol{\gamma}_0$. When $\pi_i$ and $E\{\mathbf{Z}_{mi}(\cdot)^{\otimes k} \mid \mathbf{W}_i\}$ involved in $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}$ are substituted by the estimates $\pi_i(\widehat{\boldsymbol{\alpha}})$ and $\boldsymbol{\mu}_k(\mathbf{W}_i; \widehat{\boldsymbol{\gamma}}_k)$, respectively, the resulting estimator is denoted by $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$.

### 2.3.2 Asymptotic Properties

We now develop the asymptotic theory for the augmented weighted estimators. For ease of presentation, we first state an asymptotic result assuming that the observation probabilities and the conditional expectations involved in these estimators are known.

**Theorem 2.5.** *Under Conditions 2.1 and 2.2 in Section 2.7, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{AW}} - \boldsymbol{\beta}_0)$ is asymptotically normal with mean zero and covariance matrix $\mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathrm{AW}}\mathbf{D}^{-1}$, where*

$$\boldsymbol{\Sigma}_{\mathrm{AW}} = \boldsymbol{\Sigma}_{\mathrm{SW}} - E[(1-\pi)\pi^{-1}\{E(\mathbf{M}_{\mathbf{Z}} \mid \mathbf{W})\}^{\otimes 2}] \tag{2.5}$$

$$= \boldsymbol{\Sigma} + E\{(1-\pi)\pi^{-1}\,\mathrm{Var}(\mathbf{M}_{\mathbf{Z}} \mid \mathbf{W})\}. \tag{2.6}$$

Compared with Theorem 2.1, Theorem 2.5 shows that $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}$ is in general more efficient than the simple weighted estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$. The second term in (2.5) characterizes the efficiency gain compared to $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}$, while the second term in (2.6) characterizes the efficiency loss compared to the full-data estimator.

In general, however, the observation probabilities and the conditional expectations are unknown and need to be estimated. Especially, these conditional expectations depend on the joint distribution of the covariates as well as the main model (2.1), and thus a correct form of them may not be readily seen. Nevertheless, we now present a theorem which makes the implementation of the augmented weighted estimators practical and flexible by showing that substituting the nuisance parameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$ by the sample estimates does not increase the asymptotic variance, even if these estimates are consistent at rates slower than $\sqrt{n}$. Moreover, it is shown that the augmented weighted estimators are in general no less efficient than the simple weighted estimators with estimated weights.

**Theorem 2.6.** *Under Conditions 2.1, 2.2, 2.3(i), and 2.4 in Section 2.7, $\sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0\}$ is asymptotically normal with mean zero. In addition, the asymptotic variance is the same as that of $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}$ and is no greater than that of $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})$.*

*Remark* 2.2. Since the augmentation procedure is first applied to $\mathbf{S}^{(k)}_{\mathrm{SW}}(\cdot)$, $k = 0, 1$, to form $\overline{\mathbf{Z}}_{\mathrm{AW}}(\cdot)$, and then is applied to $\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta})$ without altering $\overline{\mathbf{Z}}_{\mathrm{AW}}(\cdot)$, the resulting estimating function $\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta})$ is not a sum of independent terms. Therefore, our asymptotic results are not a direct consequence of the general theory of Robins, Rotnitzky and Zhao (1994). Similar augmented weighting procedures for the Cox model were studied by Wang and Chen (2001) and Qi, Wang and Prentice (2005). The proofs of Theorems 2.5 and 2.6 involve empirical process techniques similar to those used in the proofs of Theorems 2.1 and 2.2, and a key step is to derive an asymptotic independent-sum representation for $\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0)$.

### 2.3.3 Variance Estimation

Variance estimation for the augmented weighted estimators can be similarly carried out as for the simple weighted estimators. We present only the formulas with estimated nuisance parameters. Define

$$\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})} \int_0^t \frac{dN_i(u)}{S^{(0)}_{\mathrm{AW}}(u, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})} - \widehat{\boldsymbol{\beta}}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})^T \int_0^t \overline{\mathbf{Z}}_{\mathrm{AW}}(u, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\, du,$$

$$\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\, d\widehat{\Lambda}_0(u, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) - \widehat{\boldsymbol{\beta}}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})^T \int_0^t Y_i(u)\mathbf{Z}_i(u)\, du,$$

and further,

$$\widehat{\mathbf{M}}_{\mathbf{Z}i} = \int_0^\tau \{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\mathrm{AW}}(t, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\}\, d\widehat{M}_i(t).$$

Then the matrices $\mathbf{D}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_{\mathrm{AW}}$ appearing in Theorems 2.5 and 2.6 can be consistently estimated, respectively, by

$$\widehat{\mathbf{D}} = \frac{1}{n}\sum_{i=1}^n \frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})} \int_0^\tau Y_i(t)\{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\mathrm{AW}}(t, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\}^{\otimes 2}\, dt,$$
$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})} \int_0^\tau \{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\mathrm{AW}}(t, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\}^{\otimes 2}\, dN_i(t),$$

and

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{AW}} = \widehat{\boldsymbol{\Sigma}} + \frac{1}{n}\sum_{i=1}^n \frac{R_i\{1 - \pi_i(\widehat{\boldsymbol{\alpha}})\}}{\pi_i(\widehat{\boldsymbol{\alpha}})^2}\{\widehat{\mathbf{M}}_{\mathbf{Z}i} - E(\widehat{\mathbf{M}}_{\mathbf{Z}i} \,|\, \mathbf{W}_i; \widehat{\boldsymbol{\gamma}})\}^{\otimes 2}.$$

### 2.3.4 Modeling Auxiliary Distributions

In the implementation of the augmented weighted estimators, we need to estimate the observation probabilities $\pi_i$ and the conditional expectations $E\{\mathbf{Z}_{mi}(\cdot)^{\otimes k} \,|\, \mathbf{W}_i\}$, $k = 1, 2$; the former has been discussed in Section 2.2.4, and we now discuss the latter. Since these conditional expectations depend on the joint distribution of the covariates as well as the regression parameters $\boldsymbol{\beta}_0$ and $\Lambda_0(\cdot)$ in model (2.1), our first proposal is to model the conditional distribution of $\mathbf{Z}_m(\cdot)$ given $\mathbf{Z}_o(\cdot)$, and then estimate the desired quantities with some initial estimates of $\boldsymbol{\beta}_0$ and $\Lambda_0(\cdot)$. Under the Cox model, Wang and Chen (2001) proposed an iterative EM-type algorithm to estimate the conditional expectations. Our asymptotic results, however, suggest that one-step estimation should be sufficient in our case, provided that the initial estimates are consistent. To carry out this proposal for model (2.1), unlike the Cox model, an estimate of the baseline hazard function $\lambda_0(\cdot)$ is required. This can be obtained by, for example, the kernel estimator

$$\widehat{\lambda}_0(t) = \int_0^\tau K_h(t-s)\, d\widehat{\Lambda}_0(s),$$

where $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function with bandwidth $h$, and $\widehat{\Lambda}_0(\cdot)$ is an estimator of $\Lambda_0(\cdot)$ (Ramlau-Hansen, 1983). Some of our experience indicates that such estimates should be appropriate for moderate to large sample sizes and the quality of $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}$ seems fairly robust to the choice of smoothing parameters.

A more convenient method is to directly specify a working model for the conditional expectations. The quality of the resulting estimator is protected by the double robustness property against misspecification of this working model, provided that the model for the observation probabilities is correctly specified. In practice, one could first build a rich family of parametric models and then use model selection techniques to guide the choice of the best model. We find that this direct modeling strategy often works surprisingly well, and thus recommend it for practical use because of its simplicity.

## 2.4 Simulation Studies

In this section we present some simulation results to study the finite-sample performance of the proposed estimators and compare them to the full-data estimators and complete-case analysis. We generated data from the model

$$\lambda(t \mid Z_1, Z_2) = 0.5 + \beta_1 Z_1 + \beta_2 Z_2,$$

where $Z_1, Z_2 \sim \mathrm{Ber}(0.5)$ with $\Pr(Z_2 = 0 \mid Z_1 = 0) = \Pr(Z_2 = 1 \mid Z_1 = 1) = \varphi$; we took $\beta_1 = 0.4$, $\beta_2 = 0.5$, and $\varphi = 0.5$ or $\varphi = 0.8$. The covariate $Z_1$ is always observed and $Z_2$ is possibly missing with observation probability

$$\pi = \pi(\Delta, Z_1) = 0.1(1-\Delta)(1-Z_1) + 0.3(1-\Delta)Z_1 + 0.5\Delta(1-Z_1) + 0.7\Delta Z_1.$$

The censoring time is $C = U \wedge 1$, where $U \sim \text{Unif}(0, c)$ with constant $c$ chosen so that the censoring rate is approximately 50%. The above setting resulted in a missingness rate of about 60% for the covariate $Z_2$.

In implementing the estimators with estimated observation probabilities and/or conditional expectations, we fitted in each case a logistic regression model with $X$, $\Delta$, $Z_1$ and their pairwise interactions as predictors. In addition, to force the estimators of $\Lambda_0(\cdot)$ to be increasing, we follow the method of Lin and Ying (1994) and define the monotonic version $\widehat{\Lambda}_0^*(t) = \max_{s\in[0,t]} \widehat{\Lambda}_0(s)$. As shown in that paper, this modified version has the same asymptotic distribution as the original estimator. We took the sample size to be $n = 500$ or $n = 1000$, and for each pair of values of $\varphi$ and $n$, replicated the simulation 1000 times. Simulation results for estimating the regression coefficients and the cumulative baseline hazard function are summarized in Tables 2.1 and 2.2, respectively. For each method, we report the biases and standard errors of the estimates from the 1000 replications, the theoretically estimated standard errors averaged over the replications, and the coverage percentages of the theoretically constructed 95% confidence intervals. For the estimation of $\Lambda_0(\cdot)$, these were evaluated at three selected points $t = 0.25$, 0.5, and 0.75 to facilitate comparison.

Since the missingness depends on the failure indicator $\Delta$, we expect the estimates from a complete-case analysis to be biased; this is confirmed by the results in Table 2.1. The biases of all the other methods are minimal. We note, however, that the simple weighted estimator with estimated weights and all of the augmented weighted estimators have slightly smaller biases for estimating $\beta_1$ than the simple weighted estimator with true weights. It is also clear from Table 2.1 that, by using estimated weights, the efficiency of the simple weighted estimator for $\beta_1$ is dramatically improved. Substituting estimated nuisance parameters in the augmented weighted estimators seems to have little effect on the efficiency, and all of these estimators have slightly smaller variance for estimating $\beta_1$ than the simple weighted estimator with estimated weights. Comparing the results for different values of $\varphi$, we see that the efficiency gain of the proposed estimators is most substantial when $Z_1$ and $Z_2$ are uncorrelated, i.e., $\varphi = 0.5$, and decreases as the correlation between $Z_1$ and $Z_2$ increases. In all cases except the complete-case analysis, the theoretically estimated standard errors are close to the sample standard errors and give reasonable coverage percentages, indicating that asymptotic approximations should be appropriate for practical sample sizes. The results in Table 2.2 demonstrate similar trends in estimating $\Lambda_0(\cdot)$, except that the performance of the proposed estimators for $\Lambda_0(\cdot)$ actually improves as $Z_1$ and $Z_2$ become more correlated.

## 2.5 Example: Mouse Leukemia Data

As an illustrative example, we applied the proposed methods to analyzing a mouse leukemia data set given in Kalbfleisch and Prentice (2002, Appendix A, Data Set VI), which has been studied by a number of authors, e.g., Chen and Little (1999), Wang and Chen (2001), Qi, Wang and Prentice (2005), and Tsai (2009), using the Cox model. The data set was collected in Dr. Robert Nowinski's laboratories at the Fred Hutchinson Cancer Research Center to investigate the effects of genetic and viral factors on the development of spontaneous mouse leukemia. A total of 204 mice were followed over a 2-year period, among which 67 died of thymic leukemia and 12 died of nonthymic leukemia. The median follow-up time is 678 days. Data on sex, coat color, virus level, and a few other covariates were recorded for most of the mice. However, one of the most important covariates, Gpd-1 phenotype,

Table 2.1: Simulation results for estimating the regression coefficients in the model $\lambda(t \mid Z_1, Z_2) = 0.5 + \beta_1 Z_1 + \beta_2 Z_2$.

| | | $\beta_1$ | | | | $\beta_2$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Method | Bias | SE | Est. SE | CP | Bias | SE | Est. SE | CP |
| $\varphi = 0.5$ | Full data | 0.003 | 0.120 | 0.118 | 95.1 | 0.001 | 0.119 | 0.119 | 95.2 |
| $n = 500$ | CC | −0.362 | 0.305 | 0.289 | 75.2 | 0.099 | 0.262 | 0.277 | 95.0 |
| | SW | −0.028 | 0.259 | 0.235 | 93.1 | 0.012 | 0.246 | 0.246 | 95.1 |
| | SW($\hat{\boldsymbol{\alpha}}$) | −0.009 | 0.145 | 0.140 | 94.0 | 0.015 | 0.252 | 0.245 | 94.3 |
| | AW | −0.006 | 0.135 | 0.137 | 95.6 | 0.016 | 0.242 | 0.250 | 97.6 |
| | AW($\hat{\boldsymbol{\alpha}}$) | −0.008 | 0.139 | 0.140 | 95.9 | 0.017 | 0.254 | 0.261 | 97.9 |
| | AW($\hat{\boldsymbol{\gamma}}$) | −0.007 | 0.137 | 0.134 | 93.5 | 0.017 | 0.249 | 0.240 | 93.6 |
| | AW($\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}$) | −0.007 | 0.138 | 0.135 | 94.2 | 0.017 | 0.253 | 0.248 | 94.9 |
| $\varphi = 0.5$ | Full data | 0.004 | 0.082 | 0.083 | 94.8 | −0.002 | 0.086 | 0.084 | 95.0 |
| $n = 1000$ | CC | −0.366 | 0.199 | 0.203 | 55.8 | 0.108 | 0.195 | 0.194 | 91.7 |
| | SW | −0.013 | 0.164 | 0.162 | 95.1 | 0.011 | 0.179 | 0.172 | 94.7 |
| | SW($\hat{\boldsymbol{\alpha}}$) | −0.002 | 0.100 | 0.097 | 94.2 | 0.012 | 0.176 | 0.170 | 95.0 |
| | AW | −0.001 | 0.096 | 0.094 | 95.2 | 0.012 | 0.175 | 0.171 | 95.8 |
| | AW($\hat{\boldsymbol{\alpha}}$) | −0.002 | 0.096 | 0.094 | 95.4 | 0.012 | 0.177 | 0.174 | 95.9 |
| | AW($\hat{\boldsymbol{\gamma}}$) | −0.001 | 0.096 | 0.093 | 94.6 | 0.011 | 0.176 | 0.168 | 94.6 |
| | AW($\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}$) | −0.001 | 0.096 | 0.094 | 95.1 | 0.012 | 0.177 | 0.171 | 95.0 |
| $\varphi = 0.8$ | Full data | 0.008 | 0.153 | 0.149 | 94.6 | −0.008 | 0.145 | 0.150 | 95.4 |
| $n = 500$ | CC | −0.349 | 0.359 | 0.350 | 82.0 | 0.099 | 0.331 | 0.340 | 94.4 |
| | SW | −0.041 | 0.328 | 0.294 | 90.4 | 0.040 | 0.325 | 0.301 | 90.5 |
| | SW($\hat{\boldsymbol{\alpha}}$) | −0.035 | 0.256 | 0.229 | 89.8 | 0.046 | 0.330 | 0.299 | 88.6 |
| | AW | −0.029 | 0.251 | 0.238 | 93.1 | 0.041 | 0.327 | 0.313 | 94.5 |
| | AW($\hat{\boldsymbol{\alpha}}$) | −0.035 | 0.254 | 0.243 | 93.4 | 0.047 | 0.331 | 0.320 | 94.7 |
| | AW($\hat{\boldsymbol{\gamma}}$) | −0.032 | 0.250 | 0.227 | 90.8 | 0.045 | 0.326 | 0.297 | 88.5 |
| | AW($\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}$) | −0.034 | 0.252 | 0.228 | 90.5 | 0.047 | 0.328 | 0.301 | 89.0 |
| $\varphi = 0.8$ | Full data | 0.003 | 0.105 | 0.104 | 94.7 | 0.001 | 0.106 | 0.105 | 93.8 |
| $n = 1000$ | CC | −0.360 | 0.233 | 0.244 | 68.4 | 0.110 | 0.237 | 0.237 | 92.1 |
| | SW | −0.023 | 0.213 | 0.210 | 94.8 | 0.025 | 0.227 | 0.217 | 92.0 |
| | SW($\hat{\boldsymbol{\alpha}}$) | −0.016 | 0.172 | 0.165 | 92.6 | 0.026 | 0.229 | 0.216 | 91.7 |
| | AW | −0.016 | 0.167 | 0.166 | 94.3 | 0.026 | 0.224 | 0.218 | 93.9 |
| | AW($\hat{\boldsymbol{\alpha}}$) | −0.017 | 0.170 | 0.167 | 93.8 | 0.027 | 0.228 | 0.221 | 93.6 |
| | AW($\hat{\boldsymbol{\gamma}}$) | −0.015 | 0.170 | 0.163 | 92.0 | 0.025 | 0.228 | 0.214 | 90.6 |
| | AW($\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}$) | −0.016 | 0.171 | 0.164 | 92.9 | 0.026 | 0.229 | 0.216 | 91.8 |

Note: CC, complete-case analysis; SW, simple weighted estimator; AW, augmented weighted estimator; $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ in parentheses indicate that estimated nuisance parameters are used; SE, standard error; CP, coverage percentage (%).

Table 2.2: Simulation results for estimating the cumulative baseline hazard function in the model $\lambda(t \mid Z_1, Z_2) = 0.5 + \beta_1 Z_1 + \beta_2 Z_2$.

| | | $t = 0.25$ | | | | $t = 0.5$ | | | | $t = 0.75$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | Bias | SE | Est. SE | CP | Bias | SE | Est. SE | CP | Bias | SE | Est. SE | CP |
| $\varphi = 0.5$ | Full data | 0.002 | 0.026 | 0.028 | 95.9 | 0.003 | 0.045 | 0.045 | 95.0 | 0.002 | 0.061 | 0.062 | 94.8 |
| $n = 500$ | CC | 0.172 | 0.076 | 0.074 | 35.4 | 0.381 | 0.142 | 0.137 | 17.3 | 0.649 | 0.210 | 0.202 | 6.0 |
| | SW | 0.010 | 0.053 | 0.050 | 95.4 | 0.021 | 0.101 | 0.091 | 95.5 | 0.032 | 0.150 | 0.135 | 94.0 |
| | SW($\hat{\boldsymbol{\alpha}}$) | 0.004 | 0.036 | 0.037 | 96.2 | 0.007 | 0.063 | 0.063 | 95.4 | 0.010 | 0.087 | 0.088 | 95.5 |
| $\varphi = 0.5$ | Full data | 0.001 | 0.020 | 0.020 | 95.0 | 0.001 | 0.032 | 0.032 | 95.5 | 0.001 | 0.044 | 0.044 | 95.4 |
| $n = 1000$ | CC | 0.170 | 0.053 | 0.052 | 8.1 | 0.377 | 0.097 | 0.096 | 1.2 | 0.644 | 0.142 | 0.142 | 0.0 |
| | SW | 0.005 | 0.034 | 0.034 | 94.9 | 0.010 | 0.063 | 0.062 | 95.2 | 0.014 | 0.092 | 0.091 | 95.3 |
| | SW($\hat{\boldsymbol{\alpha}}$) | 0.002 | 0.025 | 0.025 | 96.8 | 0.003 | 0.044 | 0.043 | 95.6 | 0.003 | 0.061 | 0.060 | 95.7 |
| $\varphi = 0.8$ | Full data | 0.002 | 0.024 | 0.025 | 95.0 | 0.003 | 0.037 | 0.038 | 96.0 | 0.002 | 0.050 | 0.051 | 96.0 |
| $n = 500$ | CC | 0.169 | 0.068 | 0.068 | 27.7 | 0.370 | 0.124 | 0.121 | 9.4 | 0.634 | 0.185 | 0.179 | 1.6 |
| | SW | 0.008 | 0.044 | 0.043 | 95.9 | 0.015 | 0.081 | 0.076 | 95.0 | 0.024 | 0.121 | 0.110 | 94.2 |
| | SW($\hat{\boldsymbol{\alpha}}$) | 0.004 | 0.028 | 0.029 | 95.8 | 0.005 | 0.045 | 0.046 | 95.7 | 0.007 | 0.061 | 0.062 | 95.7 |
| $\varphi = 0.8$ | Full data | 0.001 | 0.018 | 0.017 | 94.8 | 0.000 | 0.028 | 0.027 | 94.4 | 0.000 | 0.036 | 0.036 | 95.7 |
| $n = 1000$ | CC | 0.167 | 0.050 | 0.048 | 5.2 | 0.369 | 0.089 | 0.086 | 0.5 | 0.633 | 0.130 | 0.126 | 0.0 |
| | SW | 0.005 | 0.031 | 0.030 | 94.8 | 0.009 | 0.054 | 0.052 | 94.4 | 0.012 | 0.079 | 0.076 | 94.6 |
| | SW($\hat{\boldsymbol{\alpha}}$) | 0.002 | 0.021 | 0.021 | 94.5 | 0.002 | 0.034 | 0.032 | 93.9 | 0.003 | 0.044 | 0.043 | 94.8 |

Note: CC, complete-case analysis; SW, simple weighted estimator; AW, augmented weighted estimator; $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ in parentheses indicate that estimated nuisance parameters are used; SE, standard error; CP, coverage percentage (%).

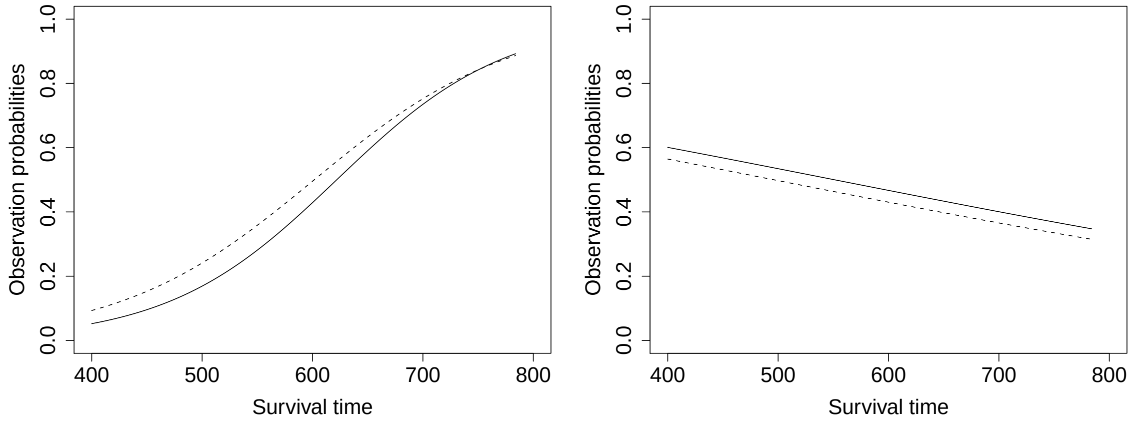

Figure 2.1: Fitted curves for observation probabilities from the mouse leukemia data. The left panel shows plots for censored subjects ($\Delta = 0$) and the right panel for failures ($\Delta = 1$), with types of endpoints being mortality due to thymic leukemia (solid) and mortality due to thymic and nonthymic leukemia (dashes).

was subject to substantial missingness because the Gpd-1 typing started midway through the study. As a result, only 100 mice were typed for Gpd-1, which were selected from those that survived at least 400 days. The selection probabilities, however, may vary among individuals and were not recorded. The MAR assumption seems reasonable here, since the missingness was decided by the experimenter without knowing the results.

