Abstract

LU, KAIFENG. Estimation of Regression Coefficients in the Competing Risks Model with Missing Cause of Failure. (Under the direction of Dr. Anastasios A. Tsiatis.)

In many clinical studies, researchers are interested in the effects of a set of prognostic factors on the hazard of death from a specific disease even though patients may die from other competing causes. Often the time to relapse is right-censored for some individuals due to incomplete follow-up. In some circumstances, it may also be the case that patients are known to die but the cause of death is unavailable. When cause of failure is missing, excluding the missing observations from the analysis or treating them as censored may yield biased estimates and erroneous inferences. Under the assumption that cause of failure is missing at random, we propose three approaches to estimate the regression coefficients. The imputation approach is straightforward to implement and allows for the inclusion of auxiliary covariates, which are not of inherent interest for modeling the cause-specific hazard of interest but may be related to the missing data mechanism. The partial likelihood approach we propose is semiparametric efficient and allows for more general relationships between the two cause-specific hazards and more general missingness mechanism than the partial likelihood approach used by others. The inverse probability weighting approach is doubly robust and highly efficient and also allows for the incorporation of auxiliary covariates. Using martingale theory and semiparametric theory for missing data problems, the asymptotic properties of these estimators are developed and the semiparametric efficiency of relevant estimators is proved. Simulation studies are carried out to assess the performance of these estimators in finite samples. The approaches are also illustrated using the data from a clinical trial in elderly women with stage II breast cancer. The inverse probability weighted doubly robust semiparametric estimator is recommended for its simplicity, flexibility, robustness and high efficiency.

KEY WORDS: Cause-specific hazard; Doubly robust; Imputation; Influence function; Inverse probability weighting; Locally efficient; Missing at random; Partial likelihood; Proportional hazards model; Semiparametric model.

ESTIMATION OF REGRESSION COEFFICIENTS IN THE COMPETING RISKS MODEL WITH MISSING CAUSE OF FAILURE

by

KAIFENG LU

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh
March, 2002

APPROVED BY:

Anastasios A. Tsiatis Chair of Advisory Committee	Marie Davidian
Sujit Ghosh	John F. Monahan

To my parents

Biography

Kaifeng Lu was born in Huaian, China to parents Wenchao Lu and Laiying Zhang on July 1, 1972. In recognition of his excellent academic performance in high school, he was exempt from the college entrance examination and entered Southeast University in August, 1991. He received a B.S. in Mathematics with a minor in Computer Engineering one year ahead of time in May, 1994. He continued to study at Southeast University for another two and a half years and received a M.S. in Statistics. In August, 1997, he entered the Ph.D. Program in Electrical Engineering at Nanjing University of Aeronautics and Astronautics and left one year later to join the Ph.D. program in Statistics at North Carolina State University. Upon completion of his doctorial degree, he will start to work at Eli Lilly as a senior statistician and begin his career in the pharmaceutical industry.

Acknowledgements

I would like to express my deepest gratitude and appreciation to my academic advisor, Dr. Anastasios A. Tsiatis, for his ceaseless inspiration and encouragement. Besides being one of leading researchers in the field of biostatistics, he is also one of the most respectable people one could ever associate with. Without his great guidance and patience, it would have been much more difficult for me to complete the doctorate degree.

I would also like to thank Dr. Sastry G. Pantula for his continuous support of my graduate study, Dr. David A. Dickey for his constant help on SAS, Dr. Marie Davidian for her useful tips on LaTeX, Dr. John F. Monahan for his insightful explanation on simulation, Dr. Sujit Ghosh for his valuable comments on the draft. I am also grateful to Terry Byron for his aid on computing, Janice Gaddy for her assistance toward graduation. I would also like to especially thank my best friend, Pei-Yun Chen, for her relentless help in numerous aspects of my life. My thanks also go to other faculty, staff, and fellow students in our department. The congenial and conducive learning environment they provide has made my study such a pleasant and fruitful experience.

Finally and foremost, I would like to thank my parents, my three sisters, and my brother for their love and support all the way in my life.