Two covariates of primary interest are the Gpd-1 phenotype and the virus level, where the former takes value 0 for $b/b$ and 1 for $b/a$ and the latter was discretized into a binary variable. As in Wang and Chen (2001), only the 175 mice with the virus level data are included in the analysis. We built a family of models for the observation probabilities and the conditional expectations by considering logistic regression models with the survival time, failure indicator, virus level, and their pairwise interactions as potential predictors, among which the best model was chosen by AIC. For the observation probabilities, the chosen model includes the survival time, failure indicator, and their interaction; the fitted curves are shown in Figure 2.1. The clear dependence of the observation probabilities on the outcome variables suggests that the complete-case estimators would be biased.

We fitted model (2.1) to this data set for two types of endpoints, mortality due to thymic leukemia and mortality due to thymic and nonthymic leukemia. The results from the regression analysis are reported in Table 2.3. All of the estimates indicate a negative association of the Gpd-1 phenotype and a positive association of the virus level with the survival time. By utilizing data from the incomplete cases, both the simple weighted estimator and the augmented weighted estimator yield stronger evidence for the effect of the virus level, while the latter indicates the most significant effect among the three methods and has a slightly smaller standard error than the former. For both types of endpoints, the large difference between the complete-case estimate and the two weighted estimates for the virus effect suggests that the former might be biased downward. Comparing with the coefficient estimates under the Cox model in the aforementioned references, the estimates presented here have much smaller absolute values because they have distinct but often more direct interpretations. For example, an estimate of $-0.000512$ for the Gpd-1 effect means that the mice with Gpd-1 phenotype $b/a$ would have on average 5.12 fewer failures than those with phenotype $b/b$ per $10,000$ subject-days of follow-up. Despite these differences, the test statistics for individual covariate effects are quite comparable between the two models, as often noted in the literature.

## 2.6 Discussion

We have developed two different procedures for inference in the additive hazards model with missing covariates. The simple weighted estimators are a convenient method of correcting the biases of complete-case analysis and producing consistent estimates. When the weights are estimated from a richly parameterized model, the efficiency of the simple weighted estimators can be very close to that of the augmented weighted estimators. The augmented method, however, enjoys the double robustness property, which allows more

Table 2.3: Results for different methods applied to the mouse leukemia data. Values shown are estimates of the regression coefficients, with standard errors in parentheses. All values are multiplied by 1000.

| | Thymic leukemia | | Thymic and nonthymic leukemia | |
|---|---|---|---|---|
| Method | Gpd-1 | Virus | Gpd-1 | Virus |
| CC | −0.531<br>(0.226) | 0.307<br>(0.151) | −0.583<br>(0.232) | 0.303<br>(0.153) |
| SW | −0.512<br>(0.218) | 0.397<br>(0.177) | −0.645<br>(0.244) | 0.401<br>(0.188) |
| AW | −0.363<br>(0.222) | 0.814<br>(0.158) | −0.574<br>(0.242) | 0.724<br>(0.172) |

Note: CC, complete-case analysis; SW, simple weighted estimator; AW, augmented weighted estimator.

freedom in modeling the observation probabilities. In addition, since an inverse probability weighted term is subtracted in the augmentation procedure, the augmented method partially alleviates the instability problem arising from dividing by too small observation probabilities in the simple weighted estimators.

As noted by Qi, Wang and Prentice (2005) in the context of Cox regression, by estimating weights nonparametrically, for instance, by the Nadaraya–Watson kernel estimator, the simple weighted estimators can reach the efficiency of the augmented weighted estimators. We have not considered the kernel-assisted method in this dissertation for at least two reasons. First, the kernel estimator may require a relatively large sample size to yield a satisfactory performance, whereas in practice one can always make some parametric assumptions and the resulting estimator usually works sufficiently well. Second, using nonparametrically estimated weights with the augmented method cannot further improve efficiency. Nevertheless, it would be worthwhile to explore the use of kernel-assisted methods in avoiding model misspecification.

The proposed estimators would not, in general, achieve the semiparametric information bound, because the full-data pseudoscore estimator is not fully efficient (Lin and Ying, 1994). An adaptive procedure was suggested by Lin and Ying (1994) and a sieve maximum likelihood approach was taken by Zeng, Yin and Ibrahim (2005) to yield a fully efficient estimator in the full-data case. The adaptation of these methods to the missing-data case, however, may not be straightforward, and the resulting estimators would not enjoy the same robustness advantages as the augmented weighted estimators. This leaves open directions for future research.

## 2.7 Regularity Conditions and Proofs

### 2.7.1 Regularity Conditions

The following regularity conditions are needed in our asymptotic results.

**Condition 2.1.** (i) $\Lambda_0(\tau) < \infty$. (ii) $\Pr\{Y(\tau) = 1\} > 0$. (iii) $E\{\sup_{t\in[0,\tau]} Y(t)\|\mathbf{Z}(t)\|^2\} < \infty$. (iv) $\mathbf{D}$ and $\boldsymbol{\Sigma}$ are positive definite. (v) The sample paths of $\mathbf{Z}(\cdot)$ are left continuous with right limits and are of uniformly bounded variation.

**Condition 2.2.** $\pi_i$ are bounded away from zero.

**Condition 2.3.** (i) $\pi(\mathbf{W};\boldsymbol{\alpha})$ is twice continuously differentiable in $\boldsymbol{\alpha}$, and there exists a compact neighborhood $\mathcal{A}$ of $\boldsymbol{\alpha}_0$ such that $E[\sup_{\boldsymbol{\alpha}\in\mathcal{A}}\{\|\pi'(\boldsymbol{\alpha})\|^2 + \|\pi''(\boldsymbol{\alpha})\|\}] < \infty$, where $\pi'(\boldsymbol{\alpha}) = \partial\pi(\mathbf{W};\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ and $\pi''(\boldsymbol{\alpha}) = \partial^2\pi(\mathbf{W};\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\boldsymbol{\alpha}^T$. (ii) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{SW}} - \boldsymbol{\beta}_0)$ and $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ are asymptotically jointly normal, and $\widehat{\boldsymbol{\alpha}}$ is asymptotically efficient.

**Condition 2.4.** (i) $E_{\boldsymbol{\gamma}}\{\mathbf{Z}_m(\cdot)\,|\,\mathbf{W}\}$ is continuously differentiable in $\boldsymbol{\gamma}$, and there exists a compact neighborhood $\Gamma$ of $\boldsymbol{\gamma}_0$ such that

$$E\left[\sup_{t\in[0,\tau],\,\boldsymbol{\gamma}\in\Gamma} Y(t)\{|f(t,\boldsymbol{\gamma};\mathbf{W})| + \|\partial f(t,\boldsymbol{\gamma};\mathbf{W})/\partial\boldsymbol{\gamma}\|\}\right] < \infty,$$

where $f(t, \boldsymbol{\gamma}; \mathbf{W}) = E_{\boldsymbol{\gamma}}\{\|\mathbf{Z}_m(t)\|^2 \mid \mathbf{W}\}$. (ii) There exist constants $\kappa_1, \kappa_2 \in (0, 1/2]$ with $\kappa_1 + \kappa_2 > 1/2$ such that $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| = O_p(n^{-\kappa_1})$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(n^{-\kappa_2})$.

Parts (i)–(iv) of Condition 2.1 are similar to those typically assumed for the Cox model, e.g., in Theorem 4.1 of Andersen and Gill (1982). Condition 2.1(v) is convenient for establishing Glivenko–Cantelli and Donsker properties, and is also mild enough to be satisfied in most practical situations; it will be clear from the proofs below, however, that it can be replaced by weaker and more abstract conditions. Condition 2.2 is commonly assumed in the missing data literature to ensure the boundedness of the inverse probability weights. Conditions 2.3 and 2.4 are regularity conditions on estimators of the nuisance parameters; note in particular that Condition 2.4(ii) allows the rates of convergence of these estimators to be slower than $\sqrt{n}$.

### 2.7.2 Proofs for Section 2.2.1

*Proof of Theorem 2.1.* We first show that $\{R\pi^{-1}Y(t)\mathbf{Z}(t)^{\otimes k} : t \in [0, \tau]\}$, $k = 0, 1, 2$, are Glivenko–Cantelli. It suffices to establish the property for each component, for example, $\{R\pi^{-1}Y(t)Z_j(t)Z_l(t) : t \in [0, \tau]\}$.

First, note that a function of bounded variation can be expressed as the difference of two increasing functions. By Condition 2.2(v) and Lemma 9.10 of Kosorok (2008), we see that $\mathcal{Z}_j \equiv \{Z_j(t) : t \in [0, \tau]\}$ is a VC-hull class associated with a VC class of index 2. Then, by Corollary 2.6.12 of van der Vaart and Wellner (1996), for any probability measure $Q$,

$$\log N(\varepsilon \|F\|_{Q,2}, \mathcal{Z}_j, L_2(Q)) \le K\left(\frac{1}{\varepsilon}\right),$$

where $F$ is an envelope of the class $\mathcal{Z}_j$ and $K$ is a constant. Thus, the uniform entropy integral of $\mathcal{Z}_j$ is

$$\begin{aligned} J(1, \mathcal{Z}_j, L_2) &= \int_0^1 \sqrt{\log \sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{Z}_j, L_2(Q))}\, d\varepsilon \\ &\le \int_0^1 \sqrt{K(1/\varepsilon)}\, d\varepsilon < \infty, \end{aligned}$$

i.e., the class $\mathcal{Z}_j$ has bounded uniform entropy integral (BUEI). Also, from Example 19.16 of van der Vaart (1998) we know that the collection of all cells $(-\infty, t]$ in the real line is a VC class. Thus, by the definition of $Y(t)$, the class $\mathcal{Y} \equiv \{Y(t) : t \in [0, \tau]\}$ is VC, and hence is BUEI. Since $Z_j(\cdot)$ and $Y(\cdot)$ are left continuous, it is easily seen that $\mathcal{Z}_j$ and $\mathcal{Y}$ are pointwise measurable (PM) (Kosorok, 2008, p. 142). Using the preservation results for BUEI and PM classes (e.g., Lemma 9.17 of Kosorok, 2008), we see that the class $\mathcal{Y}\mathcal{Z}_j\mathcal{Z}_l$ is both BUEI and PM, with integrable envelope $\sup_{t\in[0,\tau]} Y(t)|Z_j(t)||Z_l(t)|$ by Condition 2.1(iii). Thus, by Theorem 1.3, the class $\mathcal{Y}\mathcal{Z}_j\mathcal{Z}_l$ is Glivenko–Cantelli. Finally, since $R\pi^{-1}$ is bounded by Condition 2.2, the preservation results for Glivenko–Cantelli classes (e.g., Corollary 9.27 of Kosorok, 2008) imply that $\{R\pi^{-1}Y(t)Z_j(t)Z_l(t) : t \in [0, \tau]\}$ is Glivenko–Cantelli.

The Glivenko–Cantelli properties imply that

$$\sup_{t\in[0,\tau]} \|\mathbf{S}_{\mathrm{SW}}^{(k)}(t) - \mathbf{s}^{(k)}(t)\| = o_p(1), \qquad k = 0, 1, 2.$$

Consequently, since $S_{\mathrm{SW}}^{(0)}(\cdot)$ is bounded away from zero by Condition 2.1(ii), we have

$$\sup_{t\in[0,\tau]} \|\overline{\mathbf{Z}}_{\mathrm{SW}}(t) - \mathbf{e}(t)\| = o_p(1).$$

Next, since $\mathbf{U}_{\text{SW}}(\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, we have

$$\mathbf{0} = \mathbf{U}_{\text{SW}}(\widehat{\boldsymbol{\beta}}_{\text{SW}}) = \mathbf{U}_{\text{SW}}(\boldsymbol{\beta}_0) - \mathbf{V}_{\text{SW}}(\widehat{\boldsymbol{\beta}}_{\text{SW}} - \boldsymbol{\beta}_0),$$

or

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{SW}} - \boldsymbol{\beta}_0) = \mathbf{V}_{\text{SW}}^{-1}\sqrt{n}\mathbf{U}_{\text{SW}}(\boldsymbol{\beta}_0), \tag{2.7}$$

where

$$\mathbf{V}_{\text{SW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\int_0^{\tau} Y_i(t)\{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\text{SW}}(t)\}^{\otimes 2}\,dt.$$

We now derive an asymptotic representation for $\sqrt{n}\mathbf{U}_{\text{SW}}(\boldsymbol{\beta}_0)$. Adding and subtracting terms gives

$$\begin{aligned}
\sqrt{n}\mathbf{U}_{\text{SW}}(\boldsymbol{\beta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\int_0^{\tau}\{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\text{SW}}(t)\}\,dM_i(t)\\
&= \mathbb{G}_n\left[\frac{R}{\pi}\int_0^{\tau}\{\mathbf{Z}(t) - \mathbf{e}(t)\}\,dM(t)\right]\\
&\quad + \mathbb{G}_n\left[\frac{R}{\pi}\int_0^{\tau}\{\mathbf{e}(t) - \overline{\mathbf{Z}}_{\text{SW}}(t)\}\,dM(t)\right]\\
&\quad + \sqrt{n}P\left[\frac{R}{\pi}\int_0^{\tau}\{\mathbf{Z}(t) - \mathbf{e}(t)\}\,dM(t)\right]\\
&\quad + \sqrt{n}P\left[\frac{R}{\pi}\int_0^{\tau}\{\mathbf{e}(t) - \overline{\mathbf{Z}}_{\text{SW}}(t)\}\,dM(t)\right].
\end{aligned} \tag{2.8}$$

First, note that the third term is zero because

$$\begin{aligned}
P\left[\frac{R}{\pi}\int_0^{\tau}\{\mathbf{Z}(t) - \mathbf{e}(t)\}\,dM(t)\right] &= E\left[\frac{1}{\pi}\int_0^{\tau}\{\mathbf{Z}(t) - \mathbf{e}(t)\}\,dM(t)E(R \mid X, \Delta, \mathbf{Z})\right]\\
&= E\left[\int_0^{\tau}\{\mathbf{Z}(t) - \mathbf{e}(t)\}\,dM(t)\right] = E(\mathbf{M}_{\mathbf{Z}}) = 0.
\end{aligned}$$

As discussed in Section 1.3.2, the second term in (2.8) is an empirical process indexed by a class of random functions with the quantity $\overline{\mathbf{Z}}_{\text{SW}}(\cdot)$ estimated from the entire sample, and in the fourth term $P$ is shorthand for the expectation with $\bar{\mathbf{Z}}_{\mathrm{SW}}(\cdot)$ fixed. Since for any fixed function $\mathbf{f}$: $[0,\tau] \to \mathbb{R}^p$,

$$P\left[\frac{R}{\pi}\int_0^\tau \{\mathbf{e}(t) - \mathbf{f}(t)\}\, dM(t)\right] = P\left[\int_0^\tau \{\mathbf{e}(t) - \mathbf{f}(t)\}\, dM(t)\right] = 0,$$

the fourth term in (2.8) is zero.

The key step is to show that the second term in (2.8) is $o_p(1)$. To this end, let $e_j(\cdot)$ and $\bar{Z}_{\mathrm{SW},j}(\cdot)$ denote the $j$th component of $\mathbf{e}(\cdot)$ and $\bar{\mathbf{Z}}_{\mathrm{SW}}(\cdot)$, respectively. Let $\mathcal{F}_j$ denote the class of functions $f: [0,\tau] \to \mathbb{R}$ that are of uniformly bounded variation and satisfy $\sup_{t\in[0,\tau]} |f(t) - e_j(t)| \leq \delta$ for a constant $\delta$. Further, define the class of functions

$$\mathcal{G}_j = \left\{\int_0^\tau f(t)\, dM(t) \colon f \in \mathcal{F}_j\right\}.$$

By constructing $\|\cdot\|_\infty$-balls centered at piecewise constant functions on a regular grid, one can show that

$$N(\varepsilon, \mathcal{F}_j, \|\cdot\|_\infty) \leq \left(\frac{K}{\varepsilon}\right)^{K'/\varepsilon}.$$

Also, note that, for any $f_1, f_2 \in \mathcal{F}_j$,

$$\left|\int_0^\tau f_1(t)\, dM(t) - \int_0^\tau f_2(t)\, dM(t)\right| \leq \sup_{s\in[0,\tau]} |f_1(s) - f_2(s)| \int_0^\tau |dM(t)|.$$

Applying Theorem 2.7.11 of van der Vaart and Wellner (1996) yields that

$$N_{[]}(2\varepsilon\|F\|_{P,2}, \mathcal{G}_j, L_2(P)) \leq N(\varepsilon, \mathcal{F}_j, \|\cdot\|_\infty),$$

where $F = \int_0^\tau |dM(t)|$. Then the bracketing integral of $\mathcal{G}_j$ is

$$J_{[]}(1, \mathcal{G}_j, L_2(P)) = \int_0^1 \sqrt{\log N_{[]}(\varepsilon, \mathcal{G}_j, L_2(P))}\, d\varepsilon$$

$$\leq \int_0^1 \sqrt{(K'/\varepsilon)\log(K/\varepsilon)}\, d\varepsilon < \infty.$$

Thus, by Theorem 1.2, the class $\mathscr{G}_j$ is Donsker. Consequently, the class

$$\left\{\frac{R}{\pi}\int_0^\tau f(t)\, dM(t) \colon f \in \mathscr{F}_j\right\}$$

is also Donsker. The condition in Lemma 1.1 can be verified as follows:

$$\begin{aligned}
&P\left[\frac{R}{\pi}\int_0^\tau \{e_j(t) - \bar{Z}_{\mathrm{SW},j}(t)\}\, dM(t)\right]^2 \\
&\qquad \leq \sup_{s\in[0,\tau]} |e_j(s) - \bar{Z}_{\mathrm{SW},j}(s)|^2 P\left\{\frac{R}{\pi}\int_0^\tau |dM(t)|\right\}^2 = o_p(1)O_p(1) = o_p(1).
\end{aligned}$$

It follows from Lemma 1.1 that the second term in (2.8) is $o_p(1)$. Then we obtain the asymptotic representation

$$\begin{aligned}
\sqrt{n}\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0) &= \mathbb{G}_n\left[\frac{R}{\pi}\int_0^\tau \{\mathbf{Z}(t) - \mathbf{e}(t)\}\, dM(t)\right] + o_p(1) \\
&= \mathbb{G}_n\left(\frac{R}{\pi}\mathbf{M}_{\mathbf{Z}}\right) + o_p(1).
\end{aligned}$$

Thus, by the central limit theorem, $\sqrt{n}\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0)$ is asymptotically normal with mean zero and covariance matrix

$$E\left(\frac{R^2}{\pi^2}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}\right) = E\left(\frac{1}{\pi}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}\right) = \boldsymbol{\Sigma}_{\mathrm{SW}}.$$

Finally, using the proven facts that $\sup_{t\in[0,\tau]} \|\mathbf{S}_{\mathrm{SW}}^{(k)}(t) - \mathbf{s}^{(k)}(t)\| = o_p(1)$, $k = 0, 1, 2$, we have

$$\mathbf{V}_{\mathrm{SW}} - \mathbf{D} = \int_0^\tau \left\{\mathbf{S}_{\mathrm{SW}}^{(2)}(t) - \frac{\mathbf{S}_{\mathrm{SW}}^{(1)}(t)^{\otimes 2}}{S_{\mathrm{SW}}^{(0)}(t)}\right\} dt - \int_0^\tau \left\{\mathbf{s}^{(2)}(t) - \frac{\mathbf{s}^{(1)}(t)^{\otimes 2}}{s^{(0)}(t)}\right\} dt$$

$$= \int_0^\tau \{\mathbf{S}_{\mathrm{SW}}^{(2)}(t) - \mathbf{s}^{(2)}(t)\}\, dt - \int_0^\tau \left\{ \frac{\mathbf{S}_{\mathrm{SW}}^{(1)}(t)^{\otimes 2}}{S_{\mathrm{SW}}^{(0)}(t)} - \frac{\mathbf{s}^{(1)}(t)^{\otimes 2}}{s^{(0)}(t)} \right\} dt$$

$$= o_p(1).$$

An application of Slutsky's lemma completes the proof. □

*Proof of Theorem 2.2.* We first show that

$$\sup_{t\in[0,\tau]} \|\mathbf{S}_{\mathrm{SW}}^{(k)}(t, \hat{\boldsymbol{\alpha}}) - \mathbf{S}_{\mathrm{SW}}^{(k)}(t, \boldsymbol{\alpha}_0)\| = o_p(1), \qquad k = 0, 1, 2.$$

Consider, for example, the $(j, l)$th entry $S_{\mathrm{SW},jl}^{(2)}(t, \hat{\boldsymbol{\alpha}}) - S_{\mathrm{SW},jl}^{(2)}(t, \boldsymbol{\alpha}_0)$. Note that

$$\partial S_{\mathrm{SW},jl}^{(2)}(t, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha} = -\mathbb{P}_n \left\{ \frac{R}{\pi(\boldsymbol{\alpha})^2} Y(t) Z_j(t) Z_l(t) \pi'(\boldsymbol{\alpha}) \right\}. \tag{2.9}$$

By Condition 2.3(i) and Example 19.8 of van der Vaart (1998), the class $\{\pi'(\boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathcal{A}\}$ is Glivenko–Cantelli. Using arguments similar to those in the proof of Theorem 2.1, we can show that the class $\{R\pi(\boldsymbol{\alpha})^{-2} Y(t) Z_j(t) Z_l(t) \pi'(\boldsymbol{\alpha}) : t \in [0, \tau], \boldsymbol{\alpha} \in \mathcal{A}\}$ is Glivenko–Cantelli. Hence, from (2.9) we have

$$\sup_{t\in[0,\tau],\, \boldsymbol{\alpha}\in\mathcal{A}} \|\partial S_{\mathrm{SW},jl}^{(2)}(t, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}\| = O_p(1).$$

An application of the mean value theorem yields that

$$\sup_{t\in[0,\tau]} |S_{\mathrm{SW},jl}^{(2)}(t, \hat{\boldsymbol{\alpha}}) - S_{\mathrm{SW},jl}^{(2)}(t, \boldsymbol{\alpha}_0)| \le \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| \sup_{t\in[0,\tau],\, \boldsymbol{\alpha}^*\in\mathcal{A}} \|\partial S_{\mathrm{SW},jl}^{(2)}(t, \boldsymbol{\alpha}^*)/\partial \boldsymbol{\alpha}\|$$

$$= o_p(1) O_p(1) = o_p(1).$$

Therefore,

$$\sup_{t\in[0,\tau]}\|\mathbf{S}_{\mathrm{SW}}^{(k)}(t,\widehat{\boldsymbol{\alpha}})-\mathbf{S}_{\mathrm{SW}}^{(k)}(t,\boldsymbol{\alpha}_0)\|=o_p(1),\qquad k=0,1,2.$$

Since from the proof of Theorem 2.1, $\sup_{t\in[0,\tau]}\|\mathbf{S}_{\mathrm{SW}}^{(k)}(t,\boldsymbol{\alpha}_0)-\mathbf{s}^{(k)}(t)\|=o_p(1), k=0,1,2,$ it follows that

$$\sup_{t\in[0,\tau]}\|\mathbf{S}_{\mathrm{SW}}^{(k)}(t,\widehat{\boldsymbol{\alpha}})-\mathbf{s}^{(k)}(t)\|=o_p(1),\qquad k=0,1,2,$$

and consequently,

$$\sup_{t\in[0,\tau]}\|\overline{\mathbf{Z}}_{\mathrm{SW}}(t,\widehat{\boldsymbol{\alpha}})-\mathbf{e}(t)\|=o_p(1).$$

As in the proof of Theorem 2.1, write

$$\sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})-\boldsymbol{\beta}_0\}=\mathbf{V}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})^{-1}\sqrt{n}\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0,\widehat{\boldsymbol{\alpha}}). \tag{2.10}$$

We now derive an asymptotic representation for $\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0,\widehat{\boldsymbol{\alpha}})$. Differentiation and adding and subtracting terms give

$$\begin{aligned}
&\partial\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)/\partial\boldsymbol{\alpha}^T\\
&\quad=\frac{\partial}{\partial\boldsymbol{\alpha}^T}\mathbb{P}_n\left[\frac{R}{\pi(\boldsymbol{\alpha})}\int_0^\tau\{\mathbf{Z}(t)-\overline{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha})\}\,dM(t)\right]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}\\
&\quad=(\mathbb{P}_n-P)\left[\frac{R}{\pi(\boldsymbol{\alpha}_0)}\int_0^\tau\left\{-\frac{\partial}{\partial\boldsymbol{\alpha}^T}\overline{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\right\}dM(t)\right]\\
&\qquad+(\mathbb{P}_n-P)\left[-\frac{R}{\pi(\boldsymbol{\alpha}_0)^2}\int_0^\tau\{\mathbf{Z}(t)-\overline{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\}\,dM(t)\,\pi'(\boldsymbol{\alpha}_0)^T\right]\\
&\qquad+P\left[\frac{R}{\pi(\boldsymbol{\alpha}_0)}\int_0^\tau\left\{-\frac{\partial}{\partial\boldsymbol{\alpha}^T}\overline{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\right\}dM(t)\right]\\
&\qquad+P\left[-\frac{R}{\pi(\boldsymbol{\alpha}_0)^2}\int_0^\tau\{\mathbf{Z}(t)-\overline{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\}\,dM(t)\,\pi'(\boldsymbol{\alpha}_0)^T\right],
\end{aligned} \tag{2.11}$$

where

$$
\begin{aligned}
&\partial \bar{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)/\partial \boldsymbol{\alpha}^T \\
&\quad = \frac{\mathbf{S}_{\mathrm{SW}}^{(1)}(t,\boldsymbol{\alpha}_0)}{S_{\mathrm{SW}}^{(0)}(t,\boldsymbol{\alpha}_0)^2}\mathbb{P}_n\left\{\frac{R}{\pi(\boldsymbol{\alpha}_0)^2}Y(t)\pi'(\boldsymbol{\alpha}_0)^T\right\} - \frac{1}{S_{\mathrm{SW}}^{(0)}(t,\boldsymbol{\alpha}_0)}\mathbb{P}_n\left\{\frac{R}{\pi(\boldsymbol{\alpha}_0)^2}Y(t)\mathbf{Z}(t)\pi'(\boldsymbol{\alpha}_0)^T\right\}.
\end{aligned}
$$

As in the proof of Theorem 2.1, $\mathbb{P}_n$ and $P$ are shorthand for the expectations with the estimated quantity $\bar{\mathbf{Z}}_{\mathrm{SW}}(\cdot,\boldsymbol{\alpha})$ fixed. By verifying that the involved classes of functions are Glivenko–Cantelli, we can show that the first two terms in (2.11) are $o_p(1)$. Also, the third term is zero with fixed $\bar{\mathbf{Z}}_{\mathrm{SW}}(\cdot,\boldsymbol{\alpha})$. It remains to consider the fourth term:

$$
\begin{aligned}
&P\left[-\frac{R}{\pi(\boldsymbol{\alpha}_0)^2}\int_0^\tau \{\mathbf{Z}(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\}\,dM(t)\,\pi'(\boldsymbol{\alpha}_0)^T\right] \\
&\quad = P\left[-\frac{1}{\pi(\boldsymbol{\alpha}_0)}\int_0^\tau \{\mathbf{Z}(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\}\,dM(t)\,\pi'(\boldsymbol{\alpha}_0)^T\right] \\
&\quad = P\left[-\frac{1}{\pi(\boldsymbol{\alpha}_0)}\int_0^\tau \{\mathbf{Z}(t)-\mathbf{e}(t)\}\,dM(t)\,\pi'(\boldsymbol{\alpha}_0)^T\right] \\
&\qquad + P\left[-\frac{1}{\pi(\boldsymbol{\alpha}_0)}\int_0^\tau \{\mathbf{e}(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\}\,dM(t)\,\pi'(\boldsymbol{\alpha}_0)^T\right].
\end{aligned}
\tag{2.12}
$$

By definition, the first term in (2.12) is $-\mathbf{A}$. The second term is $o_p(1)$ because

$$
\begin{aligned}
&\left\|P\left[-\frac{1}{\pi(\boldsymbol{\alpha}_0)}\int_0^\tau \{\mathbf{e}(t)-\bar{\mathbf{Z}}_{\mathrm{SW}}(t,\boldsymbol{\alpha}_0)\}\,dM(t)\,\pi'(\boldsymbol{\alpha}_0)^T\right]\right\| \\
&\qquad \le \sup_{s\in[0,\tau]}\|\mathbf{e}(s)-\bar{\mathbf{Z}}_{\mathrm{SW}}(s,\boldsymbol{\alpha}_0)\|\,P\left\{\frac{1}{\pi(\boldsymbol{\alpha}_0)}\int_0^\tau |dM(t)||\pi'(\boldsymbol{\alpha}_0)^T|\right\} \\
&\qquad = o_p(1)O_p(1) = o_p(1).
\end{aligned}
$$

Thus, we obtain

$$
\partial \mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)/\partial \boldsymbol{\alpha}^T = -\mathbf{A} + o_p(1).
$$

It can be further verified that

$$\sup_{\boldsymbol{\alpha}\in\mathcal{A}} \|\partial^2 \mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\boldsymbol{\alpha}^T\| = O_p(1).$$

A Taylor expansion then gives

$$\sqrt{n}\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\alpha}}) = \sqrt{n}\mathbf{U}_{\mathrm{SW}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0) - \mathbf{A}\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1). \tag{2.13}$$

Finally, using the proven facts that $\sup_{t\in[0,\tau]} \|\mathbf{S}^{(k)}_{\mathrm{SW}}(t, \widehat{\boldsymbol{\alpha}}) - \mathbf{S}^{(k)}_{\mathrm{SW}}(t, \boldsymbol{\alpha}_0)\| = o_p(1)$, $k = 0, 1, 2$, we have

$$\begin{aligned}
&\mathbf{V}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}}) - \mathbf{V}_{\mathrm{SW}}(\boldsymbol{\alpha}_0)\\
&\quad = \int_0^\tau \left\{\mathbf{S}^{(2)}_{\mathrm{SW}}(t, \widehat{\boldsymbol{\alpha}}) - \frac{\mathbf{S}^{(1)}_{\mathrm{SW}}(t, \widehat{\boldsymbol{\alpha}})^{\otimes 2}}{S^{(0)}_{\mathrm{SW}}(t, \widehat{\boldsymbol{\alpha}})}\right\} dt - \int_0^\tau \left\{\mathbf{S}^{(2)}_{\mathrm{SW}}(t, \boldsymbol{\alpha}_0) - \frac{\mathbf{S}^{(1)}_{\mathrm{SW}}(t, \boldsymbol{\alpha}_0)^{\otimes 2}}{S^{(0)}_{\mathrm{SW}}(t, \boldsymbol{\alpha}_0)}\right\} dt\\
&\quad = \int_0^\tau \{\mathbf{S}^{(2)}_{\mathrm{SW}}(t, \widehat{\boldsymbol{\alpha}}) - \mathbf{S}^{(2)}_{\mathrm{SW}}(t, \boldsymbol{\alpha}_0)\}\, dt - \int_0^\tau \left\{\frac{\mathbf{S}^{(1)}_{\mathrm{SW}}(t, \widehat{\boldsymbol{\alpha}})^{\otimes 2}}{S^{(0)}_{\mathrm{SW}}(t, \widehat{\boldsymbol{\alpha}})} - \frac{\mathbf{S}^{(1)}_{\mathrm{SW}}(t, \boldsymbol{\alpha}_0)^{\otimes 2}}{S^{(0)}_{\mathrm{SW}}(t, \boldsymbol{\alpha}_0)}\right\} dt\\
&\quad = o_p(1).
\end{aligned}$$

Since from the proof of Theorem 2.1, $\mathbf{V}_{\mathrm{SW}}(\boldsymbol{\alpha}_0) - \mathbf{D} = o_p(1)$, it follows that

$$\mathbf{V}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}}) - \mathbf{D} = o_p(1). \tag{2.14}$$

Substituting (2.13) and (2.14) back into (2.10), we have

$$\sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}}) - \boldsymbol{\beta}_0\} = \sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{SW}} - \boldsymbol{\beta}_0) - \mathbf{D}^{-1}\mathbf{A}\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + o_p(1).$$

By Condition 2.3(ii), an application of the result of Pierce (1982) completes the proof. □

### 2.7.3 Proofs for Section 2.2.2

*Proof of Theorem 2.3.* By substituting

$$dN_i(u) = dM_i(u) + Y_i(u)\,d\Lambda_0(u) + Y_i(u)\boldsymbol{\beta}_0^T \mathbf{Z}_i(u)\,du$$

into the expression of $\widehat{\Lambda}_0(t)$, we have

$$\begin{aligned}
&\sqrt{n}\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_i}{\pi_i} \int_0^t \frac{dN_i(u)}{S_{\text{SW}}^{(0)}(u)} - \sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{SW}}^T \int_0^t \overline{\mathbf{Z}}_{\text{SW}}(u)\,du - \sqrt{n}\Lambda_0(t) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_i}{\pi_i} \int_0^t \frac{dM_i(u)}{S_{\text{SW}}^{(0)}(u)} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_i}{\pi_i} \int_0^t \frac{Y_i(u)}{S_{\text{SW}}^{(0)}(u)}\,d\Lambda_0(u) \\
&\qquad + \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_i}{\pi_i} \int_0^t \frac{Y_i(u)}{S_{\text{SW}}^{(0)}(u)} \boldsymbol{\beta}_0^T \mathbf{Z}_i(u)\,du - \sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{SW}}^T \int_0^t \overline{\mathbf{Z}}_{\text{SW}}(u)\,du - \sqrt{n}\Lambda_0(t) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_i}{\pi_i} \int_0^t \frac{dM_i(u)}{S_{\text{SW}}^{(0)}(u)} + \sqrt{n}\Lambda_0(t) + \sqrt{n}\boldsymbol{\beta}_0^T \int_0^t \overline{\mathbf{Z}}_{\text{SW}}(u)\,du \\
&\qquad - \sqrt{n}\widehat{\boldsymbol{\beta}}_{\text{SW}}^T \int_0^t \overline{\mathbf{Z}}_{\text{SW}}(u)\,du - \sqrt{n}\Lambda_0(t) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{R_i}{\pi_i} \int_0^t \frac{dM_i(u)}{S_{\text{SW}}^{(0)}(u)} - \sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{SW}} - \boldsymbol{\beta}_0)^T \int_0^t \overline{\mathbf{Z}}_{\text{SW}}(u)\,du.
\end{aligned}$$

Adding and subtracting terms gives

$$\begin{aligned}
&\sqrt{n}\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\} \\
&\quad = \mathbb{G}_n\left[\frac{R}{\pi} \int_0^t \left\{\frac{dM(u)}{S_{\text{SW}}^{(0)}(u)} - \frac{dM(u)}{s^{(0)}(u)}\right\}\right] + \mathbb{G}_n\left\{\frac{R}{\pi} \int_0^t \frac{dM(u)}{s^{(0)}(u)}\right\} \\
&\qquad + \sqrt{n}P\left[\frac{R}{\pi} \int_0^t \left\{\frac{dM(u)}{S_{\text{SW}}^{(0)}(u)} - \frac{dM(u)}{s^{(0)}(u)}\right\}\right] + \sqrt{n}P\left\{\frac{R}{\pi} \int_0^t \frac{dM(u)}{s^{(0)}(u)}\right\} \qquad (2.15) \\
&\qquad - \sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{SW}} - \boldsymbol{\beta}_0)^T \int_0^t \{\overline{\mathbf{Z}}_{\text{SW}}(u) - \mathbf{e}(u)\}\,du - \sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{SW}} - \boldsymbol{\beta}_0)^T \int_0^t \mathbf{e}(u)\,du.
\end{aligned}$$

Using arguments similar to those in the proof of Theorem 2.1, we can show that the first and fifth terms in (2.15) are $o_p(1)$ uniformly in $t \in [0, \tau]$, and the third and fourth terms are zero. Also, from the proof of Theorem 2.1, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{SW}} - \boldsymbol{\beta}_0) = \mathbf{D}^{-1}\mathbb{G}_n\left(\frac{R}{\pi}\mathbf{M}_{\mathbf{Z}}\right) + o_p(1).$$

Substituting into (2.15), we obtain the asymptotic representation

$$\sqrt{n}\{\widehat{\Lambda}_0(t) - \Lambda_0(t)\} = \mathbb{G}_n\left\{\frac{R}{\pi}M_Y(t)\right\} - \mathbf{h}(t)^T\mathbf{D}^{-1}\mathbb{G}_n\left(\frac{R}{\pi}\mathbf{M}_{\mathbf{Z}}\right) + o_p(1),$$

where $o_p(1)$ is uniform in $t \in [0, \tau]$. Thus, $\sqrt{n}\{\widehat{\Lambda}(t) - \Lambda_0(t)\}$ converges weakly in $\ell^\infty[0, \tau]$ to a zero-mean Gaussian process with covariance function

$$\begin{aligned} C_{\mathrm{SW}}(s,t) = {} & E\left\{\frac{1}{\pi}M_Y(s)M_Y(t)\right\} + \mathbf{h}(s)^T\mathbf{D}^{-1}E\left(\frac{1}{\pi}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}\right)\mathbf{D}^{-1}\mathbf{h}(t) \\ & - \mathbf{h}(s)^T\mathbf{D}^{-1}E\left\{\frac{1}{\pi}M_Y(t)\mathbf{M}_{\mathbf{Z}}\right\} - \mathbf{h}(t)^T\mathbf{D}^{-1}E\left\{\frac{1}{\pi}M_Y(s)\mathbf{M}_{\mathbf{Z}}\right\} \\ = {} & B(s,t) + \mathbf{h}(s)^T\mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathrm{SW}}\mathbf{D}^{-1}\mathbf{h}(t) - \mathbf{h}(s)^T\mathbf{D}^{-1}\mathbf{k}(t) - \mathbf{h}(t)^T\mathbf{D}^{-1}\mathbf{k}(s). \quad \square \end{aligned}$$

*Proof of Theorem 2.4.* As in the proof of Theorem 2.3, we can write

$$\begin{aligned} & \sqrt{n}\{\widehat{\Lambda}_0(t, \widehat{\boldsymbol{\alpha}}) - \Lambda_0(t)\} \\ & \quad = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^t \frac{dM_i(u)}{S_{\mathrm{SW}}^{(0)}(u, \widehat{\boldsymbol{\alpha}})} - \sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}}) - \boldsymbol{\beta}_0\}^T\int_0^t \overline{\mathbf{Z}}_{\mathrm{SW}}(u, \widehat{\boldsymbol{\alpha}})\,du \\ & \quad = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^t \left\{\frac{dM_i(u)}{S_{\mathrm{SW}}^{(0)}(u, \widehat{\boldsymbol{\alpha}})} - \frac{dM_i(u)}{s^{(0)}(u)}\right\} + \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^t \frac{dM_i(u)}{s^{(0)}(u)} \\ & \qquad - \sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}}) - \boldsymbol{\beta}_0\}^T\int_0^t \{\overline{\mathbf{Z}}_{\mathrm{SW}}(u, \widehat{\boldsymbol{\alpha}}) - \mathbf{e}(u)\}\,du \qquad (2.16) \\ & \qquad - \sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}}) - \boldsymbol{\beta}_0\}^T\int_0^t \mathbf{e}(u)\,du. \end{aligned}$$

Using the facts that $\sup_{t\in[0,\tau]}|S_{\mathrm{SW}}^{(0)}(t,\widehat{\boldsymbol{\alpha}})-s^{(0)}(t)|=o_p(1)$ and $\sup_{t\in[0,\tau]}\|\overline{\mathbf{Z}}_{\mathrm{SW}}(t,\widehat{\boldsymbol{\alpha}})-\mathbf{e}(t)\|=o_p(1)$ from the proof of Theorem 2.2, we can show that the first and third terms in (2.16) are $o_p(1)$ uniformly in $t\in[0,\tau]$. A Taylor expansion of the second term gives

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}\\&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\boldsymbol{\alpha}_0)}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}\\&\qquad-\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\boldsymbol{\alpha}_0)^2}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}\pi_i'(\boldsymbol{\alpha}_0)^T\sqrt{n}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)+o_p(1).\end{aligned}\tag{2.17}$$

Also, note that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\boldsymbol{\alpha}_0)^2}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}\pi_i'(\boldsymbol{\alpha}_0)\\&\quad=\mathbb{P}_n\left\{\frac{R}{\pi(\boldsymbol{\alpha}_0)^2}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}\pi_i'(\boldsymbol{\alpha}_0)\right\}=P\left\{\frac{R}{\pi(\boldsymbol{\alpha}_0)^2}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}\pi_i'(\boldsymbol{\alpha}_0)\right\}+o_p(1)\\&\quad=P\left\{\frac{1}{\pi(\boldsymbol{\alpha}_0)}M_Y(t)\pi_i'(\boldsymbol{\alpha}_0)\right\}+o_p(1)=\mathbf{m}(t)+o_p(1).\end{aligned}$$

Substituting back into (2.17), we then have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\widehat{\boldsymbol{\alpha}})}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}\\&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\boldsymbol{\alpha}_0)}\int_0^t\frac{dM_i(u)}{s^{(0)}(u)}-\mathbf{m}(t)^T\sqrt{n}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)+o_p(1).\end{aligned}\tag{2.18}$$

From the proof of Theorem 2.2, we have

$$\sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})-\boldsymbol{\beta}_0\}=\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}-\boldsymbol{\beta}_0)-\mathbf{D}^{-1}\mathbf{A}\sqrt{n}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)+o_p(1).\tag{2.19}$$

Substituting (2.18) and (2.19) back into the second and fourth terms in (2.16), respectively, and comparing with the representation of $\sqrt{n}\{\widehat{\Lambda}(t)-\Lambda_0(t)\}$, we obtain

$$\begin{aligned}&\sqrt{n}\{\widehat{\Lambda}_0(t,\widehat{\boldsymbol{\alpha}})-\Lambda_0(t)\}\\&\qquad=\sqrt{n}\{\widehat{\Lambda}_0(t)-\Lambda_0(t)\}-\{\mathbf{m}(t)^T-\mathbf{h}(t)^T\mathbf{D}^{-1}\mathbf{A}\}\sqrt{n}(\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)+o_p(1).\end{aligned}$$

The proof is complete by applying the result of Pierce (1982). □

### 2.7.4 Proofs for Section 2.3.2

*Proof of Theorem 2.5.* First, using arguments similar to those in the proof of Theorem 2.1, we can show that $\sup_{t\in[0,\tau]}\|\mathbf{S}_{\mathrm{SW}}^{(k)}(t)-\mathbf{s}^{(k)}(t)\|=o_p(1)$, $k=0,1,2$, and

$$\sup_{t\in[0,\tau]}\|\bar{\mathbf{Z}}_{\mathrm{SW}}(t)-\mathbf{e}(t)\|=o_p(1).$$

Similarly, write

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}-\boldsymbol{\beta}_0)=\mathbf{V}_{\mathrm{AW}}^{-1}\sqrt{n}\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0),$$

where

$$\begin{aligned}\mathbf{V}_{\mathrm{AW}}=&\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\int_0^\tau Y_i(t)\{\mathbf{Z}_i(t)-\bar{\mathbf{Z}}_{\mathrm{AW}}(t)\}^{\otimes 2}\,dt\\&-\frac{1}{n}\sum_{i=1}^{n}\frac{R_i-\pi_i}{\pi_i}\int_0^\tau Y_i(t)[E\{\mathbf{Z}_i(t)^{\otimes 2}\,|\,\mathbf{W}_i\}-\bar{\mathbf{Z}}_{\mathrm{AW}}(t)E\{\mathbf{Z}_i(t)^T\,|\,\mathbf{W}_i\}]\,dt.\end{aligned}$$