Contents

List of Tables

List of Abbreviations

CC	complete case
CLT	central limit theorem
CP	coverage probability
CPBT	coverage probability via bootstrap
DR	doubly robust
EPL	efficient partial likelihood
ER	estrogen receptor
GR	Goetghebeur and Ryan
IPW	inverse probability weighted
IPWCC	inverse probability weighted complete-case
IPWDR	inverse probability weighted doubly robust
IPWLE	inverse probability weighted locally efficient
LHS	left hand side
LIE	law of iterated expectations
MAR	missing at random
MI	multiple imputation
MLE	maximum likelihood estimator
MPLE	maximum partial likelihood estimator
RHS	right hand side
SE	standard error
SEE	standard error estimate
SEEBT	standard error estimate via bootstrap
SI	single imputation
SSE	sampling standard error
WLLN	weak law of large numbers

Chapter 1

Multiple Imputation Approach

1.1 Introduction

In many clinical studies where time to failure is of primary interest, patients may fail or die from one of many causes. For example, in a clinical trial that compares different therapies for breast cancer, interest may focus on death from breast cancer even though patients may die from other causes. A routine objective is to assess the effects of a set of prognostic covariates on the hazard rate of time to failure due to the cause of interest. In many studies, the cause of death information may be censored due to incomplete follow-up. In some circumstances, it may also be the case that patients are known to die but the cause of death is unavailable, e.g., whether death is attributable to the cause of interest or other causes may require documentation with information that is not collected or lost or cause may be difficult for investigators to determine for some patients (Andersen, Goetghebeur, and Ryan, 1996). If there were no missing cause of failure, the standard proportional hazards model can be used to model the cause-specific hazard of interest and the regression coefficients can be estimated using maximum partial likelihood estimators (Cox, 1972, 1975). However, when cause of failure is missing, excluding the missing observations from the analysis or treating them as censored may yield biased estimates and erroneous inferences. With missing cause of failure, Goetghebeur and Ryan (1995) proposed an approach by making assumptions directly on the relationship between the cause-specific hazard of interest and that of competing causes; these are assumed proportional, although this may be relaxed.

In this article, we use parametric models to model the probability that a missing cause is that of interest while allowing the inclusion of additional auxiliary covariates and we use multiple imputation procedures (e.g., Rubin, 1987, 1996; Wang and Robins, 1998) to impute missing cause of failure. On the basis of each of several imputed data sets, maximum partial likelihood estimators are computed and combined. In Section 1.2, we describe notation and the assumption of missing at random. In Section 1.3, we outline the imputation procedure.

In Section 1.4, we state asymptotic properties of imputation estimators with proofs sketched out in the Appendix. In Section 1.5, we provide simulation results to show the relevance of the theory in finite samples. In Section 1.6, we illustrate the results using data from a clinical trial in stage II breast cancer. In Section 1.7, we give a brief discussion.

1.2 Notation and Assumptions

In this article, we will consider the situation where individuals may fail or die from one of two specific causes, one of which is of interest. The cause of interest will be referred to as cause 2 and all other causes of death will be combined and referred to as cause 1. If there was no censoring, the cause of death data could be summarized as (T^*, Δ^*), where T^* denotes the time to death and Δ^* denotes the cause of death, taking on values one or two. A set of covariates X is also defined with the primary goal of modeling the cause-specific hazard for the cause of interest to these covariates, namely

$$\lambda^*(t|x) = \lim_{h \to 0} h^{-1} P(t \le T^* < t+h, \Delta^* = 2 | T^* \ge t, X = x).$$

A popular model for this relationship is the proportional hazards model, which assumes that

$$\lambda^*(t|x) = \lambda(t) e^{\beta^T x}, \tag{1.1}$$

where β is the q-dimensional vector of regression coefficients and $\lambda(t)$ is the unspecified baseline hazard for the cause of interest. For example, X may represent the indicator variable for treatment assignment and other baseline characteristics.

Because of incomplete follow-up, cause of death data are often censored by a variable C, in which case the data we observe can be summarized by the variables $T = \min(T^*, C)$ and Δ, which equals Δ^* if $T^* \le C$ and equals zero if $T^* > C$, i.e., T is the time to failure or censoring and Δ is the failure-censoring indicator taking on values zero, one, or two. To avoid nonidentifiability problems, we assume that C is conditionally independent of (T^*, Δ^*) given X, in which case the observable cause-specific hazards for causes 1 and 2, in the presence of censoring, defined as

$$\lambda_d(t|x) = \lim_{h \to 0} h^{-1} P(t \le T < t+h, \Delta = d | T \ge t, X = x),\ d = 1, 2,$$

are the same as the cause-specific hazards of interest. In particular, $\lambda(t|x) = \lambda^*(t|x)$. With a sample of data $(T_i, \Delta_i, X_i), i = 1, \ldots, n$, the parameter β in the proportional hazards model can be estimated using the maximum partial likelihood estimator after treating the values T_i observed for individuals whose Δ_i is equal to zero or one as censored times.