We then follow the same lines of arguments to show that

$$\sqrt{n}\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0)=\mathbb{G}_n\left\{\frac{R}{\pi}\mathbf{M}_{\mathbf{Z}}-\frac{R-\pi}{\pi}E(\mathbf{M}_{\mathbf{Z}}\,|\,\mathbf{W})\right\}+o_p(1)$$

and $\mathbf{V}_{\text{AW}} - \mathbf{D} = o_p(1)$. The asymptotic normality follows from the central limit theorem and Slutsky's lemma. The proof is complete by noting that

$$\begin{aligned}
&\operatorname{Var}\left\{\frac{R}{\pi}\mathbf{M}_{\mathbf{Z}} - \frac{R-\pi}{\pi}E(\mathbf{M}_{\mathbf{Z}} \mid \mathbf{W})\right\} \\
&\quad = E\left(\frac{1}{\pi^2}\mathbf{M}_{\mathbf{Z}}^{\otimes 2}\right) - 2E\left[\frac{1-\pi}{\pi}\{E(\mathbf{M}_{\mathbf{Z}} \mid \mathbf{W})\}^{\otimes 2}\right] + E\left[\frac{1-\pi}{\pi}\{E(\mathbf{M}_{\mathbf{Z}} \mid \mathbf{W})\}^{\otimes 2}\right] \\
&\quad = \boldsymbol{\Sigma}_{\text{SW}} - E\left[\frac{1-\pi}{\pi}\{E(\mathbf{M}_{\mathbf{Z}} \mid \mathbf{W})\}^{\otimes 2}\right] = \boldsymbol{\Sigma}_{\text{AW}}. \qquad \square
\end{aligned}$$

*Proof of Theorem 2.6.* First, using arguments similar to those in the proof of Theorem 2.2, we can show that $\sup_{t\in[0,\tau]} \|\mathbf{S}_{\text{SW}}^{(k)}(t, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) - \mathbf{s}^{(k)}(t)\| = o_p(1)$, $k = 0, 1, 2$, and

$$\sup_{t\in[0,\tau]} \|\overline{\mathbf{Z}}_{\text{SW}}(t, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) - \mathbf{e}(t)\| = o_p(1).$$

Similarly, write

$$\sqrt{n}\{\widehat{\boldsymbol{\beta}}_{\text{AW}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}_0\} = \mathbf{V}_{\text{AW}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})^{-1}\sqrt{n}\mathbf{U}_{\text{AW}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}).$$

We now derive an asymptotic representation for $\sqrt{n}\mathbf{U}_{\text{AW}}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$. By differentiation and adding and subtracting terms, we have

$$\begin{aligned}
&\partial\mathbf{S}_{\text{AW}}^{(k)}(t, \boldsymbol{\alpha}, \boldsymbol{\gamma})/\partial\boldsymbol{\alpha}^T \\
&\quad = -\mathbb{P}_n\left\{\frac{R}{\pi(\boldsymbol{\alpha})^2}Y(t)\mathbf{Z}(t)^{\otimes k}\pi'(\boldsymbol{\alpha})^T\right\} + \mathbb{P}_n\left[\frac{R}{\pi(\boldsymbol{\alpha})^2}Y(t)E_{\boldsymbol{\gamma}}\{\mathbf{Z}(t)^{\otimes k} \mid \mathbf{W}\}\pi'(\boldsymbol{\alpha})^T\right] \\
&\quad = -(\mathbb{P}_n - P)\left\{\frac{R}{\pi(\boldsymbol{\alpha})^2}Y(t)\mathbf{Z}(t)^{\otimes k}\pi'(\boldsymbol{\alpha})^T\right\} \\
&\qquad + (\mathbb{P}_n - P)\left[\frac{R}{\pi(\boldsymbol{\alpha})^2}Y(t)E_{\boldsymbol{\gamma}}\{\mathbf{Z}(t)^{\otimes k} \mid \mathbf{W}\}\pi'(\boldsymbol{\alpha})^T\right] \qquad (2.20) \\
&\qquad + P\left\{\frac{\pi(\boldsymbol{\alpha}_0)}{\pi(\boldsymbol{\alpha})^2}Y(t)[E_{\boldsymbol{\gamma}}\{\mathbf{Z}(t)^{\otimes k} \mid \mathbf{W}\} - E\{\mathbf{Z}(t)^{\otimes k} \mid \mathbf{W}\}]\pi'(\boldsymbol{\alpha})^T\right\}, \qquad k = 0, 1.
\end{aligned}$$

By verifying that the involved classes of functions are Donsker, we can show that the first two terms in (2.20) are $O_p(n^{-1/2})$. Also, it follows from the mean value theorem that the third term is $O_p(n^{-\kappa_2})$. Therefore, we have

$$\sup_{t\in[0,\tau],\,\boldsymbol{\alpha}\in\mathcal{A},\,\boldsymbol{\gamma}\in\Gamma} \|\partial \mathbf{S}_{\mathrm{AW}}^{(k)}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\alpha}^T\| = O_p(n^{-\kappa_2}), \qquad k=0,1,$$

and hence

$$\sup_{t\in[0,\tau],\,\boldsymbol{\alpha}\in\mathcal{A},\,\boldsymbol{\gamma}\in\Gamma} \|\partial \overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\alpha}^T\| = O_p(n^{-\kappa_2}). \tag{2.21}$$

Next, write

$$\begin{aligned}
&\partial\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\alpha}^T\\
&\quad= -(\mathbb{P}_n-P)\left[\frac{R}{\pi(\boldsymbol{\alpha})^2}\int_0^\tau\{\mathbf{Z}(t)-\overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})\}\,dM(t)\,\pi'(\boldsymbol{\alpha})^T\right]\\
&\qquad+(\mathbb{P}_n-P)\left\{\frac{R}{\pi(\boldsymbol{\alpha})^2}\int_0^\tau[E_{\boldsymbol{\gamma}}\{\mathbf{Z}(t)\,dM(t)\,|\,\mathbf{W}\}\right.\\
&\qquad\qquad\left.-\overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})E_{\boldsymbol{\gamma}}\{dM(t)\,|\,\mathbf{W}\}]\pi'(\boldsymbol{\alpha})^T\right\}\\
&\qquad-(\mathbb{P}_n-P)\left\{\frac{R}{\pi(\boldsymbol{\alpha})}\int_0^\tau\frac{\partial}{\partial\boldsymbol{\alpha}^T}\overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})\,dM(t)\right\}\\
&\qquad+(\mathbb{P}_n-P)\left[\frac{R-\pi(\boldsymbol{\alpha})}{\pi(\boldsymbol{\alpha})}\int_0^\tau\frac{\partial}{\partial\boldsymbol{\alpha}^T}\overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})E_{\boldsymbol{\gamma}}\{dM(t)\,|\,\mathbf{W}\}\right]\\
&\qquad+P\left\{\frac{\pi(\boldsymbol{\alpha}_0)}{\pi(\boldsymbol{\alpha})^2}\int_0^\tau[E_{\boldsymbol{\gamma}}\{\mathbf{Z}(t)\,dM(t)\,|\,\mathbf{W}\}-E\{\mathbf{Z}(t)\,dM(t)\,|\,\mathbf{W}\}]\pi'(\boldsymbol{\alpha})^T\right\}\\
&\qquad-P\left\{\frac{\pi(\boldsymbol{\alpha}_0)}{\pi(\boldsymbol{\alpha})^2}\int_0^\tau\overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})[E_{\boldsymbol{\gamma}}\{dM(t)\,|\,\mathbf{W}\}-E\{dM(t)\,|\,\mathbf{W}\}]\pi'(\boldsymbol{\alpha})^T\right\}\\
&\qquad-P\left\{\frac{R}{\pi(\boldsymbol{\alpha})}\int_0^\tau\frac{\partial}{\partial\boldsymbol{\alpha}^T}\overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})\,dM(t)\right\}\\
&\qquad+P\left[\frac{R-\pi(\boldsymbol{\alpha})}{\pi(\boldsymbol{\alpha})}\int_0^\tau\frac{\partial}{\partial\boldsymbol{\alpha}^T}\overline{\mathbf{Z}}_{\mathrm{AW}}(t,\boldsymbol{\alpha},\boldsymbol{\gamma})E_{\boldsymbol{\gamma}}\{dM(t)\,|\,\mathbf{W}\}\right].
\end{aligned}$$

The first four terms are easily shown to be $O_p(n^{-1/2})$ by verifying that the involved classes of functions are Donsker. By the mean value theorem, the fifth and sixth terms are $O_p(n^{-\kappa_2})$.

The last two terms are $O_p(n^{-\kappa_2})$ by (2.21). Then we have

$$\sup_{\boldsymbol{\alpha}\in\mathcal{A},\,\boldsymbol{\gamma}\in\Gamma} \|\partial \mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\alpha}^T\| = O_p(n^{-\kappa_2}).$$

We can similarly show that

$$\sup_{\boldsymbol{\alpha}\in\mathcal{A},\,\boldsymbol{\gamma}\in\Gamma} \|\partial \mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^T\| = O_p(n^{-\kappa_1}).$$

An application of the mean value theorem yields that

$$\begin{aligned}\sqrt{n}\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0,\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}}) &= \sqrt{n}\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) + \sqrt{n}\,O_p(n^{-(\kappa_1+\kappa_2)})\\ &= \sqrt{n}\mathbf{U}_{\mathrm{AW}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) + o_p(1).\end{aligned}$$

Finally, as before, we can show that $\mathbf{V}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}}) - \mathbf{D} = o_p(1)$. It follows that $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}})$ is asymptotically normal with the same asymptotic variance as $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}$.

To see the optimality of $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}})$, consider a class of estimators with the asymptotic representation

$$\mathbb{G}_n\left\{\frac{R}{\pi(\widehat{\boldsymbol{\alpha}})}\mathbf{M}_{\mathbf{Z}} - \frac{R-\pi(\widehat{\boldsymbol{\alpha}})}{\pi(\widehat{\boldsymbol{\alpha}})}c(\mathbf{W})\right\}$$

for any function $c$. One can proceed as in Example 25.43 of van der Vaart (1998) to show that the asymptotic variance is minimized by taking $c(\mathbf{W}) = E(\mathbf{M}_{\mathbf{Z}}\,|\,\mathbf{W})$. Since $\widehat{\boldsymbol{\beta}}_{\mathrm{SW}}(\widehat{\boldsymbol{\alpha}})$ is also in this class with $c = 0$, its asymptotic variance is no less than that of $\widehat{\boldsymbol{\beta}}_{\mathrm{AW}}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}})$. □

# CHAPTER 3
# High-Dimensional Variable Selection in Additive Hazards Regression

## 3.1 Introduction

Continuing our discussion in Section 1.2.2, in this chapter we study the problem of high-dimensional variable selection in additive hazards regression. We propose a general class of regularized estimators which combine the nonconcave penalized likelihood approach (Fan and Li, 2001) and the pseudoscore method for the additive hazards model (Lin and Ying, 1994). To justify the superior performance of the proposed method, we consider a very general high-dimensional setting, where the dimension of covariates may grow fast, possibly nonpolynomially, with the sample size. Under very mild and easily interpretable conditions, we show that the proposed estimators enjoy the weak oracle property (Lv and Fan, 2009) and the oracle property (Fan and Li, 2001). Our high-dimensional analysis is innovative in that it involves empirical process techniques that have not been previously used in the survival analysis literature, and sheds light on the model selection properties of regularization methods for survival models.

Recently, Bradic, Fan and Jiang (2011) and Huang and Ma (2010) studied regularized estimation for variable selection in the Cox model and the accelerated failure time model, respectively, and obtained some theoretical results in which the dimension of covariates may grow nonpolynomially with the sample size. Besides model assumptions, an important difference from our results, however, is that these authors imposed some *random* conditions on an *empirical* covariance matrix; specifically, these are Condition 8 in Bradic, Fan and Jiang (2011) and Condition (A4) in Huang and Ma (2010). Although such results are useful, it would be very natural to ask the question whether the regularized estimators still enjoy the desired properties if similar conditions are imposed on the *population* version of the matrix. Since the empirical covariance matrix involves the outcome variables, as is generally the case for survival models, *nonrandom* conditions on the population covariance matrix seem to be more natural and will provide more confirmative performance guarantees for the regularized estimators. Such conditions also have the benefit that they can be viewed as high-dimensional extensions of the classical asymptotic regularity conditions in the low-dimensional setting, which are imposed on the population covariance matrix.

The extension from the empirical covariance matrix to its population counterpart turns out to be very nontrivial. Due to high dimensionality and dependency among matrix entries, this is not a direct consequence of classical random matrix theory. A similar difficulty was noted by Ravikumar, Wainwright and Lafferty (2010) in the context of graphical model selection. Our case, however, is more intricate in that the usual Hoeffding's inequality is not applicable, since each entry of the covariance matrix in question, to be defined in Section 3.3, is not an independent sum. Thus, we resort to a functional Hoeffding-type inequality (Lemma 1.4) and some empirical process techniques, and first establish some concentration results that will be useful in our proofs. Although we focus on the additive hazards model in this dissertation, the ideas could be extended, for instance, to the Cox model and the accelerated failure time model.

In Section 3.2 we propose a class of regularized estimators and discuss choices of the penalty function. The theoretical properties of these estimators are studied in Section 3.3.

In Section 3.4 we describe an implementation of the proposed estimators by an iterative coordinate descent algorithm. Simulation studies and a real data example are presented in Sections 3.5 and 3.6, respectively. Some discussion is offered in Section 3.7, and all proofs are deferred to Section 3.8.

## 3.2 Method

### 3.2.1 Regularized Estimators

We begin with the problem setup. Let $T$ denote the failure time and $C$ the censoring time. We observe the censored failure time $X = T \wedge C$ and the failure indicator $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Let $\mathbf{Z}(\cdot)$ be a $p$-vector of predictable covariate processes and assume that $T$ and $C$ are conditionally independent given $\mathbf{Z}(\cdot)$. The data consist of $(X_i, \Delta_i, \mathbf{Z}_i(\cdot))$, $i = 1, \ldots, n$, which are independent copies of $(X, \Delta, \mathbf{Z}(\cdot))$. We consider the additive hazards model

$$\lambda(t \,|\, \mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^T \mathbf{Z}(t), \tag{3.1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0$ is a $p$-vector of regression coefficients.

We adopt the usual counting process notation. Define the observed-failure counting process $N_i(t) = I(X_i \leq t, \Delta_i = 1)$, the at-risk indicator $Y_i(t) = I(X_i \geq t)$, and the counting process martingale

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\{\lambda_0(s) + \boldsymbol{\beta}_0^T \mathbf{Z}_i(s)\}\, ds.$$

We will also use $N(t)$, $Y(t)$, and $M(t)$ to denote the generic versions of these processes.

The pseudoscore function of Lin and Ying (1994) is defined as

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\beta}^T\mathbf{Z}_i(t)\,dt\},$$

where

$$\bar{\mathbf{Z}}(t) = \frac{\sum_{j=1}^{n} Y_j(t)\mathbf{Z}_j(t)}{\sum_{j=1}^{n} Y_j(t)},$$

and $\tau$ is the maximum follow-up time. This estimating function is linear in $\boldsymbol{\beta}$; through some algebraic manipulation, we can write $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{b} - \mathbf{V}\boldsymbol{\beta}$, where

$$\mathbf{b} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}\,dN_i(t)$$

and

$$\mathbf{V} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\,dt. \tag{3.2}$$

Since $\mathbf{V}$ is positive semidefinite, integrating $-\mathbf{U}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ leads to the least-squares-type loss function

$$L(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta} - \mathbf{b}^T\boldsymbol{\beta}.$$

Using this loss function as the basis of regularized estimators in model (3.1) has been suggested by a number of authors, e.g., Leng and Ma (2007) and Martinussen and Scheike (2009). The latter authors also noted that $L(\boldsymbol{\beta})$ has an appealing interpretation that it is the empirical prediction error, up to a constant, for the part of the model orthogonal to the at-risk indicator.

We now define the regularized estimator $\widehat{\boldsymbol{\beta}}$ as the solution to the problem

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{L(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|)\right\}, \tag{3.3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and $p_\lambda(\cdot)$ is a penalty function that depends on the regularization parameter $\lambda > 0$, which controls the strength of regularization. When the minimization problem (3.3) is not convex, we will consider a local minimizer. It is often convenient to rewrite the penalty function as $p_\lambda(\cdot) = \lambda\rho_\lambda(\cdot)$. Without the penalty term, $\widehat{\boldsymbol{\beta}}$ reduces to the pseudoscore estimator of Lin and Ying (1994). When dimensionality is high, however, some form of regularization is required, and the performance of the regularized estimator depends critically on the choice of the penalty function. Thus, in the following we will first define a general class of penalty functions and discuss several popular choices among the class, and then present some theory to gain further insight into these choices.

### 3.2.2 Penalty Functions

To answer the question on what kind of penalty functions are ideal for model selection, Fan and Li (2001) advocated penalty functions giving rise to estimators with three desired properties: sparsity, unbiasedness, and continuity. These properties have motivated consideration of the class of penalty functions that satisfy the following condition.

**Condition 3.1.** The function $\rho_\lambda(\theta)$ is increasing and concave in $\theta \in [0, \infty)$, and has a continuous derivative $\rho_\lambda'(\theta)$ on $(0, \infty)$. In addition, $\rho_\lambda'(\theta)$ is increasing in $\lambda$ and $\rho'(0+) > 0$ is independent of $\lambda$.

Some intuition for Condition 3.1 is as follows: The singularity at the origin encourages sparsity; the concavity assumption aims to reduce the estimation bias; the requirement that $\rho_\lambda'(\theta)$ is increasing in $\lambda$ allows $\lambda$ to effectively control the overall strength of the penalty. It should be noted that we do *not* require *strict* concavity or monotonicity, so that a wide range of penalty functions, including those that do not lead to all of the aforementioned three properties, are included in this class, which will facilitate our comparisons between different penalty functions. In the contexts of (generalized) linear models, this class of penalty functions have been studied by Lv and Fan (2009) and Fan and Lv (2011). Of particular interest are the following examples.

- The lasso (Tibshirani, 1996) uses the $L_1$ penalty, i.e., $\rho(\theta) = \theta, \theta \geq 0$.

- The *smoothly clipped absolute deviation* (SCAD) penalty (Fan, 1997; Fan and Li, 2001) is given by the derivative

  $$\rho'_\lambda(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda), \qquad \theta \geq 0,$$

  where $a > 2$ is a shape parameter. The penalty function takes off at the origin as the $L_1$ penalty and then gradually levels off until its derivative reaches zero.

- The *minimax concave penalty* (MCP) proposed by Zhang (2010) has the derivative

  $$\rho'_\lambda(\theta) = \frac{(a\lambda - \theta)_+}{a\lambda}, \qquad \theta \geq 0,$$

  where $a > 1$ is a shape parameter. In a similar spirit to SCAD, the penalty function gradually decreases its derivative to zero, except that it drops the $L_1$ part of SCAD.

- The *smooth integration of counting and absolute deviation* (SICA) penalty (Lv and Fan, 2009) takes the form

  $$\rho(\theta) = \frac{(a+1)\theta}{a+\theta}, \qquad \theta \geq 0,$$

  where $a > 0$ is a shape parameter. With $a$ varying from 0 to $\infty$, the family can be viewed as a smooth homotopy between the $L_0$ and $L_1$ penalties, with each penalty function pinned at the point $(1, 1)$ and having $\rho'(0+) = 1 + a^{-1}$.

The $L_1$ penalty is a convex relaxation of the $L_0$ penalty and falls at the boundary of the class of penalty functions that satisfy Condition 3.1. Although widely believed to be computationally simple, it suffers from several drawbacks which have motivated a number of improvements. The SCAD penalty was originally proposed to alleviate the bias caused by the $L_1$ approach, and has been shown to enjoy the *oracle property*, i.e., the resulting estimator performs asymptotically as well as the oracle who knows the true submodel in advance. Such procedures are appealing in that they allow simultaneous estimation and variable selection. Still, one can question whether this two-in-one feature is indeed a great advantage, since one would expect that the large estimation bias resulted from a procedure without the oracle property could be remedied by refitting the model with only the selected variables. This is not always possible, however, if some important variables have been incorrectly excluded in the selection step. In fact, the estimation bias can interfere with variable selection; as a result, more stringent conditions such as the irrepresentable condition (Zhao and Yu, 2006) are typically required for consistent variable selection. The advantages of concave penalties regarding model selection consistency have recently been revealed and justified by a number of authors. Zhang (2010) showed that the MCP penalty enjoys certain minimax optimality which enables it to find a balance between the superior theoretical properties of concave penalties and the computational costs of nonconvex optimization problems. By investigating a nonasymptotic weak oracle property, Lv and Fan (2009) showed that the regularity conditions needed for the $L_1$ approach can be substantially relaxed by using concave penalties. The SICA family they proposed has the remarkable feature that it can be used to define a sequence of regularization problems with varying theoretical performance and computational complexity.

## 3.3 Theory

Besides the choice of the penalty function, the performance of the regularized estimators depends on a variety of factors, such as dimensionality of the model, dependency among the covariates, and the choice of the regularization parameter. In order to determine how these factors interact with each other and together affect the performance of the proposed estimators, in this section we rigourously develop a high-dimensional theory and discuss some of its implications.

We begin by introducing some notation to be used in our theoretical results. For any vector $\mathbf{v}$, write $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$, and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. Define

$$\begin{aligned} \mathbf{S}^{(k)}(t) &= \frac{1}{n}\sum_{j=1}^{n} Y_j(t)\mathbf{Z}_j(t)^{\otimes k}, \\ \mathbf{s}^{(k)}(t) &= E\{Y(t)\mathbf{Z}(t)^{\otimes k}\}, \qquad k = 0, 1, 2, \\ \mathbf{e}(t) &= \mathbf{s}^{(1)}(t)/s^{(0)}(t), \\ \mathbf{D} &= E\left[\int_0^{\tau} Y(t)\{\mathbf{Z}(t) - \mathbf{e}(t)\}^{\otimes 2}\, dt\right], \end{aligned}$$

and

$$\mathbf{\Sigma} = E\left[\int_0^{\tau} \{\mathbf{Z}(t) - \mathbf{e}(t)\}^{\otimes 2}\, dN(t)\right].$$

It is worthwhile to note that $\mathbf{D}$ is the population counterpart of the matrix $\mathbf{V}$ defined in (3.2), while $\mathbf{\Sigma}$ is the population counterpart of the matrix

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t)\}^{\otimes 2}\, dN_i(t).$$

These matrices characterize the covariance structure of the model and will play a key role in our high-dimensional analysis.

Furthermore, define the *active set* $A = \{j : \beta_{0j} \neq 0\}$, where $\beta_{0j}$ is the $j$th component of $\boldsymbol{\beta}_0$. Let $s = |A|$, i.e., the number of nonzero coefficients in $\boldsymbol{\beta}_0$, and we allow the dimension triple $(n, p, s)$ to vary freely. Similarly, define the *estimated active set* $\widehat{A} = \{j : \widehat{\beta}_j \neq 0\}$. Denote the complement of a set $B$ by $B^c$. We will use sets to index vectors and matrices; for example, $\boldsymbol{\beta}_{0A}$ is the vector formed by the components $\beta_{0j}$ of $\boldsymbol{\beta}_0$ with $j \in A$, and $\mathbf{D}_{A^cA}$ is the matrix formed by the entries $D_{ij}$ of $\mathbf{D}$ with $i \in A^c$ and $j \in A$. Define the (half) *minimum signal*

$$d = \frac{1}{2} \min_{j \in A} |\beta_{0j}|.$$

For any $\boldsymbol{\theta} \in \mathbb{R}^q$ with $\theta_j \neq 0$ for all $j$, define the *local concavity* of the function $\rho_\lambda(\cdot)$ as

$$\kappa(\rho_\lambda; \boldsymbol{\theta}) = \lim_{\varepsilon \to 0+} \max_{1 \le j \le q} \sup_{|\theta_j| - \varepsilon < t_1 < t_2 < |\theta_j| + \varepsilon} \left\{ -\frac{\rho'_\lambda(t_2) - \rho'_\lambda(t_1)}{t_2 - t_1} \right\}.$$

Finally, define

$$\kappa_0 = \sup\{\kappa(\rho_\lambda; \boldsymbol{\theta}) : \|\boldsymbol{\theta} - \boldsymbol{\beta}_{0A}\| \le d\},$$

$$\varphi = \|\mathbf{D}_{AA}^{-1}\|_\infty,$$

and

$$\mu = \Lambda_{\min}(\mathbf{D}_{AA}) - \lambda \kappa_0,$$

where $\Lambda_{\min}(\cdot)$ is the minimum eigenvalue. It is important to note that all the quantities defined above can depend on the sample size $n$, and we have suppressed this dependency for notational simplicity.

### 3.3.1 Weak Oracle Property

Lv and Fan (2009) introduced the weak oracle property. An estimator is said to have the *weak oracle property* if it is both consistent in model selection and uniformly consistent in estimation. This notion is weaker than the oracle property considered by Fan and Li (2001) and hence can be satisfied by a broader class of estimators. In high-dimensional statistical problems, nonasymptotic results, in which the dimension parameters appear as they are, are often desirable, since they characterize the influence of the parameters explicitly and allow a high-dimensional setting as general as possible. To derive a nonasymptotic result regarding the weak oracle property, we need to impose the following conditions.

**Condition 3.2.** (i) $\int_0^\tau \lambda_0(t)\,dt < \infty$. (ii) $\Pr\{Y(\tau) = 1\} > 0$. (iii) There exist constants $D, K, r > 0$ such that

$$\Pr\left(\sup_{t\in[0,\tau]} |Z_j(t)| > x\right) \le D\exp(-Kx^r)$$

for all $x > 0$ and $j = 1, \dots, p$. (iv) The sample paths of $\mathbf{Z}(\cdot)$ are of uniformly bounded variation.

*Remark* 3.1. In Condition 3.2, parts (i) and (ii) are standard for survival models. For $r \ge 1$, part (iii) is equivalent to saying that $\max_{j=1}^{p} \sup_{t\in[0,\tau]} |Z_j(t)|$ has a finite Orlicz norm $\|\cdot\|_{\psi_r}$ with $\psi_r(x) = e^{x^r} - 1$ (Kosorok, 2008, Lemma 8.1); this condition controls the tail behavior of the covariates and is trivially satisfied for bounded covariates. Part (iv) is a very mild technical assumption that will facilitate entropy calculations.

**Condition 3.3.** There exist constants $\alpha \in (0, 1]$, $\gamma \in [0, 1/2]$, and $c > 0$ such that

$$\|\mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1}\|_\infty \le \left((1-\alpha)\frac{\rho'(0+)}{\rho'_\lambda(d)}\right) \wedge (cn^\gamma).$$

*Remark* 3.2. Condition 3.3 is an analog of Condition (35) in Lv and Fan (2009) for regularized least squares, which is in turn a generalization of condition (15) in Wainwright (2009) for the lasso. Very often for linear regression, such conditions are first imposed on the deterministic Gram matrix, and then a variety of random design matrices such as Gaussian ensembles can be further considered. For survival models such as model (3.1), however, there is no exact analog of the deterministic Gram matrix; here the matrix $\mathbf{V}$, which plays the same role as the Gram matrix in linear regression, involves the at-risk indicators which are *not* deterministic. Thus, we impose conditions directly on the matrix $\mathbf{D}$, which is the population counterpart of the matrix $\mathbf{V}$. Also, we are not restricted to the cases where the covariates are bounded or Gaussian.

*Remark* 3.3. The right-hand side of Condition 3.3 consists of two parts: The first part is an upper bound that reflects the intrinsic capability of the penalty function for variable selection; the second part is at most $O(\sqrt{n})$, where the parameter $\gamma$ needs to be determined by other conditions to be presented later. For the $L_1$ penalty, the first part is bounded by constant 1, which is quite stringent; for concave penalties, the upper bound is generally relaxed, since strict concavity implies that $\rho'(0+) > \rho'_\lambda(d)$. When signals are fairly strong so that $d \gg \lambda$, the first part imposes no constraint for the SCAD and MCP penalties, since $\rho'_\lambda(d) = 0$ in that case. Also, the upper bound for the SICA penalty can be substantially relaxed by choosing a small value of $a$.

Since Condition 3.3 and definitions of $\varphi$ and $\mu$ involve the matrices $\mathbf{D}_{A^c A}\mathbf{D}_{AA}^{-1}$, $\mathbf{D}_{AA}^{-1}$, and $\mathbf{D}_{AA}$, a key step to establishing the weak oracle property is to show that these matrices are close to their empirical counterparts in some sense. This intermediate result is provided by the following lemma, which gives explicit probability bounds for similar conditions to hold for the empirical matrices. In what follows, let $\Omega_L$ denote the event that $\max_{i=1}^{p} \sup |Z_j(t)| \le L$.

**Lemma 3.1.** *Under Conditions 3.1–3.3, if $\mu > 0$ and $\varphi \vee \mu^{-1} = O(\sqrt{n}/s)$, then there exist constants $D, K > 0$ such that*

$$\Pr(\|\mathbf{V}_{AA}^{-1}\|_\infty \ge 2\varphi \mid \Omega_L) \le s^2 D \exp\left\{-K \frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2} \wedge 1\right)\right\}, \tag{3.4}$$

$$\begin{aligned}\Pr\Bigl\{\|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\|_\infty \ge \Bigl(\Bigl(1-\frac{\alpha}{2}\Bigr)\frac{\rho'(0+)}{\rho'_\lambda(d)}\Bigr) \wedge (2cn^\gamma) \mid \Omega_L\Bigr\}&\\ \le (p-s)sD\exp\Bigl\{-K\frac{n}{L^4}\Bigl(\frac{\bigl(\rho'_\lambda(d)^{-1}\wedge n^\gamma\bigr)^2}{\varphi^2 s^2}\wedge 1\Bigr)\Bigr\}&\\ + s^2D\exp\Bigl\{-K\frac{n}{L^4}\Bigl(\frac{1}{\varphi^2 s^2}\wedge 1\Bigr)\Bigr\},&\end{aligned} \tag{3.5}$$

*and*

$$\Pr(\Lambda_{\min}(\mathbf{V}_{AA}) \le \lambda\kappa_0 \mid \Omega_L) \le s^2D\exp\Bigl\{-K\frac{n}{L^4}\Bigl(\frac{\mu^2}{s^2}\wedge 1\Bigr)\Bigr\}. \tag{3.6}$$

*Remark* 3.4. Inequalities (3.4) and (3.5) show that there would not be much difference if we had defined the quantity $\varphi$ or imposed Condition 3.3 on the empirical matrices. The eigenvalue condition $\Lambda_{\min}(\mathbf{V}_{AA}) > \lambda\kappa_0$ is needed for identification of a strict local minimizer of problem (3.3); inequality (3.6) says that this condition holds with high probability if $\Lambda_{\min}(\mathbf{D}_{AA})$ and $\lambda\kappa_0$ have a positive gap $\mu$ that does not shrink to zero too fast.

We now state our main theoretical result regarding the weak oracle property of the proposed estimators.

**Theorem 3.1 (Weak oracle property).** *In addition to Conditions 3.1–3.3, assume that the following conditions hold:*

$$\frac{n\bigl(\rho'_\lambda(d)^{-1}\wedge n^\gamma\bigr)^2}{\varphi^2 s^2(\log p)^{r_1}} \to \infty, \qquad \frac{n(\varphi^{-1}\wedge\mu)^2}{s^2(\log s)^{r_1}} \to \infty, \tag{3.7}$$

$$\frac{n\lambda^2}{(\log p)^{r_1}} \to \infty, \qquad \frac{n^{1-2\gamma}\lambda^2}{(\log s)^{r_1}} \to \infty, \tag{3.8}$$

*and*

$$d \ge c_1\varphi\lambda\rho'(0+), \tag{3.9}$$

*where $\mu > 0$, $r_1 = (r+4)/r$, and $c_1 = 2 + 1/(4c)$. Then, for some constants $D, K > 0$, with probability at least*

$$1 - D\exp\left\{-Kn^{1/r_1}\left(\frac{(\varphi^{-1}\wedge\mu)^2}{s^2}\wedge 1\right)^{1/r_1}\right\} - D\exp\left\{-Kn^{1/r_1}\left(\frac{\lambda^2}{n^{2\gamma}}\wedge 1\right)^{1/r_1}\right\} \to 1,$$

*there exists a regularized estimator $\widehat{\boldsymbol{\beta}}$ that satisfies the following properties:*

*(a) (Sparsity) $\widehat{\boldsymbol{\beta}}_{A^c} = \mathbf{0}$.*

*(b) ($L_\infty$ loss) $\|\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\|_\infty \le c_1\varphi\lambda\rho'(0+)$.*

*Remark* 3.5. To develop intuition for the two conditions in (3.7), we consider some simplified cases. First, concavity implies that $\rho'_\lambda(d) \le \rho'(0+)$; thus, a sufficient condition for the first condition in (3.7) to hold is

$$\frac{n}{\varphi^2 s^2(\log p)^{r_1}} \to \infty. \tag{3.10}$$

Consider the second condition in (3.7) and recall that $\mu = \Lambda_{\min}(\mathbf{D}_{AA}) - \lambda\kappa_0$. For the $L_1$ penalty, $\kappa_0 = 0$; for SCAD and MCP, $\lambda\kappa_0 = (a-1)^{-1}$ and $a^{-1}$, respectively. Thus, for this condition to hold for these penalties, it suffices to assume that $\Lambda_{\min}(\mathbf{D}_{AA})$ is bounded away from zero and that

$$\frac{n}{\varphi^2 s^2(\log s)^{r_1}} \to \infty,$$

where the latter is implied by (3.10). Therefore, conditions in (3.7) are primarily constraints on the growth rates of the model dimensions $p$ and $s$ and certain matrix norms of $\mathbf{D}_{AA}^{-1}$. On the other hand, if we assume, for simplicity, that $\varphi$ is constant, then (3.10) gives a lower bound for the number of observations that are needed for guaranteed sparse recovery, $n \gg s^2(\log p)^{r_1}$. This is an interesting setting, since it shows that the proposed estimators can handle a nonpolynomially growing dimension of covariates as high as $\log p = o(n^{1/r_1})$, while the dimension of the submodel grows as $s = o(\sqrt{n})$. In particular, for bounded covariates, we can take $r_1 = 1$ and thus allow $\log p = o(n)$.

*Remark* 3.6. For simplicity, consider the case of bounded covariates, i.e., $r_1 = 1$. The two conditions in (3.8) give a lower bound for the regularization parameter $\lambda$,

$$\lambda \gg \sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}}.$$

Thus, in view of (3.9), we see that $\lambda$ should be chosen to satisfy

$$\sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}} \ll \lambda \le \frac{d}{c_1 \varphi \rho'(0+)}.$$

For such choices of $\lambda$ to exist, the minimum signal $d$ must satisfy

$$d \gg \varphi\left(\sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}}\right). \tag{3.11}$$

Recall that $\gamma \in [0, 1/2]$ has appeared in Condition 3.3. More insight can be gained by comparing the two parts on the right-hand side of (3.11): The first part will dominate if $n^\gamma \ll \sqrt{(\log p)/(\log s)}$, and in this case, Theorem 3.1 guarantees recovery of signals that satisfy $d \gg \varphi\sqrt{(\log p)/n}$, independent of $\gamma$; otherwise, the second part will dominate, and the weakest recoverable signal will depend on the correlation between the two set of variables as reflected by the value of $\gamma$. Of course, for the $L_1$ penalty, since the first part in Condition 3.3 always dominates the second, we can simply take $\gamma = 0$, and thus the value of $\gamma$ plays no role in determining the lower bound for $d$.

### 3.3.2 Oracle Property

In addition to model selection consistency, the oracle property requires the regularized estimator to be asymptotically as efficient as the oracle estimator with the active set known a priori. For this purpose, clearly, some extra eigenvalue conditions are needed. Define $\Lambda_1 = \Lambda_{\min}(\mathbf{D}_{AA})$, $\Lambda_2 = \Lambda_{\min}(\mathbf{\Sigma}_{AA})$, and $\Lambda_3 = \Lambda_{\min}(\mathbf{D}_{AA}^{-1}\mathbf{\Sigma}_{AA}\mathbf{D}_{AA}^{-1})$. The oracle property of the proposed estimators is stated in the following result.

**Theorem 3.2 (Oracle Property).** *Assume that all the conditions for Theorem 3.1 hold. In addition, assume that*

$$\frac{n\Lambda_1^2}{s^2(\log s)^{r_1}} \to \infty, \qquad \frac{n\Lambda_2^2}{s^2} \to \infty, \qquad \frac{n\Lambda_1^4\Lambda_3}{s^3} \to \infty, \tag{3.12}$$

*and*

$$\frac{ns\lambda^2\rho_\lambda'(d)^2}{\Lambda_1^2\Lambda_3} \to 0, \tag{3.13}$$

*where $r_1 = (r+4)/r$. Then, for some constants $D, K > 0$, with probability at least*

$$1 - D\exp\left\{-Kn^{1/r_1}\left(\frac{\Lambda_1^2}{s^2} \wedge 1\right)^{1/r_1}\right\} \to 1,$$

*there exists a regularized estimator $\widehat{\boldsymbol{\beta}}$ that satisfies the following properties:*

*(a) (Sparsity) $\widehat{\boldsymbol{\beta}}_{A^c} = \mathbf{0}$.*

*(b) (Asymptotic normality) For every $\mathbf{u} \in \mathbb{R}^s$ with $\|\mathbf{u}\|_2 = 1$,*

$$\sqrt{n}\mathbf{u}^T\mathbf{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{0A}) \to_d N(0,1).$$

*Remark* 3.7. The three conditions in (3.12) relate the sparsity dimension $s$ to eigenvalue bounds for the matrices $\mathbf{D}_{AA}$, $\mathbf{\Sigma}_{AA}$, and $\mathbf{D}_{AA}^{-1}\mathbf{\Sigma}_{AA}\mathbf{D}_{AA}^{-1}$. If we consider the special case where the eigenvalues of these matrices are all bounded away from zero, then these conditions are trivially satisfied for $s = o(n^{1/3})$. The form of (3.12), however, deals with much more general situations and is very meaningful in that the eigenvalue bounds are closely related to the difficulty of the estimation problem.

*Remark* 3.8. In the context of linear regression, it is well known that the $L_1$ penalty does not have the oracle property (Zou, 2006; Wainwright, 2009). It is clear from (3.13) that this is also the case for the problem considered here. To see this, consider the case where $\Lambda_1$ and $\Lambda_3$ are fixed, and note that $\rho'(d) \equiv 1$ for the $L_1$ penalty. Conditions in (3.8) imply that $n\lambda^2 \to \infty$, and hence (3.13) cannot hold. For the SCAD and MCP penalties, (3.13) is trivially satisfied for $d \gg \lambda$, since $\rho'_\lambda(d) = 0$ in that case. For the SICA penalty, we have $\rho'(d) = a(a+1)/(a+d)^2$; thus, in order to obtain an oracle property, we should take $a \to 0+$ at a rate such that $d \gg a$ and $\sqrt{ns}\lambda a/d^2 \to 0$. This is reasonable since the SICA penalty approaches the $L_0$ penalty as $a \to 0+$.

## 3.4 Implementation: Iterative Coordinate Descent

In this section we describe an efficient implementation of the proposed estimators by iterative coordinate descent. The idea of using coordinate optimization for solving regularized least-squares problems was proposed by Fu (1998) and Daubechies, Defrise and De Mol (2004), but its potential for producing surprisingly fast algorithms in large sparse problems was not widely noticed until convincingly demonstrated by Friedman et al. (2007) and Wu and Lange (2008), among others. Recently, Fan and Lv (2011) and Breheny and Huang (2011) generalized the idea to regularized problems with concave penalties and showed that it works equally well.

We propose an iterative coordinate descent algorithm for solving problem (3.3), which is a slight modification of the iterative coordinate ascent algorithm given in Fan and Lv (2011). The idea is to optimize over one coordinate at a time, with the other coordinates fixed at their current values. In order to speed up convergence, in stead of cycling through all the coordinates, the algorithm maintains an active index set and adds indices that violate the optimality conditions, as given in Lemma 3.2, to this set at the end of each iteration; the next iteration will then cycle through only the active set. To produce a solution path, the above procedure can be performed with a decreasing sequence of regularization parameters, and for each parameter value, the solution from the previous step can be used as a warm start. The algorithm is stated as follows.

**Algorithm 3.1 (Iterative Coordinate Descent).**

1. Set $\lambda_0 = \|\mathbf{b}\|_\infty / \rho'(0+)$ and $\widehat{\boldsymbol{\beta}}^{\lambda_0} = \mathbf{0}$. Sample a decreasing sequence $(\lambda_1, \dots, \lambda_K)$.
2. For each $k = 1, \dots, K$, set $\widehat{\boldsymbol{\beta}}^{\lambda_k} = \widehat{\boldsymbol{\beta}}^{\lambda_{k-1}}$ and $\widehat{A} = \{j : \widehat{\beta}_j^{\lambda_{k-1}} \neq 0\}$.
3. Successively for each $j \in \widehat{A}$, compute the minimizer $\widehat{\beta}_j^{\lambda_k}$ of the objective function in (3.3) with all $\widehat{\beta}_{j'}^{\lambda_k}$, $j' \neq j$, fixed at their current values.
4. Set $\widehat{A} = \{j : \widehat{\beta}_j^{\lambda_k} \neq 0\} \cup \{j : |U_j(\widehat{\boldsymbol{\beta}}^{\lambda_k})| > \lambda\rho'(0+)\}$. Repeat Steps 3 and 4 until convergence.
5. Continue Steps 2–4 with the next $k$ until all $\widehat{\boldsymbol{\beta}}^{\lambda_k}$ are obtained.

Several computational considerations are in order. The success of Algorithm 3.1 relies largely on efficiently solving the one-dimensional minimization problem in Step 3. Fortunately, since $L(\boldsymbol{\beta})$ is linear, closed-form solutions to this one-dimensional problem exist for many commonly used penalty functions, including all the examples considered in Section 3.2.2; we provide detailed formulas for these penalties in the Appendix.

The tuning parameter $\lambda$ can be chosen by, for example, $M$-fold cross-validation. The function $L(\boldsymbol{\beta})$ can be used to define a cross-validation score, since it has a prediction error interpretation. It should be noted that $L(\boldsymbol{\beta})$ is negative if the model fits the data better than the trivial model with no variables selected. An alternative loss function is the (integrated)

Brier score for survival data introduced by Graf et al. (1999), which can be viewed as the average quadratic loss of the predicted survival function. The Brier score is a more intuitive measure of prediction accuracy in that it ranges from 0 to 1, and the trivial prediction 0.5 at all time yields a Brier score of 0.25. Our experience, however, does not indicate much difference between using these two loss functions with cross-validation. We therefore use $L(\boldsymbol{\beta})$ in the cross-validation procedure because of its computational simplicity.

The advantages of coordinate descent algorithms are most obvious when most of the estimated coefficients are zero. Too small values of $\lambda$ may result in too dense models and hence incur a heavy computational burden. For computational efficiency, we recommend the following two-stage sampling strategy: First, start with the maximum value $\lambda_0$ and successively take bracketing triples $(\lambda_{k-1}, \lambda_k, \lambda_{k+1})$ with $\lambda_k = \lambda_0 \rho^k$ for some $\rho \in (0, 1)$, until an increase in the cross-validation curve is observed; then sample a grid of points in the last bracketing interval $(\lambda_{k-1}, \lambda_{k+1})$. A similar two-stage sampling strategy was also suggested by Wu and Lange (2008).

## 3.5 Simulation Studies

In this section we report on some simulation studies to evaluate the performance of the proposed estimators with the several penalties considered in Section 3.2.2. In the first simulation, we set $(n, p, s) = (200, 50, 6)$. The data were generated from model (3.1), where we took $\lambda_0(t) \equiv 1$ and the first eight components of $\boldsymbol{\beta}_0$ to be $(1, -1, 0, 1, -1, 0, 1, -1)^T$. The time-independent covariate vector $\mathbf{Z}$ was generated from $N(\mathbf{0}, (\rho^{|i-j|})_{i,j=1}^{p})$, where we took $\rho = 0.2$ or $\rho = 0.5$. Since the unbounded covariates may cause the conditional hazard function to be negative, we dropped such data and continued generating new ones until the desired sample size was reached. The censoring time was generated from Unif$(0, c_0)$ with constant $c_0$ chosen so that the censoring rate is about 25%. We implemented the proposed estimators with the lasso, SCAD, MCP, and SICA penalties using Algorithm 3.1, and chose the regularization parameter $\lambda$ by fivefold cross-validation. To avoid a two-dimensional cross-validation, we took $a = 3.7$ for both SCAD and MCP, which was suggested by Fan and Li (2001). For the SICA penalty, a two-stage approach was used, in which we first set $a = 1$ and then adjusted it to $a = 10^{-4}$. The rationale behind this approach is that by setting $a$ to a larger value, we allow the SICA method to approximate a solution more stably, and then by adjusting $a$ to a smaller value, force it to select a sparser model. We replicated the simulation 100 times for each setting.

The results for the first simulation are summarized in Table 3.1, where, for each method, medians and median absolute deviations of five measures are reported. The prediction error is defined as $L(\widehat{\boldsymbol{\beta}})$ calculated from an independent testing sample of size 10,000. The $L_2$ loss and $L_1$ loss refer to $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1$, respectively, which are performance measures on estimation accuracy. The last two measures, the number of variables in the selected model and the number of missed true variables, pertain to sparsity and model selection consistency. To better illustrate the distribution of these measures, boxplots for three of the five measures are also shown in Figure 3.1.

From Table 3.1 and Figure 3.1, we see that all the concave penalties greatly improve on the performance of the lasso, in all of the five measures. The SCAD and MCP have very comparable performance, as expected from their similarity. The SICA method achieves a surprisingly good performance in terms of sparsity; in most of the time, it identified exactly the six true variables. This substantial improvement over the other methods is primarily due to the advantages of the two-stage approach discussed above.