If an individual dies and cause of failure information is not collected, classification is uncertain; hence, we define the missingness indicator $R_i = 1$ if Δ_i is known and $R_i = 0$

otherwise. Assume that, if a subject is censored, this is known, so $\Delta_i = 0$ implies $R_i = 1$ and $R_i = 0$ implies $\Delta_i = 1$ or 2. Unlike previous methods, we may also define auxiliary covariates A_i, which are not of inherent interest for modeling the cause-specific hazard of interest but may be related to the missingness mechanism. For example, A_i may be some post-treatment variable that may be related to the reason why the cause of death information was not collected but that would not be included in the model because it may affect the causal interpretation associated with the parameters for treatment effects. The observed data are then $O_i = (R_i, T_i, \Delta_i, X_i, A_i)$ if $R_i = 1$ and $O_i = (R_i, T_i, X_i, A_i)$ if $R_i = 0$, independent across i.

The imputation procedure relies on the assumption of missing at random, or the probability that cause of failure is missing given $\Delta_i(> 0)$ and $W_i = (T_i, X_i, A_i)$ depends only on W_i, the information always observed on all subjects, and not on the unobserved Δ_i,

$$P(R_i = 0|W_i, \Delta_i > 0, \Delta_i) = P(R_i = 0|W_i, \Delta_i > 0).$$

This assumption stipulates that R_i and Δ_i are independent given $\{W_i, I(\Delta_i > 0)\}$, expressed equivalently as

$$\begin{aligned} P(\Delta_i = 2|W_i, \Delta_i > 0, R_i = 0) &= P(\Delta_i = 2|W_i, \Delta_i > 0, R_i = 1) \\ &= P(\Delta_i = 2|W_i, \Delta_i > 0). \end{aligned} \tag{1.2}$$

The proposed imputation method exploits (1.2) as discussed in the next section.

Ordinarily, if no causes of failure are missing, auxiliary covariates are not used in estimating β. When cause of failure is missing for some subjects, the assumption that it is missing at random depending only on $I(\Delta_i > 0)$ and (T_i, X_i) may be untenable. However, it may be possible to identify auxiliary covariates such that the missing at random assumption is plausible if A_i is included as above. The proposed approach allows information from such A_i to be exploited to impute missing causes.

1.3 Imputation Procedure

As is customary, to form a completed data set, missing $D_i = I(\Delta_i = 2)$ values are imputed from the distribution of D_i conditional on the observed data. This distribution is Bernoulli with success probability $P(\Delta_i = 2|W_i, \Delta_i > 0, R_i = 0)$, which, by (1.2), equals $P(\Delta_i = 2|W_i, \Delta_i > 0) = \varrho(W_i)$, say. We will assume that $\varrho(W_i)$ may be specified as a parametric model in terms of a few unknown parameters γ and $\varrho(W_i) = \varrho(W_i, \gamma_0)$, where γ_0 is the true value of γ. A natural choice is the logistic regression model, logit $\varrho(W_i, \gamma) = W_i^T\gamma$, but we can make the model as flexible as necessary to give a reasonable fit to the true model of $\varrho(W_i)$, induced above, by choosing a suitable parametric model. For example, we might include higher order polynomials and interaction terms for $\varrho(W_i, \gamma)$.

From (1.2), the success probability for the imputation, which equals $\varrho(W_i, \gamma_0)$, is identical to $P(\Delta_i = 2|W_i, \Delta_i > 0, R_i = 1)$. This suggests that $\varrho(W_i, \gamma_0)$, and hence the imputation probability, may be deduced from the completed cases for whom $(R_i = 1, \Delta_i > 0)$. In particular, under the parametric model $\varrho(W_i, \gamma)$, the maximum likelihood estimator $\hat{\gamma}$ of γ may be obtained by fitting the model to the completed cases only, thus providing an estimate of $P(\Delta_i = 2|W_i, \Delta_i > 0, R_i = 0)$.