In the second simulation, we examined the performance of the proposed estimators in a high-dimensional setting, where $(n, p, s) = (500, 1000, 6)$. The other settings were unchanged, except that, to reduce computational burden, we generated a testing sample of size 500. The simulation results are summarized in Table 3.2 and boxplots are shown in

Table 3.1: Simulation results for different methods with $(n, p, s) = (200, 50, 6)$. Values shown are medians of each measure, with median absolute deviations in parentheses.

| | Measure | Lasso | SCAD | MCP | SICA | Oracle |
|---|---|---|---|---|---|---|
| $\rho = 0.2$ | PE | −0.215 | −0.283 | −0.279 | −0.294 | −0.299 |
| | | (0.030) | (0.020) | (0.019) | (0.017) | (0.013) |
| | $L_2$ loss | 1.129 | 0.496 | 0.520 | 0.449 | 0.408 |
| | | (0.201) | (0.149) | (0.142) | (0.137) | (0.112) |
| | $L_1$ loss | 3.469 | 1.215 | 1.293 | 0.908 | 0.834 |
| | | (0.594) | (0.378) | (0.395) | (0.310) | (0.276) |
| | #Selected | 19 | 11 | 10 | 6 | 6 |
| | | (3) | (2) | (2) | (0) | (0) |
| | #Missed | 0 | 0 | 0 | 0 | 0 |
| | | (0) | (0) | (0) | (0) | (0) |
| $\rho = 0.5$ | PE | −0.144 | −0.226 | −0.225 | −0.240 | −0.244 |
| | | (0.036) | (0.016) | (0.017) | (0.015) | (0.011) |
| | $L_2$ loss | 1.211 | 0.472 | 0.483 | 0.427 | 0.387 |
| | | (0.218) | (0.148) | (0.144) | (0.146) | (0.140) |
| | $L_1$ loss | 3.931 | 1.233 | 1.211 | 0.875 | 0.814 |
| | | (0.710) | (0.384) | (0.398) | (0.351) | (0.281) |
| | #Selected | 20 | 12 | 10 | 6 | 6 |
| | | (3) | (2) | (1) | (0) | (0) |
| | #Missed | 0 | 0 | 0 | 0 | 0 |
| | | (0) | (0) | (0) | (0) | (0) |

Note: PE, prediction error; #Selected, number of variables in the selected model; #Missed, number of missed true variables.

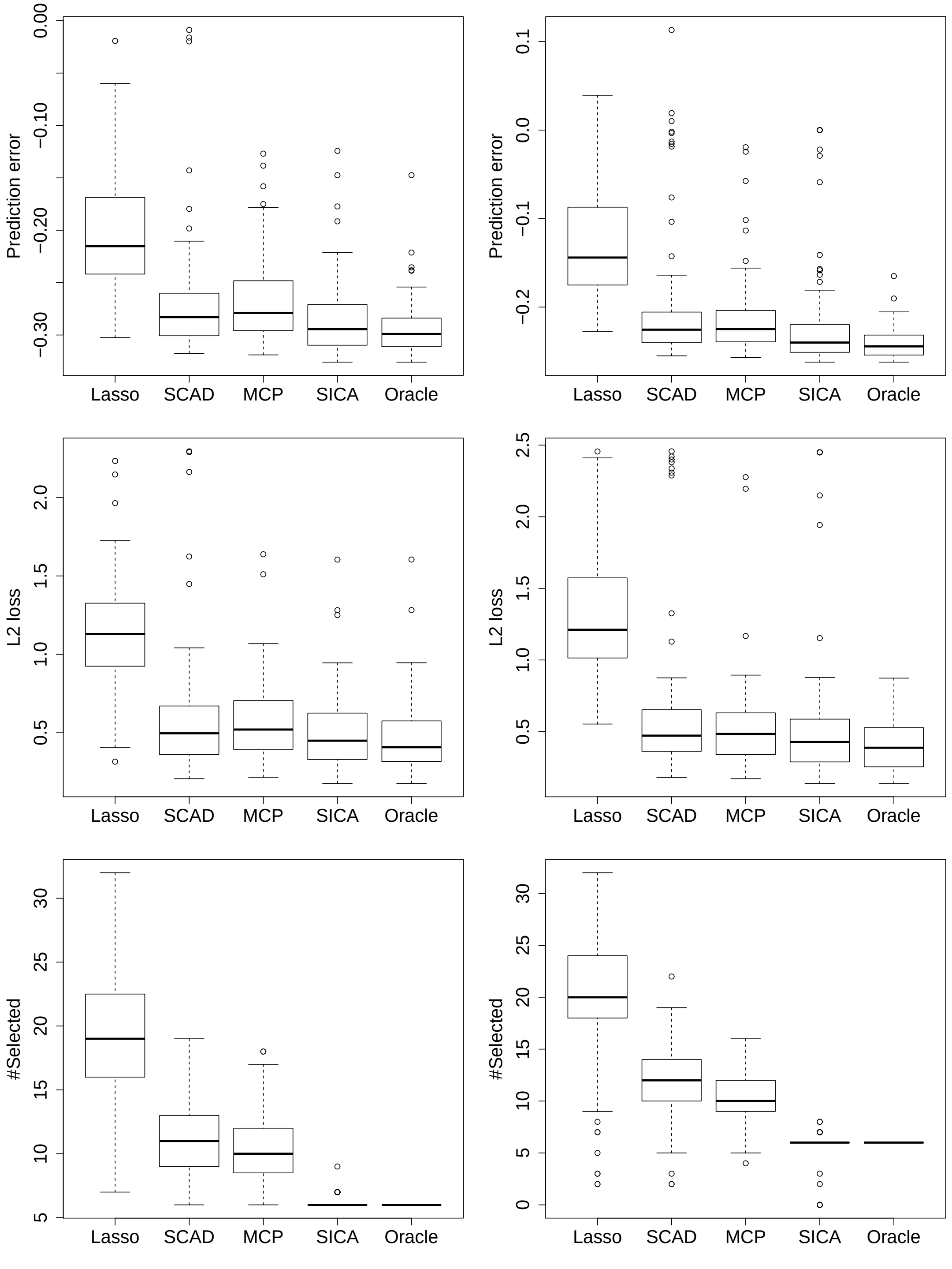

Figure 3.1: Boxplots of three measures for the simulation results with $(n, p, s) = (200, 50, 6)$. The left panel is for $\rho = 0.2$ and the right panel is for $\rho = 0.5$.

Table 3.2: Simulation results for different methods with $(n, p, s) = (500, 1000, 6)$. Values shown are medians of each measure, with median absolute deviations in parentheses.

| | Measure | Lasso | SCAD | MCP | SICA | Oracle |
|---|---|---|---|---|---|---|
| $\rho = 0.2$ | PE | −0.211 | −0.343 | −0.334 | −0.352 | −0.352 |
| | | (0.027) | (0.009) | (0.012) | (0.005) | (0.005) |
| | $L_2$ loss | 1.468 | 0.328 | 0.388 | 0.221 | 0.219 |
| | | (0.142) | (0.087) | (0.119) | (0.067) | (0.064) |
| | $L_1$ loss | 4.692 | 1.153 | 1.439 | 0.466 | 0.460 |
| | | (0.415) | (0.392) | (0.528) | (0.149) | (0.143) |
| | #Selected | 44 | 28 | 26 | 6 | 6 |
| | | (8.5) | (9) | (10) | (0) | (0) |
| | #Missed | 0 | 0 | 0 | 0 | 0 |
| | | (0) | (0) | (0) | (0) | (0) |
| $\rho = 0.5$ | PE | −0.095 | −0.273 | −0.274 | −0.286 | −0.286 |
| | | (0.040) | (0.010) | (0.011) | (0.004) | (0.004) |
| | $L_2$ loss | 1.856 | 0.325 | 0.309 | 0.215 | 0.215 |
| | | (0.236) | (0.088) | (0.076) | (0.071) | (0.071) |
| | $L_1$ loss | 5.883 | 1.318 | 1.090 | 0.453 | 0.453 |
| | | (0.297) | (0.407) | (0.363) | (0.178) | (0.177) |
| | #Selected | 51 | 35 | 23 | 6 | 6 |
| | | (13.5) | (8.5) | (6) | (0) | (0) |
| | #Missed | 0 | 0 | 0 | 0 | 0 |
| | | (0) | (0) | (0) | (0) | (0) |

Note: PE, prediction error; #Selected, number of variables in the selected model; #Missed, number of missed true variables.

Figure 3.2, from which we see similar trends as in the first simulation. In particular, we also observe a superior performance of the SICA method, which is quite close to that of the oracle estimator, whereas the other methods selected much larger models due to the high dimensionality.

## 3.6 Example: DLBCL Data

We now illustrate the proposed method by an application to the diffuse large-B-cell lymphoma (DLBCL) data reported by Rosenwald et al. (2002). The data set consists of gene

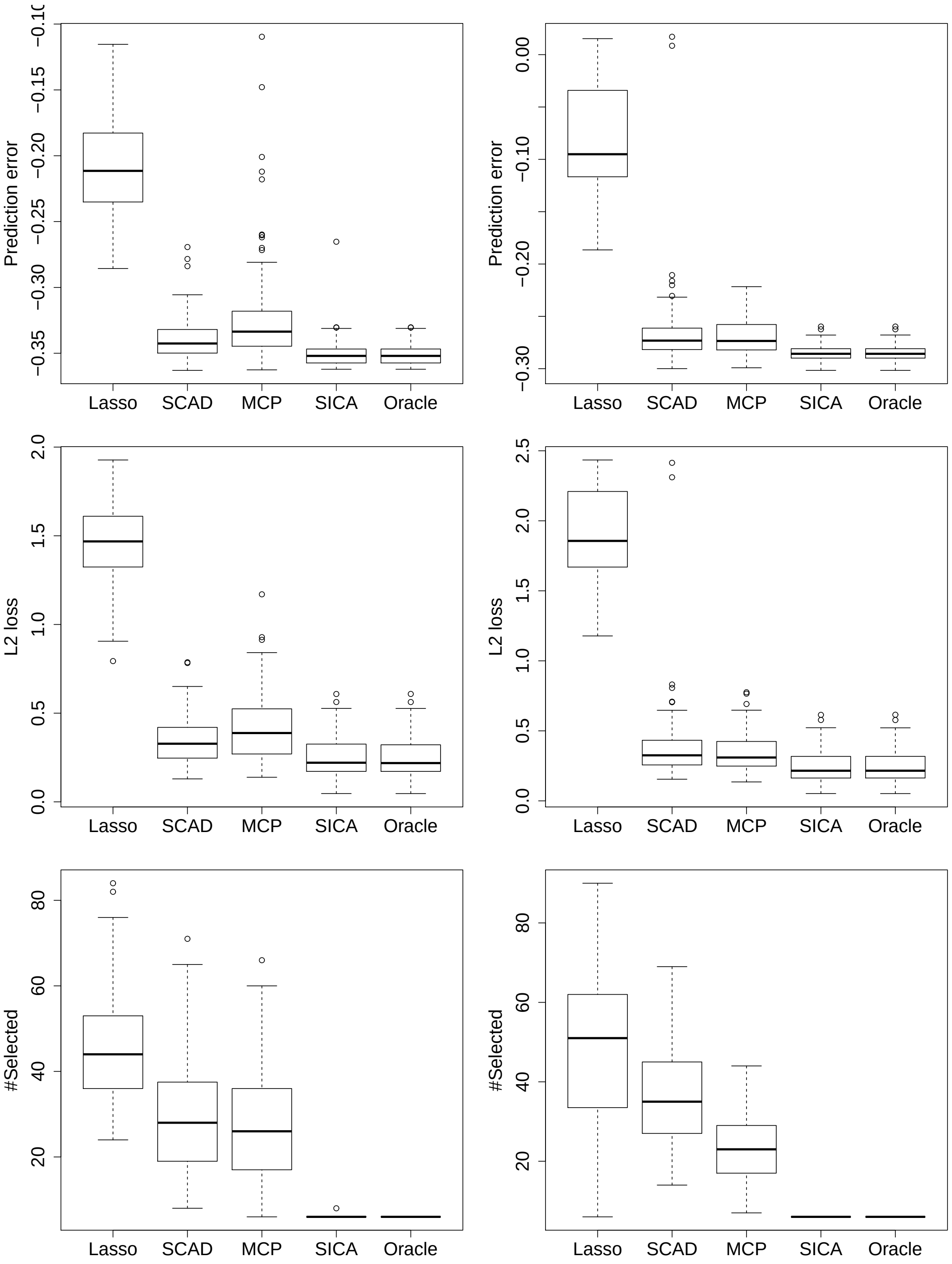

Figure 3.2: Boxplots of three measures for the simulation results with $(n, p, s) = (500, 1000, 6)$. The left panel is for $\rho = 0.2$ and the right panel is for $\rho = 0.5$.

Table 3.3: Results for different methods applied to the DLBCL data. Prediction errors are multiplied by 1000.

| | Lasso | SCAD | MCP | SICA |
|---|---|---|---|---|
| Selected model size | 31 | 12 | 3 | 3 |
| Prediction error | 4.88 | −4.11 | −2.95 | −3.00 |

expression measurements on 7,399 genes and survival outcomes after chemotherapy for 240 patients. The goal of the study is to formulate a molecular predictor of survival after chemotherapy for DLBCL. The whole data set was randomly divided into a training sample of size 160 and a testing sample of size 80. Our simulation studies suggest that the regularized estimators may not work well with such a high dimensionality. Therefore, to reduce the dimensionality to a reasonable scale, we first performed a univariate screening by picking out 160 genes with the largest regression coefficients in marginal regression. We then applied the proposed method with the lasso, SCAD, MCP, and SICA penalties to the data with the selected 160 genes.

For each method, the selected model size and the prediction error calculated from the testing sample are reported in Table 3.3. From these results, we see that all the concave penalties showed advantages in both sparsity and prediction accuracy over the lasso method. In this instance, SCAD performed best among the four methods, whereas MCP and SICA seem to have selected too sparse models. These observations, however, are not conclusive, since a more careful parameter tuning may be needed to improve the performance and enhance numerical stability of these methods.

## 3.7 Discussion

Although we have focused on the additive hazards model in this chapter, the techniques used in establishing our high-dimensional theory are not difficult to be generalized to other survival models. In particular, one could modify the empirical process arguments used here to develop a similar theory for the Cox model; a key step is to control the convergence rate of the empirical information matrix for the regression coefficients to its population counterpart. Since the partial likelihood score function is nonlinear, however, some extra effort in dealing with the technical details may be required.

Our theory indicates that the proposed method can potentially handle a very high dimensionality for survival data, even if the dimension of covariates grows nonpolynomially with the sample size. However, since we do not have accurate control of the constants that appear in our nonasymptotic results, as is common in the statistics literature, the curse of dimensionality might be more severe than the theory suggests. In particular, we note that survival data may require a relatively large sample size for the task of variable selection. In view of the fact that many clinical studies that involve high-throughput data have a limited sample size, it would be worthwhile to explore more effective strategies to combine the proposed method with other inference methods for high-dimensional survival data.

## 3.8 Proofs

In this section we provide the proofs of our theoretical results. The first lemma, presented in Section 3.8.1, gives conditions that characterize the solution to problem (3.3). In Section 3.8.2 we establish several concentration results that are essential to the main proofs. The proofs of Lemma 3.1 and Theorem 3.1 are given in Section 3.8.3, and the proof of Theorem 3.2 is given in Section 3.8.4.

### 3.8.1 Optimality Conditions

The following lemma provides sufficient optimality conditions that will be needed in the proof of Theorem 3.1.

**Lemma 3.2.** *Under Condition 3.1, $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is a strict local minimizer of problem* (3.3) *if the following conditions hold:*

$$\mathbf{U}_{\widehat{A}}(\widehat{\boldsymbol{\beta}}) - \lambda\rho'_\lambda(|\widehat{\boldsymbol{\beta}}_{\widehat{A}}|) \circ \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_{\widehat{A}}) = \mathbf{0}, \tag{3.14}$$

$$\|\mathbf{U}_{\widehat{A}^c}(\widehat{\boldsymbol{\beta}})\|_\infty < \lambda\rho'(0+), \tag{3.15}$$

*and*

$$\Lambda_{\min}(\mathbf{V}_{\widehat{A}\widehat{A}}) > \lambda\kappa(\rho_\lambda; \widehat{\boldsymbol{\beta}}_{\widehat{A}}), \tag{3.16}$$

*where $\circ$ is the Hadamard (entrywise) product.*

*Proof.* First, consider the $|\widehat{A}|$-dimensional subspace $\mathscr{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_{\widehat{A}^c} = \mathbf{0}\}$. Condition (3.16) implies that the objective function

$$Q(\boldsymbol{\beta}) \equiv L(\boldsymbol{\beta}) + \sum_{j=1}^{n} p_\lambda(|\beta_j|)$$

is strictly convex in a neighborhood of $\widehat{\boldsymbol{\beta}}$ in $\mathscr{B}$. Then (3.14) implies that $\widehat{\boldsymbol{\beta}}$ is a stationary point and hence a strict local minimizer of $Q(\boldsymbol{\beta})$ in the subspace $\mathscr{B}$.

It remains to show that, for any $\boldsymbol{\beta}_1 \in \mathbb{R}^p \setminus \mathscr{B}$ that lies in a small neighborhood of $\widehat{\boldsymbol{\beta}}$, we still have $Q(\boldsymbol{\beta}_1) > Q(\widehat{\boldsymbol{\beta}})$. To this end, let $\boldsymbol{\beta}_2$ be the projection of $\boldsymbol{\beta}_1$ onto the subspace $\mathscr{B}$. Since $Q(\boldsymbol{\beta}_2) \geq Q(\widehat{\boldsymbol{\beta}})$ from the preceding paragraph, it suffices to show that $Q(\boldsymbol{\beta}_1) > Q(\boldsymbol{\beta}_2)$. By the mean value theorem, we have

$$\begin{aligned} Q(\boldsymbol{\beta}_1) - Q(\boldsymbol{\beta}_2) &= \sum_{j \in \widehat{A}^c : \beta_{1j} \neq 0} \frac{\partial Q(\boldsymbol{\beta}^*)}{\partial \beta_j} \beta_{1j} \\ &= \sum_{j \in \widehat{A}^c : \beta_{1j} \neq 0} \{-U_j(\boldsymbol{\beta}^*) + \lambda\rho'_\lambda(|\beta^*_j|) \operatorname{sgn}(\beta^*_j)\}\beta_{1j}, \end{aligned} \tag{3.17}$$

where $\boldsymbol{\beta}^*$ is a point on the line segment between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. It follows from condition (3.15) and continuity that

$$|U_j(\boldsymbol{\beta}^*)| < \lambda \rho'_\lambda(|\beta^*_j|)\,\mathrm{sgn}(\beta^*_j), \qquad j \in \widehat{A}^c,$$

provided that $\boldsymbol{\beta}_1$, and hence $\boldsymbol{\beta}^*$, is sufficiently close to $\widehat{\boldsymbol{\beta}}$. Using this fact and that $\mathrm{sgn}(\beta^*_j) = \mathrm{sgn}(\beta_{1j})$, we see that each term in (3.17) is positive, and thus $Q(\boldsymbol{\beta}_1) > Q(\boldsymbol{\beta}_2)$. □

### 3.8.2 Concentration Results

We now establish a series of concentration results. Our primary mathematical tools are the maximal and concentration inequalities introduced in Section 1.3.3. We begin with the following lemma, on which the other results will be based.

**Lemma 3.3.** *Under Condition 3.2, there exist constants $C, K > 0$ such that, for every $x > 0$,*

$$\Pr\Big(\sup_{t\in[0,\tau]} |S^{(0)}(t) - s^{(0)}(t)| \geq Cn^{-1/2}(1+x)\Big) \leq \exp(-Kx^2), \qquad (3.18)$$

$$\Pr\Big(\sup_{t\in[0,\tau]} |S^{(1)}_j(t) - s^{(1)}_j(t)| \geq Cn^{-1/2}(1+x) \,|\, \Omega_L\Big) \leq \exp(-Kx^2/L^2), \qquad (3.19)$$

*and*

$$\Pr\Big(\sup_{t\in[0,\tau]} |S^{(2)}_{ij}(t) - s^{(2)}_{ij}(t)| \geq Cn^{-1/2}(1+x) \,|\, \Omega_L\Big) \leq \exp(-Kx^2/L^4), \qquad (3.20)$$

*for $i, j = 1, \ldots, p$.*

*Proof.* We only show (3.19), and the other two inequalities follow similarly. Denote $R_j = \sup_{t\in[0,\tau]} |S^{(1)}_j(t) - s^{(1)}_j(t)|$. The main idea is to apply Lemma 1.4; to this end, we need to control the term $ER_j$.

We first show that the class of functions $\{Y(t)Z_j(t)\colon t \in [0,\tau]\}$ has bounded uniform entropy integral (BUEI). Since a function of bounded variation can be expressed as the difference of two increasing functions, it follows from Condition 3.2(iv) and Lemma 9.10 of Kosorok (2008) that $\mathcal{Z}_j \equiv \{Z_j(t)\colon t \in [0,\tau]\}$ is a VC-hull class associated with a VC class of index 2. Then, by Corollary 2.6.12 of van der Vaart and Wellner (1996), for any probability measure $Q$,

$$\log N(\varepsilon \|F\|_{Q,2}, \mathcal{Z}_j, L_2(Q)) \le K\left(\frac{1}{\varepsilon}\right),$$

where $F$ is an envelope of the class $\mathcal{Z}_j$. Thus, the uniform entropy integral of $\mathcal{Z}_j$ is

$$\begin{aligned} J(1, \mathcal{Z}_j, L_2) &= \int_0^1 \sqrt{\log \sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{Z}_j, L_2(Q))}\, d\varepsilon \\ &\le \int_0^1 \sqrt{K(1/\varepsilon)}\, d\varepsilon < \infty. \end{aligned}$$

Also, from the definition of $Y(t)$ and Example 19.16 of van der Vaart (1998), we see that the class $\mathcal{Y} \equiv \{Y(t)\colon t \in [0,\tau]\}$ is VC and hence is BUEI. Thus, by the preservation results for BUEI classes (e.g., Theorem 9.15 of Kosorok, 2008), the class $\mathcal{Y}\mathcal{Z}_j$ is also BUEI.

Now an application of Lemma 1.3 gives

$$ER_j \lesssim n^{-1/2} J(1, \mathcal{Y}\mathcal{Z}_j, L_2) \|F\|_{P,2} \le Cn^{-1/2},$$

where the envelope $F$ is taken as $\sup_{t\in[0,\tau]} Y(t)|Z_j(t)|$. We apply Lemma 1.4 to conclude that

$$\begin{aligned} \Pr(R_j \ge Cn^{-1/2}(1+x) \mid \Omega_L) &\le \Pr(R_j \ge ER_j + Cn^{-1/2}x \mid \Omega_L) \\ &\le \exp(-Kx^2/L^2). \end{aligned}$$ □

The next lemma characterizes the tail behavior of the pseudoscore function at the true parameter value, and will be key to the proofs of Theorems 3.1 and 3.2.

**Lemma 3.4.** *Under Conditions 3.2, there exist constants* $C, D, K > 0$ *such that*

$$\Pr(\|\mathbf{U}(\boldsymbol{\beta}_0)\|_\infty \geq Cn^{-1/2}(1+x) \mid \Omega_L) \leq D\exp\left(-K\frac{x^2 \wedge n}{L^4}\right).$$

*Proof.* First, write

$$\begin{aligned} U_j(\boldsymbol{\beta}_0) &= \mathbb{P}_n \int_0^\tau \{Z_j(t) - \bar{Z}_j(t)\}\, dM(t) \\ &= \mathbb{P}_n \int_0^\tau Z_j(t)\, dM(t) - \mathbb{P}_n \int_0^\tau \bar{Z}_j(t)\, dM(t) \\ &\equiv T_1 - T_2. \end{aligned}$$

Since $T_1$ is an i.i.d. sum of mean-zero random variables, an application of Hoeffding's inequality yields that

$$\Pr(|T_1| \geq n^{-1/2}x \mid \Omega_L) \leq 2\exp(-Kx^2/L^4).$$

We will apply Lemma 1.4 to bound $T_2$. First, from (3.18) and (3.19) in Lemma 3.3, we have

$$\Pr\left(\sup_{t\in[0,\tau]} |S^{(0)}(t) - s^{(0)}(t)| \geq \delta\right) \leq \exp(-Kn)$$

and

$$\Pr\left(\sup_{t\in[0,\tau]} |S_j^{(1)}(t) - s_j^{(1)}(t)| \geq \delta \mid \Omega_L\right) \leq \exp(-Kn/L^2),$$

for a constant $\delta > 0$ and $j = 1, \dots, p$. Since these two tail probabilities are bounded by $\exp(-Kn/L^4)$, it suffices to consider the case where

$$\sup_{t\in[0,\tau]} |S^{(0)}(t) - s^{(0)}(t)| \le \delta \qquad \text{and} \qquad \sup_{t\in[0,\tau]} |S_j^{(1)}(t) - s_j^{(1)}(t)| \le \delta,$$

for a small $\delta > 0$ and $j = 1, \dots, p$. Write

$$\bar{Z}_j(t) - e_j(t) = \frac{1}{S^{(0)}(t)}\{S^{(1)}(t) - s^{(1)}(t)\} - \frac{s^{(1)}(t)}{S^{(0)}(t)s^{(0)}(t)}\{S^{(0)}(t) - s^{(0)}(t)\}. \quad (3.21)$$

Note that $s^{(0)}(\cdot)$ is bounded away from zero by Condition 3.2(ii). Then (3.21) implies that

$$\sup_{t\in[0,\tau]} |\bar{Z}_j(t) - e_j(t)| \le \delta'$$

for a constant $\delta' > 0$. Let $\mathscr{F}_j$ denote the class of functions $f\colon [0, \tau] \to \mathbb{R}$ that are of uniformly bounded variation and satisfy $\sup_{t\in[0,\tau]} |f(t) - e_j(t)| \le \delta'$. Define

$$\mathscr{G}_j = \left\{\int_0^\tau f(t)\, dM(t)\colon f \in \mathscr{F}_j\right\}$$

and $G_j = \|\mathbb{P}_n - P\|_{\mathscr{G}_j} = \|\mathbb{P}_n\|_{\mathscr{G}_j}$.

We need to control the term $EG_j$. By constructing $\|\cdot\|_\infty$-balls centered at piecewise constant functions on a regular grid, one can show that

$$N(\varepsilon, \mathscr{F}_j, \|\cdot\|_\infty) \le \left(\frac{K}{\varepsilon}\right)^{K'/\varepsilon}.$$

Also, note that, for any $f_1, f_2 \in \mathscr{F}_j$,

$$\left|\int_0^\tau f_1(t)\, dM(t) - \int_0^\tau f_2(t)\, dM(t)\right| \le \sup_{s\in[0,\tau]} |f_1(s) - f_2(s)| \int_0^\tau |dM(t)|.$$

From Theorem 2.7.11 of van der Vaart and Wellner (1996), it follows that

$$N_{[]}(2\varepsilon\|F\|_{P,2}, \mathcal{G}_j, L_2(P)) \le N(\varepsilon, \mathcal{F}_j, \|\cdot\|_\infty),$$

where $F = \int_0^\tau |dM(t)|$. Then the bracketing integral of $\mathcal{G}_j$ is

$$\begin{aligned} J_{[]}(1, \mathcal{G}_j, L_2(P)) &= \int_0^1 \sqrt{\log N_{[]}(\varepsilon, \mathcal{G}_j, L_2(P))}\, d\varepsilon \\ &\le \int_0^1 \sqrt{(K'/\varepsilon)\log(K/\varepsilon)}\, d\varepsilon < \infty. \end{aligned}$$

Thus, by Lemma 1.2, we have

$$EG_j \lesssim n^{-1/2} J_{[]}(\|G\|_{P,2}, \mathcal{G}_j, L_2(P)) \le Cn^{-1/2},$$

where $G$ is an envelope of $\mathcal{G}_j$. Now apply Lemma 1.4 to obtain

$$\Pr(|T_2| \ge Cn^{-1/2}(1+x) \mid \Omega_L) \le \exp(-Kx^2/L^4).$$

Putting the bounds for $T_1$ and $T_2$ together, the inequality follows. □

We now turn to concentration results regarding two important matrices. The first result pertains to entrywise concentration of the matrix $\mathbf{V}$ around its population counterpart $\mathbf{D}$, and will be useful in the proofs of Lemmas 3.1 and Theorem 3.2.

**Lemma 3.5.** *Under Condition 3.2, there exist constants $C, D, K > 0$ such that*

$$\Pr(|V_{ij} - D_{ij}| \ge Cn^{-1/2}(1+x) \mid \Omega_L) \le D \exp\left(-K\frac{x^2 \wedge n}{L^4}\right)$$

*for $i, j = 1, \dots, p$.*

*Proof.* First, write

$$\begin{aligned} V_{ij} - D_{ij} &= \int_0^\tau \left\{ S_{ij}^{(2)}(t) - \frac{S_i^{(1)}(t) S_j^{(1)}(t)}{S^{(0)}(t)} \right\} dt - \int_0^\tau \left\{ s_{ij}^{(2)}(t) - \frac{s_i^{(1)}(t) s_j^{(1)}(t)}{s^{(0)}(t)} \right\} dt \\ &= \int_0^\tau \{ S_{ij}^{(2)}(t) - s_{ij}^{(2)}(t) \}\, dt + \int_0^\tau \left\{ \frac{S_i^{(1)}(t) S_j^{(1)}(t)}{S^{(0)}(t)} - \frac{s_i^{(1)}(t) s_j^{(1)}(t)}{s^{(0)}(t)} \right\} dt \\ &\equiv T_1 + T_2. \end{aligned}$$

From (3.20) in Lemma 3.3, we have

$$\begin{aligned} &\Pr(|T_1| \geq C n^{-1/2}(1+x) \mid \Omega_L) \\ &\qquad \leq \Pr\left( \sup_{t \in [0,\tau]} |S_{ij}^{(2)}(t) - s_{ij}^{(2)}(t)| \geq C' n^{-1/2}(1+x) \mid \Omega_L \right) \leq \exp(-Kx^2/L^4). \end{aligned}$$

To bound $T_2$, write

$$\begin{aligned} &\frac{S_i^{(1)}(t) S_j^{(1)}(t)}{S^{(0)}(t)} - \frac{s_i^{(1)}(t) s_j^{(1)}(t)}{s^{(0)}(t)} \\ &\qquad = \frac{S_j^{(1)}(t)}{S^{(0)}(t)} \{ S_i^{(1)}(t) - s_i^{(1)}(t) \} + \frac{s_i^{(1)}(t)}{S^{(0)}(t)} \{ S_j^{(1)}(t) - s_j^{(1)}(t) \} \\ &\qquad\quad - \frac{s_i^{(1)}(t) s_j^{(1)}(t)}{S^{(0)}(t) s^{(0)}(t)} \{ S^{(0)}(t) - s^{(0)}(t) \}. \end{aligned}$$

By the same arguments as in the proof of Lemma 3.4, it suffices to consider the case where $S^{(0)}(\cdot)$ is bounded away from zero and $S_j^{(1)}(\cdot)$ are bounded. Then, from the preceding display and (3.18) and (3.19) in Lemma 3.3, it follows easily that

$$\Pr(|T_2| \geq C n^{-1/2}(1+x) \mid \Omega_L) \leq 3 \exp(-Kx^2/L^2).$$

Combining the bounds for $T_1$ and $T_2$ yields the desired inequality. □

The next result characterizes entrywise concentration of the matrix $\mathbf{W}$ around its population counterpart $\boldsymbol{\Sigma}$, and will be needed in the proof of Theorem 3.2.

**Lemma 3.6.** *Under Condition 3.2, there exist constants $C, D, K > 0$ such that*

$$\Pr(|W_{ij} - \Sigma_{ij}| \geq C n^{-1/2}(1 + x) \,|\, \Omega_L) \leq D \exp\biggl(-K \frac{x^2 \wedge n}{L^4}\biggr)$$

*for $i, j = 1, \ldots, p$.*

*Proof.* First, write

$$\begin{aligned}
W_{ij} - \Sigma_{ij} &= (\mathbb{P}_n - P) \int_0^\tau Z_i(t) Z_j(t) \, dN(t) \\
&\quad - \biggl\{\mathbb{P}_n \int_0^\tau \bar{Z}_i(t) Z_j(t) \, dN(t) - P \int_0^\tau e_i(t) Z_j(t) \, dN(t)\biggr\} \\
&\quad - \biggl\{\mathbb{P}_n \int_0^\tau Z_i(t) \bar{Z}_j(t) \, dN(t) - P \int_0^\tau Z_i(t) e_j(t) \, dN(t)\biggr\} \\
&\quad + \biggl\{\mathbb{P}_n \int_0^\tau \bar{Z}_i(t) \bar{Z}_j(t) \, dN(t) - P \int_0^\tau e_i(t) e_j(t) \, dN(t)\biggr\} \\
&\equiv T_1 - T_2 - T_3 + T_4.
\end{aligned}$$

Since $T_1$ is an i.i.d. sum, an application of Hoeffding's inequality gives

$$\Pr(|T_1| \geq n^{-1/2} x \,|\, \Omega_L) \leq 2 \exp(-K x^2 / L^4).$$

To bound $T_2$, write

$$\begin{aligned}
T_2 &= (\mathbb{P}_n - P) \int_0^\tau \bar{Z}_i(t) Z_j(t) \, dN(t) + P \int_0^\tau \{\bar{Z}_i(t) - e_i(t)\} Z_j(t) \, dN(t) \\
&\equiv T_{21} + T_{22}.
\end{aligned}$$

Using arguments similar to those for bounding $T_2$ in the proof of Lemma 3.4, we have

$$\Pr(|T_{21}| \geq Cn^{-1/2}(1+x) \,|\, \Omega_L) \leq D\exp\biggl(-K\frac{x^2 \wedge n}{L^4}\biggr).$$

Also, note that

$$|T_{22}| \leq \sup_{s\in[0,\tau]} |\bar{Z}_i(s) - e_i(s)| P \int_0^\tau |Z_j(t)|\, dN(t).$$

Then, by (3.21) and Lemma 3.3, one can easily show that

$$\Pr(|T_{22}| \geq Cn^{-1/2}(1+x) \,|\, \Omega_L) \leq D\exp\biggl(-K\frac{x^2 \wedge n}{L^4}\biggr).$$

Combining the bounds for $T_{21}$ and $T_{22}$ yields a similar bound for $T_2$. We can similarly bound $T_3$ and $T_4$, and arrive at the desired inequality. □

### 3.8.3 Proof of the Weak Oracle Property

*Proof of Lemma 3.1.* By the union bound and Lemma 3.5, we have

$$\begin{aligned}
&\Pr\biggl(\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty \geq \frac{1}{2\varphi} \,\bigg|\, \Omega_L\biggr) \\
&\quad = \Pr\biggl(\max_{i\in A}\sum_{j\in A}|V_{ij} - D_{ij}| \geq \frac{1}{2\varphi} \,\bigg|\, \Omega_L\biggr) \leq \sum_{i\in A}\Pr\biggl(\sum_{j\in A}|V_{ij} - D_{ij}| \geq \frac{1}{2\varphi} \,\bigg|\, \Omega_L\biggr) \\
&\quad \leq \sum_{i\in A}\sum_{j\in A}\Pr\biggl(|V_{ij} - D_{ij}| \geq \frac{1}{2\varphi s} \,\bigg|\, \Omega_L\biggr) \leq s^2 D\exp\biggl\{-K\frac{n}{L^4}\biggl(\frac{1}{\varphi^2 s^2} \wedge 1\biggr)\biggr\}.
\end{aligned}$$

By an error bound for matrix inversion (Horn and Johnson, 1985, p. 336), if $\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\| < 1/(2\varphi)$, then

$$\frac{\|\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}\|_\infty}{\varphi} \leq \frac{\varphi\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty}{1 - \varphi\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty} < 1,$$

which implies that

$$\|\mathbf{V}_{AA}^{-1}\|_\infty \le \varphi + \|\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}\|_\infty < 2\varphi.$$

Then inequality (3.4) follows.

To show (3.5), write

$$\begin{aligned}
&\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1} - \mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1} \\
&\qquad = (\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA})\mathbf{V}_{AA}^{-1} + \mathbf{D}_{A^cA}(\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}) \\
&\qquad = (\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA})\mathbf{V}_{AA}^{-1} - \mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1}(\mathbf{V}_{AA} - \mathbf{D}_{AA})\mathbf{V}_{AA}^{-1} \\
&\qquad \equiv T_1 - T_2.
\end{aligned}$$

Similarly as above, by the union bound and Lemma 3.5, we have

$$\begin{aligned}
&\Pr\left\{\|\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA}\|_\infty \ge \frac{1}{2\varphi}\left(\frac{\alpha}{4}\frac{\rho'(0+)}{\rho_\lambda'(d)}\right) \wedge \left(\frac{c}{2}n^\gamma\right) \,\Big|\, \Omega_L\right\} \\
&\qquad \le (p-s)sD\exp\left\{-K\frac{n}{L^4}\left(\frac{\left(\rho_\lambda'(d)^{-1} \wedge n^\gamma\right)^2}{\varphi^2 s^2} \wedge 1\right)\right\}.
\end{aligned}$$

This, along with (3.4), gives

$$\begin{aligned}
&\Pr\left\{\|T_1\|_\infty \ge \left(\frac{\alpha}{4}\frac{\rho'(0+)}{\rho_\lambda'(d)}\right) \wedge \left(\frac{c}{2}n^\gamma\right) \,\Big|\, \Omega_L\right\} \\
&\qquad \le \Pr\left\{\|\mathbf{V}_{A^cA} - \mathbf{D}_{A^cA}\|_\infty \ge \frac{1}{2\varphi}\left(\frac{\alpha}{4}\frac{\rho'(0+)}{\rho_\lambda'(d)}\right) \wedge \left(\frac{c}{2}n^\gamma\right) \,\Big|\, \Omega_L\right\} \\
&\qquad\qquad + \Pr\left(\|\mathbf{V}_{AA}^{-1}\|_\infty \ge 2\varphi \mid \Omega_L\right) \\
&\qquad \le (p-s)sD\exp\left\{-K\frac{n}{L^4}\left(\frac{\left(\rho_\lambda'(d)^{-1} \wedge n^\gamma\right)^2}{\varphi^2 s^2} \wedge 1\right)\right\} \\
&\qquad\qquad + s^2 D\exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2} \wedge 1\right)\right\}.
\end{aligned}$$

Also, by Condition 3.3 and (3.4), we have

$$\begin{aligned}
&\Pr\left\{\|T_2\|_\infty \geq \left(\frac{\alpha}{4}\frac{\rho'(0+)}{\rho'_\lambda(d)}\right) \wedge \left(\frac{c}{2}n^\gamma\right) \,\Big|\, \Omega_L\right\} \\
&\quad \leq \Pr\left\{\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_\infty \geq \frac{1}{2\varphi}\left(\frac{\alpha}{4(1-\alpha)} \wedge \frac{1}{2}\right) \,\Big|\, \Omega_L\right\} \\
&\qquad + \Pr\big(\|\mathbf{V}_{AA}^{-1}\|_\infty \geq 2\varphi \mid \Omega_L\big) \\
&\quad \leq s^2 D \exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2} \wedge 1\right)\right\}.
\end{aligned}$$

Putting the bounds for $T_1$ and $T_2$ together, we obtain

$$\begin{aligned}
&\Pr\left\{\|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1} - \mathbf{D}_{A^cA}\mathbf{D}_{AA}^{-1}\|_\infty \geq \left(\frac{\alpha}{2}\frac{\rho'(0+)}{\rho'_\lambda(d)}\right) \wedge (cn^\gamma) \mid \Omega_L\right\} \\
&\quad \leq (p-s)sD \exp\left\{-K\frac{n}{L^4}\left(\frac{\big(\rho'_\lambda(d)^{-1} \wedge n^\gamma\big)^2}{\varphi^2 s^2} \wedge 1\right)\right\} \\
&\qquad + s^2 D \exp\left\{-K\frac{n}{L^4}\left(\frac{1}{\varphi^2 s^2} \wedge 1\right)\right\}.
\end{aligned}$$

Then (3.5) follows from Condition 3.3 and the triangle inequality.

Finally, to show (3.6), by Corollary 6.3.8 of Horn and Johnson (1985), we have

$$\begin{aligned}
&|\Lambda_{\min}(\mathbf{V}_{AA}) - \Lambda_{\min}(\mathbf{D}_{AA})| \\
&\quad \leq \left\{\sum_{j=1}^{s} |\Lambda_{(j)}(\mathbf{V}_{AA}) - \Lambda_{(j)}(\mathbf{D}_{AA})|^2\right\}^{1/2} \leq \|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_2 \leq \|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_F,
\end{aligned}$$

where $\Lambda_{(j)}(\cdot)$ is the $j$th smallest eigenvalue and $\|\cdot\|_F$ is the Frobenius norm. Then, by the union bound and Lemma 3.5, we have

$$\begin{aligned}
&\Pr(|\Lambda_{\min}(\mathbf{V}_{AA}) - \Lambda_{\min}(\mathbf{D}_{AA})| \geq \mu \mid \Omega_L) \\
&\quad \leq \Pr(\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_F \geq \mu \mid \Omega_L) = \Pr\left(\sum_{i,j\in A} |V_{ij} - D_{ij}|^2 \geq \mu^2 \mid \Omega_L\right)
\end{aligned}$$

$$\le \sum_{i,j\in A} \Pr\left( |V_{jk} - D_{jk}| \ge \frac{\mu}{s} \,\middle|\, \Omega_L \right) \le s^2 D \exp\left\{ -K\frac{n}{L^4}\left( \frac{\mu^2}{s^2} \wedge 1 \right) \right\},$$

which, by the definition of $\mu$, implies (3.6). □

*Proof of Theorem 3.1.* By the union bound and Lemma 3.4, we have

$$\begin{aligned} &\Pr\left( \|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty \ge \frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+) \mid \Omega_L \right) \\ &\qquad \le \sum_{j\in A} \Pr\left( |U_j(\boldsymbol{\beta}_0)| \ge \frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+) \mid \Omega_L \right) \le sD \exp\left\{ -K\frac{n}{L^4}\left( \frac{\lambda^2}{n^{2\gamma}} \wedge 1 \right) \right\}. \end{aligned} \tag{3.22}$$

Similarly, we have

$$\Pr\left( \|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty \ge \frac{\alpha}{4}\lambda\rho'(0+) \mid \Omega_L \right) \le (p-s)D \exp\left\{ -K\frac{n}{L^4}(\lambda^2 \wedge 1) \right\}. \tag{3.23}$$

Also, Condition 3.2(iii) and the union bound imply that

$$\Pr(\Omega_L^c) \le \sum_{j=1}^{p} \Pr\left( \sup_{t\in[0,\tau]} |Z_j(t)| > L \right) \le pD\exp(-KL^r). \tag{3.24}$$

It follows from (3.22), (3.23), (3.24), and Lemma 3.1 that, with probability at least

$$\begin{aligned} 1 - (p-s)sD \exp&\left\{ -K\frac{n}{L^4}\left( \frac{\left(\rho'_\lambda(d)^{-1} \wedge n^\gamma\right)^2}{\varphi^2 s^2} \wedge 1 \right) \right\} \\ &- s^2 D \exp\left\{ -K\frac{n}{L^4}\left( \frac{(\varphi^{-1} \wedge \mu)^2}{s^2} \wedge 1 \right) \right\} - sD \exp\left\{ -K\frac{n}{L^4}\left( \frac{\lambda^2}{n^{2\gamma}} \wedge 1 \right) \right\} \\ &\qquad - (p-s)D \exp\left\{ -K\frac{n}{L^4}(\lambda^2 \wedge 1) \right\} - pD\exp(-KL^r), \end{aligned} \tag{3.25}$$

the following inequalities hold:

$$\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty < \frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(0+), \qquad \|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty < \frac{\alpha}{4}\lambda\rho'(0+),$$
$$\|\mathbf{V}_{AA}^{-1}\|_\infty < 2\varphi, \qquad \|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\|_\infty < \left(\left(1-\frac{\alpha}{2}\right)\frac{\rho'(0+)}{\rho_\lambda'(d)}\right) \wedge (2cn^\gamma),$$

and

$$\Lambda_{\min}(\mathbf{V}_{AA}) > \lambda\kappa_0. \tag{3.26}$$

Now assume that these inequalities hold. It suffices to find a $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ that satisfies all the conditions in Lemma 3.2 and the desired properties. To this end, take $\widehat{\boldsymbol{\beta}}_{A^c} = \mathbf{0}$, and we will determine $\widehat{\boldsymbol{\beta}}_A$ by using condition (3.14). Since $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{b} - \mathbf{V}\boldsymbol{\beta}$, we have

$$\begin{aligned}\mathbf{U}_A(\widehat{\boldsymbol{\beta}}) &= \mathbf{U}_A(\boldsymbol{\beta}_0) - \mathbf{V}_{AA}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) - \mathbf{V}_{AA^c}(\widehat{\boldsymbol{\beta}}_{A^c} - \boldsymbol{\beta}_{0A^c})\\ &= \mathbf{U}_A(\boldsymbol{\beta}_0) - \mathbf{V}_{AA}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}).\end{aligned}$$

Substituting this into the equation $\mathbf{U}_A(\widehat{\boldsymbol{\beta}}) - \lambda\rho_\lambda'(|\widehat{\boldsymbol{\beta}}_A|) \circ \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) = \mathbf{0}$ gives

$$\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A} = \mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) - \lambda\rho_\lambda'(|\widehat{\boldsymbol{\beta}}_A|) \circ \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)\}. \tag{3.27}$$

Define the function $f\colon \mathbb{R}^s \to \mathbb{R}^s$ by

$$f(\boldsymbol{\theta}) = \boldsymbol{\beta}_{0A} + \mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) - \lambda\rho_\lambda'(|\boldsymbol{\theta}|) \circ \operatorname{sgn}(\boldsymbol{\theta})\},$$

and let $\mathcal{K}$ denote the hypercube $\{\boldsymbol{\theta} \in \mathbb{R}^s\colon \|\boldsymbol{\theta} - \boldsymbol{\beta}_{0A}\|_\infty \le c_1\varphi\lambda\rho'(0+)\}$. By the inequalities we have assumed, for $\boldsymbol{\theta} \in \mathcal{K}$,

$$\|f(\boldsymbol{\theta}) - \boldsymbol{\beta}_{0A}\|_\infty \le \|\mathbf{V}_{AA}^{-1}\|_\infty\{\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty + \lambda\rho'(0+)\}$$

$$\leq 2\varphi\left(\frac{1}{2cn^{\gamma}}\frac{\alpha}{4}\lambda\rho'(0+) + \lambda\rho'(0+)\right) \leq c_1\varphi\lambda\rho'(0+),$$

i.e., $f(\mathcal{K}) \subset \mathcal{K}$. An application of Brouwer's fixed point theorem yields that equation (3.27) has a solution $\widehat{\boldsymbol{\beta}}_A$ in the hypercube $\mathcal{K}$. Then, by condition (3.9), we have $\|\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\| \leq d$, which implies that $\operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) = \operatorname{sgn}(\boldsymbol{\beta}_{0A})$ and hence $\widehat{A} = A$. Thus, we have found a $\widehat{\boldsymbol{\beta}}$ that satisfies condition (3.14) and the desired properties.