For given γ, let $D_{ij}(R_i, \gamma)$ be the imputation of D_i from the jth imputed data set. If cause of failure is known $(R_i = 1)$, we take $D_{ij}(R_i, \gamma)$ to be D_i. If cause of failure is missing $(R_i = 0)$, we randomly choose $D_{ij}(R_i, \gamma)$ to be one or zero with probabilities $\varrho(W_i, \gamma)$ and $\{1 - \varrho(W_i, \gamma)\}$, respectively.

The joint distribution of (W_i, D_i) and $\{W_i, D_{ij}(R_i, \gamma_0)\}$ may be seen to be the same. When $R_i = 1$, $D_{ij}(R_i, \gamma_0) = D_i$, and when $R_i = 0$, $\{W_i, I(\Delta_i > 0)\}$ arise from the distribution of the observed data and $D_{ij}(R_i, \gamma_0)$ from the conditional distribution of D_i given the observed data, so that $D_{ij}(R_i, \gamma_0)$ is a draw from the joint distribution of the full data. Therefore, if true parameters and a parametric model for ϱ were known, then a single imputation of any missing data is as good as if you could conduct the experiment with no missingness.

Since $\hat{\gamma}$ is the maximum likelihood estimator for γ, then for a correctly specified model $\varrho(W_i, \gamma)$, it is consistent and we can treat it as if it were the true parameter. Because we can now generate data that are asymptotically as good as the original experiment, we can fit the proportional hazards model to a completed data set. We can carry out the imputation procedure multiple times and average the maximum partial likelihood estimators. The resulting estimator is the multiple imputation estimator we propose. Because each estimate is consistent, their average is also.

Although Rubin (1987) suggests a method for estimating the variance of the average of quantities from m imputed data sets, it is not appropriate here; because we generate imputations from the conditional distribution of missing data given the observed evaluated at $\hat{\gamma}$, where $\hat{\gamma}$ is held fixed across j, our imputation is not proper in the sense of Rubin (1987). Results of Wang and Robins (1998) indicate that under these conditions, which they refer to as type B multiple imputation, Rubin's variance expression will yield an inconsistent estimator for the true sampling variance. Consequently, we derive a variance estimator directly which accounts for all sources of variability, including the variability in $\hat{\gamma}$.

We would like to make a few remarks regarding the probability model $\varrho(W_i, \gamma)$. It is easy to show that $\varrho(W_i) = P(\Delta_i = 2|W_i, \Delta_i > 0)$ is related to the ratio of cause-specific hazards for the two failure types, conditional on (X, A) and with $w = (t, x, a)$, by

$$\frac{\lambda(t|x,a)}{\lambda_1(t|x,a)} = \left\{\frac{\varrho(w)}{1-\varrho(w)}\right\}. \tag{1.3}$$

This implies that the functional relationship of $\varrho(W_i)$ to W_i is induced from the ratio of cause-specific hazards. Note that the cause-specific hazards in (1.3) are conditional on both the covariates of interest X and the auxiliary covariates A and may not necessarily be the same as the cause-specific hazard of interest given in (1.1), which only conditions on X. For convenience, we have used a parametric model to model a relationship which is of no inherent interest to us and one which would be left arbitrary if there were no missing cause of death information. Therefore, as pointed out by Satten, Datta, and Williamson (1998), it will be important to examine the robustness of our estimator to misspecification of this probability model. This issue, although difficult to establish theoretical properties for, will be considered empirically in Section 5.

1.4 Asymptotic Properties

In establishing the consistency and asymptotic normality of imputation estimators, we assume that both the proportional hazards model (1.1) and the model for the probability that a missing cause is that of interest $\varrho(W_i, \gamma)$ are correctly specified. The results are listed below while the proofs are outlined in Appendix A.