To verify that $\widehat{\boldsymbol{\beta}}$ satisfies condition (3.15), write

$$\begin{aligned}\mathbf{U}_{A^c}(\widehat{\boldsymbol{\beta}}) &= \mathbf{U}_{A^c}(\boldsymbol{\beta}_0) - \mathbf{V}_{A^cA}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A})\\ &= \mathbf{U}_{A^c}(\boldsymbol{\beta}_0) - \mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) - \lambda\rho_\lambda'(|\widehat{\boldsymbol{\beta}}_A|) \circ \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)\},\end{aligned}$$

where we have substituted (3.27). Since $\|\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\| \leq d$, by the definition of $d$, we have

$$\|\widehat{\boldsymbol{\beta}}_A\|_\infty = \|\boldsymbol{\beta}_{0A} + (\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A})\|_\infty \geq 2d - d = d.$$

The triangle inequality, concavity of $\rho_\lambda'(\cdot)$, and the inequalities we have assumed together imply that

$$\begin{aligned}\|\mathbf{U}_{A^c}(\widehat{\boldsymbol{\beta}})\|_\infty &\leq \|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty + \|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\|_\infty\{\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty + \lambda\rho_\lambda'(|\widehat{\boldsymbol{\beta}}_A|)\}\\ &\leq \|\mathbf{U}_{A^c}(\boldsymbol{\beta}_0)\|_\infty + \|\mathbf{V}_{A^cA}\mathbf{V}_{AA}^{-1}\|_\infty\{\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_\infty + \lambda\rho_\lambda'(d)\}\\ &< \frac{\alpha}{4}\lambda\rho'(0+) + 2cn^{\gamma}\frac{1}{2cn^{\gamma}}\frac{\alpha}{4}\lambda\rho'(0+) + \left(1 - \frac{\alpha}{2}\right)\frac{\rho'(0+)}{\rho_\lambda'(d)}\lambda\rho_\lambda'(d)\\ &= \frac{\alpha}{4}\lambda\rho'(0+) + \frac{\alpha}{4}\lambda\rho'(0+) + \left(1 - \frac{\alpha}{2}\right)\lambda\rho'(0+)\\ &= \lambda\rho'(0+).\end{aligned}$$

Finally, since $\|\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}\| \leq d$, it follows from (3.26) and the definition of $\kappa_0$ that

$$\Lambda_{\min}(\mathbf{V}_{AA}) > \lambda\kappa_0 \geq \lambda\kappa(\rho_\lambda; \widehat{\boldsymbol{\beta}}_A),$$

which verifies condition (3.16).

To complete the proof, choose $L$ by matching the exponential terms in (3.25), and note that the probability tends to 1 by (3.7) and (3.8). □

### 3.8.4 Proof of the Oracle Property

*Proof of Theorem 3.2.* First, by the same arguments as for (3.6) in the proof of Lemma 3.1, one can obtain

$$\begin{aligned}
&\Pr(\Lambda_{\min}(\mathbf{V}_{AA}) \leq \Lambda_1/2 \,|\, \Omega_L) \\
&\quad = \Pr(|\Lambda_{\min}(\mathbf{V}_{AA}) - \Lambda_1| \geq \Lambda_1/2 \,|\, \Omega_L) \leq s^2 D \exp\left\{-K\frac{n}{L^4}\left(\frac{\Lambda_1^2}{s^2} \wedge 1\right)\right\}.
\end{aligned}$$

Thus, with probability at least

$$1 - s^2 D \exp\left\{-K\frac{n}{L^4}\left(\frac{\Lambda_1^2}{s^2} \wedge 1\right)\right\} - pD\exp(-KL^r), \tag{3.28}$$

it holds that $\Lambda_{\min}(\mathbf{V}_{AA}) > \Lambda_1/2$, and hence

$$\|\mathbf{V}_{AA}^{-1}\|_2 = 1/\Lambda_{\min}(\mathbf{V}_{AA}) < 2/\Lambda_1. \tag{3.29}$$

Now assume that (3.29) holds. Since sparsity is implied by Theorem 3.1, we only need to show the asymptotic normality. By substituting (3.27) from the proof of Theorem 3.1, we can write

$$
\begin{aligned}
&\sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_{0A}) \\
&\quad = \sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\mathbf{V}_{AA}^{-1}\{\mathbf{U}_A(\boldsymbol{\beta}_0) - \lambda\rho_\lambda'(|\widehat{\boldsymbol{\beta}}_A|)\circ \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A)\} \\
&\quad = \sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{U}_A(\boldsymbol{\beta}_0) + \sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}(\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1})\mathbf{U}_A(\boldsymbol{\beta}_0) \\
&\qquad - \sqrt{n}\mathbf{u}^T\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\mathbf{V}_{AA}^{-1}\lambda\rho_\lambda'(|\widehat{\boldsymbol{\beta}}_A|)\circ \operatorname{sgn}(\widehat{\boldsymbol{\beta}}_A) \\
&\quad \equiv T_1 + T_2 - T_3.
\end{aligned}
$$

First consider $T_2$. Since $\|\mathbf{u}\|_2 = 1$, we have

$$
\begin{aligned}
|T_2| &\le \sqrt{n}\|\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2\|\mathbf{V}_{AA}^{-1} - \mathbf{D}_{AA}^{-1}\|_2\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_2 \\
&= \sqrt{n}\|\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2\|\mathbf{D}_{AA}^{-1}\|_2\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_2\|\mathbf{V}_{AA}^{-1}\|_2\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_2.
\end{aligned}
$$

It follows from Lemma 3.5 that

$$
\|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_2 \le \|\mathbf{V}_{AA} - \mathbf{D}_{AA}\|_F \le \left(s^2 \max_{i,j\in A}|V_{ij} - D_{ij}|\right)^{1/2} = sO_p(n^{-1/2}),
$$

and similarly, by Lemma 3.4, $\|\mathbf{U}_A(\boldsymbol{\beta}_0)\|_2 = \sqrt{s}O_p(n^{-1/2})$. Using also

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2 &= \sqrt{1/\Lambda_{\min}(\mathbf{D}_{AA}^{-1}\boldsymbol{\Sigma}_{AA}\mathbf{D}_{AA}^{-1})} = \Lambda_3^{-1/2}, \\
\|\mathbf{D}_{AA}^{-1}\|_2 &= 1/\Lambda_{\min}(\mathbf{D}_{AA}) = 1/\Lambda_1,
\end{aligned}
$$

and (3.29), we then obtain

$$
|T_2| \le \sqrt{n}\Lambda_3^{-1/2}\Lambda_1^{-1}sO_p(n^{-1/2})2\Lambda_1^{-1}\sqrt{s}O_p(n^{-1/2}) = \frac{2s^{3/2}}{\Lambda_1^2\Lambda_3^{1/2}}O_p(n^{-1/2}),
$$

which is $o_p(1)$ by the third condition in (3.12).

Then consider $T_3$, and the concavity of $\rho'_\lambda(\cdot)$ and condition (3.13) imply that

$$\begin{aligned}|T_3| &\le \sqrt{n}\|\mathbf{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2\|\mathbf{V}_{AA}^{-1}\|_2\lambda\|\rho'_\lambda(|\widehat{\boldsymbol{\beta}}_A|)\|_2\\ &\le \sqrt{n}\|\mathbf{\Sigma}_{AA}^{-1/2}\mathbf{D}_{AA}\|_2\|\mathbf{V}_{AA}^{-1}\|_2\lambda\sqrt{s}\rho'_\lambda(d)\\ &= \frac{2\sqrt{ns}\lambda\rho'_\lambda(d)}{\Lambda_1\Lambda_3^{1/2}} \to 0.\end{aligned}$$

Now it remains to show that $T_1$ is asymptotically normal. To this end, note that

$$\mathbf{u}^T\mathbf{\Sigma}_{AA}^{-1/2}\mathbf{W}_{AA}\mathbf{\Sigma}_{AA}^{-1/2}\mathbf{u} = 1 + \mathbf{u}^T\mathbf{\Sigma}_{AA}^{-1/2}(\mathbf{W}_{AA} - \mathbf{\Sigma}_{AA})\mathbf{\Sigma}_{AA}^{-1/2}\mathbf{u}.$$

By Lemma 3.6, we have $\|\mathbf{W}_{AA} - \mathbf{\Sigma}_{AA}\|_2 = sO_p(n^{-1/2})$. Then the second term in the preceding display is bounded by

$$\|\mathbf{\Sigma}_{AA}^{-1/2}\|_2\|\mathbf{W}_{AA} - \mathbf{\Sigma}_{AA}\|_2\|\mathbf{\Sigma}_{AA}^{-1/2}\|_2 = \Lambda_2^{-1/2}sO_p(n^{-1/2})\Lambda_2^{-1/2} = \frac{s}{\Lambda_2}O_p(n^{-1/2}),$$

which is $o_p(1)$ by the second condition in (3.12). An application of the martingale central limit theorem yields that $T_1$ is asymptotically standard normal.

Finally, we match the exponential terms in (3.28) and choose the optimal $L$. The probability tends to 1 by the first condition in (3.12). □

# CHAPTER 4
# Future Work

In this dissertation we have contented ourselves with inference in the additive hazards model. The two problems we have considered, however, are relevant in much more general contexts. We now list a number of directions for future work.

On missing covariates:

- *Maximum likelihood approach.* The maximum likelihood approach has far-reaching impact in survival analysis, and the resulting estimators typically enjoy some form of optimality. It would be of interest to explore its use for handling missing data in a large class of of survival models to which the approach is applicable.

- *Semiparametric and nonparametric covariate effects.* There has been much effort to consider more flexible, for example, time-varying, partially linear, nonparametric additive, and fully nonparametric, covariate effects. Missing data problems in these more general models are still widely open.

- *Model checking.* Model checking techniques are valuable in practice, especially when conclusions drawn under different model assumptions disagree. The presence of missing data can make the development of such techniques more difficult, which has not yet been addressed in the literature.

- *Nonignorable missingness.* Most of the previous work has made the missing at random assumption, which may be violated in many practical situations. Dealing with these complex cases may require new methodological development.

On high-dimensional inference:

- *Methods for very high dimensionality.* Due to the intrinsic difficulty underlying high-dimensional problems, the existing methods are far from satisfactory, and as noted in Chapter 3, survival data add more intricacy to these problems.
- *Incorporating structural information.* Various types of structural information on the covariates may be supplied in advance. The role of such structural information in high-dimensional inference has not yet been clarified.
- *Semiparametric and nonparametric covariate effects.* Several smoothing techniques could be combined with the existing methodology to relax parametric assumptions on the covariate effects, and new theory and methods may be required.
- *Misspecified models and robust inference.* Model misspecification is often a serious concern in practice. Robust methods that are suitable for high-dimensional inference need to be invented.

# References

Aalen, O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical Statistics and Probability Theory: Proceedings of the Sixth International Conference* (W. Klonecki, A. Kozek and J. Rosiński, eds.). Springer, New York, 1–25.

Aalen, O. O., Borgan, O. and Gjessing, H. K. (2008). *Survival and Event History Analysis*. Springer, New York.

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

Antoniadis, A., Fryzlewicz, P. and Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics* **37**, 531–552.

Begun, J. M., Hall, W. J., Huang, W.-M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics* **11**, 432–452.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis* **6**, 39–58.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, to appear.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5**, 232–253.

Breslow, N. E. and Day, N. E. (1987). *Statistical Models in Cancer Research, 2: The Design and Analysis of Cohort Studies*. IARC, Lyon.

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303–316.

Cai, T., Huang, J. and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65**, 394–404.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when $p$ is much larger than $n$ (with discussion). *The Annals of Statistics* **35**, 2313–2404.

Chen, H. Y. (2002). Double-semiparametric method for missing covariates in Cox regression models. *Journal of the American Statistical Association* **97**, 565–576.

Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* **94**, 896–908.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Ser. B* **34**, 187–220.

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

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Daubechies, I., Defrise, M. and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* **57**, 1413–1457.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Ser. B* **39**, 1–37.

Fan, J. (1997). Comments on "Wavelets in statistics: A review," by A. Antoniadis. *Journal of the Italian Statistical Society* **6**, 131–138.

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

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 74–99.

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

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, to appear.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302–332.

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

Gijbels, I., Lin, D. and Ying, Z. (2007). Non- and semi-parametric analysis of failure time data with missing failure indicators. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond* (R. Liu, W. Strawderman and C.-H. Zhang, eds.). Institute of Mathematical Statistics, 203–223.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *The Annals of Statistics* **20**, 1903–1928.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika* **91**, 929–941.

Herring, A. H. and Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association* **96**, 292–302.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, New York.

Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis* **16**, 176–195.

Huang, J., Ma, S. and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. and Lauer, M. S. (2010). High-dimensional variable selection for survival data. *Journal of the American Statistical Association* **105**, 205–217.

Jiang, J. and Zhou, H. (2007). Additive hazard regression with auxiliary covariates. *Biometrika* **94**, 359–369.

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society, Ser. B* **70**, 351–370.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. Wiley, Hoboken, NJ.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

Kulich, M. and Lin, D. Y. (2000a). Additive hazards regression for case-cohort studies. *Biometrika* **87**, 73–87.

Kulich, M. and Lin, D. Y. (2000b). Additive hazards regression with covariate measurement error. *Journal of the American Statistical Association* **95**, 238–248.

Kulich, M. and Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* **99**, 832–844.

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

Ledoux, M. (1995). On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics* **1**, 63–87.

Leng, C. and Ma, S. (2007). Path consistent model selection in additive risk model via Lasso. *Statistics in Medicine* **26**, 3753–3770.

Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* **88**, 1341–1349.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

Lin, D. Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *The Annals of Statistics* **23**, 1712–1734.

Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* **54**, 1002–1013.

Luo, X., Tsai, W. Y. and Xu, Q. (2009). Pseudo-partial likelihood estimators for the Cox regression model with missing covariates. *Biometrika* **96**, 617–633.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37**, 3498–3528.

Martinussen, T. and Scheike, T. H. (2009). Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics* **36**, 602–619.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability* **28**, 863–884.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika* **81**, 501–514.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association* **95**, 449–485.

Nan, B. (2004). Efficient estimation for case-cohort studies. *Canadian Journal of Statistics* **32**, 403–419.

Nan, B., Kalbfleisch, J. D. and Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *The Annals of Statistics* **37**, 2351–2376.

Paik, M. C. (1997). Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis* **3**, 289–298.

Paik, M. C. and Tsai, W.-Y. (1997). On using the Cox proportional hazards model with missing covariates. *Biometrika* **84**, 579–593.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *The Annals of Statistics* **10**, 475–478.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Pugh, M., Robins, J., Lipsitz, S. and Harrington, D. (1993). Inference in the Cox proportional hazards model with missing covariate data. Techical Report 758Z, Dept. of Biostatistics, Harvard School of Public Health.

Qi, L., Wang, C. Y. and Prentice, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association* **100**, 1250–1263.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics* **11**, 453–466.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *The Annals of Statistics* **38**, 1287–1319.

Robins, J. M. and Rotnitzky, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer" by P. J. Bickel and J. Kwon **11**, 920–936.

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

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. and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine* **346**, 1937–1947.

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

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics* **16**, 64–81.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

Thomas, D. C. (1977). Addendum to "Methods of cohort analysis: Appraisal by application to asbestos mining," by F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *Journal of the Royal Statistical Society, Ser. A* **140**, 483–485.

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

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Tsai, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96**, 601–15.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, New York.

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

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* **55**, 2183–2202.

Wang, C. Y. and Chen, H. Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics* **57**, 414–419.

Wang, S., Nan, B., Zhu, J. and Beer, D. G. (2008). Doubly penalized Buckley–James method for survival data with high-dimensional covariates. *Biometrics* **64**, 132–140.

Witten, D. M. and Tibshirani, R. (2010). Survival analysis with high-dimensional covariates. *Statistical Methods in Medical Research* **19**, 29–51.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* **2**, 224–244.

Xu, Q., Paik, M. C., Luo, X. and Tsai, W.-Y. (2009). Reweighting estimators for Cox regression with missing covariates. *Journal of the American Statistical Association* **104**, 1155–1167.

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

Zeng, D., Yin, G. and Ibrahim, J. G. (2005). Inference for a class of transformed hazards models. *Journal of the American Statistical Association* **100**, 1000–1008.

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

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

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

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

# Appendix: Regularized Least Squares in One Dimension

In this appendix we provide solution formulas for regularized least squares in one dimension with the lasso, SCAD, MCP, and SICA penalties. The purpose is to supply implementation details for Algorithm 3.1, as indicated in Section 3.4.

Consider the one-dimensional regularized least squares problem

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}} \left\{ \frac{1}{2}(\theta - \theta_0)^2 + p_\lambda(|\theta|) \right\},$$

where the form of $p_\lambda(\cdot)$ for these penalties can be found in Section 3.2.2. The formulas for the lasso, SCAD, and MCP penalties are well known or easily derived, and so they are given below directly.

- Lasso:

$$\widehat{\theta} = \operatorname{sgn}(\theta_0)(|\theta_0| - \lambda)_+.$$

- SCAD:

$$\widehat{\theta} = \begin{cases} \operatorname{sgn}(\theta_0)(|\theta_0| - \lambda)_+, & |\theta_0| \le 2\lambda, \\ \operatorname{sgn}(\theta_0)\dfrac{|\theta_0| - a\lambda/(a-1)}{1 - 1/(a-1)}, & 2\lambda < |\theta_0| \le a\lambda, \\ \theta_0, & |\theta_0| > a\lambda. \end{cases}$$

- MCP:

$$\widehat{\theta} = \begin{cases} \operatorname{sgn}(\theta_0)\dfrac{(|\theta_0| - \lambda)_+}{1 - 1/a}, & |\theta| \le a\lambda, \\ \theta_0, & |\theta| > a\lambda. \end{cases}$$

For the SICA penalty, we have

$$p'(\theta) = \lambda \frac{a(a+1)}{(a+\theta)^2}.$$

To find the nonzero critical points, we need to solve the equation

$$\theta - \theta_0 + \lambda\, \operatorname{sgn}(\theta) \frac{a(a+1)}{(a+|\theta|)^2} = 0.$$

Noting that $\operatorname{sgn}(\theta) = \operatorname{sgn}(\theta_0)$ and canceling it from the above equation give

$$|\theta| - |\theta_0| + \lambda \frac{a(a+1)}{(a+|\theta|)^2} = 0,$$

or

$$|\theta|^3 + (2a - |\theta_0|)\theta^2 + (a^2 - 2a|\theta_0|)|\theta| + \lambda a(a+1) - a^2|\theta_0| = 0.$$

Consider the positive roots of the cubic equation

$$t^3 + c_2 t^2 + c_1 t + c_0 = 0, \tag{A.1}$$

where $c_2 = 2a - |\theta_0|$, $c_1 = a^2 - 2a|\theta_0|$, and $c_0 = \lambda a(a+1) - a^2|\theta_0|$. Let

$$q = \frac{c_2^2 - 3c_1}{9} \qquad \text{and} \qquad r = \frac{2c_2^3 - 9c_1c_2 + 27c_0}{54}.$$

If $r^2 \geq q^3$, then $\widehat{\theta} = 0$. Otherwise, consider the two largest roots of equation (A.1),

$$t_1 = -2\sqrt{q}\cos\left(\frac{\alpha - 2\pi}{3}\right) - \frac{c_2}{3} \qquad \text{and} \qquad t_2 = -2\sqrt{q}\cos\left(\frac{\alpha + 2\pi}{3}\right) - \frac{c_2}{3},$$

where $\alpha = \arccos(r/\sqrt{q^3})$ and $t_1 < t_2$. We discuss the following cases:

(a) If $t_1 > 0$, then $t_1$ is a local maximum and $t_2$ is a local minimum, and we need to compare the values of the objective function at $t_2$ and 0:

$$\begin{aligned} J(t_2) - J(0) &= \frac{1}{2}(t_2 - \theta_0)^2 + \lambda\frac{(a+1)t_2}{a + t_2} - \frac{1}{2}\theta_0^2 \\ &= \frac{1}{2}t_2^2 - \theta_0 t_2 + \lambda\frac{(a+1)t_2}{a + t_2}. \end{aligned}$$

Thus, if

$$\frac{1}{2}t_2 + \lambda\frac{a+1}{a+t_2} < \theta_0,$$

then $J(t_2) < J(0)$, and $\widehat{\theta} = \text{sgn}(\theta_0)t_2$; otherwise, $\widehat{\theta} = 0$.

(b) If $t_1 < 0$ and $t_2 > 0$, then $t_2$ is a local minimum, and $\widehat{\theta} = \text{sgn}(\theta_0)t_2$.

(c) If $t_2 < 0$, then $\widehat{\theta} = 0$.