Let

$$\mu_X(t) = \frac{E\{Xe^{\beta_0^T X} I(T \geq t)\}}{E\{e^{\beta_0^T X} I(T \geq t)\}}.$$

Also denote $\varrho_\gamma(W)$ as the derivative of $\varrho(W, \gamma)$ with respect to γ evaluated at γ_0, and

$$I_\gamma = E\left[\{\varrho_\gamma(W)\}^{\otimes 2}\frac{P(R=1, \Delta > 0|W)}{\varrho(W)\{1-\varrho(W)\}}\right].$$

Proposition 1 *Each single imputation estimator, $\hat{\beta}_j (j = 1, \ldots, m)$, is consistent and $n^{1/2}(\hat{\beta}_j - \beta_0)$ is asymptotically normal with asymptotic variance equal to $V_S^{-1} V_{SI} V_S^{-1}$, where*

$$\begin{aligned} V_{SI} &= V_S + E[\{X - \mu_X(T)\}P(\Delta > 0|W)\varrho_\gamma^T(W)]I_\gamma^{-1} \\ &\quad E[\varrho_\gamma(W)P(\Delta > 0|W)\{X - \mu_X(T)\}^T] \\ &\quad -E[\{X - \mu_X(T)\}P(R=1, \Delta > 0|W)\varrho_\gamma^T(W)]I_\gamma^{-1} \\ &\quad E[\varrho_\gamma(W)P(R=1, \Delta > 0|W)\{X - \mu_X(T)\}^T], \end{aligned}$$

and $V_S = \int E[\{X - \mu_X(t)\}^{\otimes 2} e^{\beta_0^T X} I(T \geq t)]\lambda(t)dt$ is the asymptotic variance of the score in the absence of missing cause of failure.

The variability of $\hat{\gamma}$ plays a role in both the second term and the third term, while the missingness contributes to their nonnegative difference. Without missing cause of failure, the second term and the third term are identical and would vanish, leaving $V_{SI} = V_S$, which leads to the familiar asymptotic results for partial likelihood estimators for the proportional hazards model.

Proposition 2 *The multiple imputation estimator, $\hat{\beta}$, is consistent and $n^{1/2}(\hat{\beta} - \beta_0)$ is asymptotically normal with asymptotic variance equal to $V_S^{-1} V_{MI} V_S^{-1}$, where*

$$V_{MI} = V_{SI} - (1 - m^{-1}) E[\{X - \mu_X(T)\}^{\otimes 2} P(R = 0|W) \varrho(W)\{1 - \varrho(W)\}].$$

It is evident that the second term, which measures the reduction in variability of the multiple imputation estimator over the single imputation estimator, is introduced through imputing the missing data multiple times. The more imputation we use, the greater the reduction in the asymptotic variance. The relative magnitude of V_{SI} and the second term will determine the number of imputations we might use.

Note that the estimate of the asymptotic variance can be obtained easily by manipulating readily available statistical software output. For example, $\hat{I}_\gamma$ can be obtained by inverting the variance estimate of $\hat{\gamma}$ and dividing it by n; $\hat{V}_S$ can be obtained by inverting the variance estimates from the m imputed data sets, dividing them by n, and averaging them across the m imputations; and all other quantities can be consistently estimated using their sample analogs.

1.5 Simulation Study

Several simulations were carried out to evaluate the performance of imputation estimators. We considered the case where the treatment indicator X_i is the only prognostic covariate with $P(X_i = 1) = P(X_i = 0) = 1/2$, and we also considered a single auxiliary covariate A_i, drawn from the standard normal distribution, independently of X_i. For each subject i, we took $T_i = \min(T_{2i}, T_{1i}, C_i)$, where T_{2i}, T_{1i}, and C_i were generated independently, conditional on (X_i, A_i), as described below; the resulting hazards for T_2, T_1, and C were thus the same as the cause-specific hazards $\lambda_j(t|x, a)$ for $j = 2, 1, 0$, respectively. Conditional on $(X_i = x, A_i = a)$, T_{2i} was generated from the exponential distribution with hazard function $\lambda(t|x, a) = \lambda(t|x) = \phi e^{\beta x}$, where $\phi = 1$, $\beta = -0.2$. Let logit $\varrho(W_i, \gamma) = \gamma_1 + \gamma_2 T_i + \gamma_3 X_i + \gamma_4 A_i$, where $\gamma = (1, -0.2, 0.5, 2)$. Then by (1.3), T_{1i} follows the Gompertz distribution with hazard function $\lambda_1(t|x, a) = \alpha \exp\{-\gamma_1 - \gamma_2 t - (\gamma_3 - \beta)x - \gamma_4 a\}$. In order to simplify the simulations, we let $\lambda(t|x, a) = \lambda(t|x)$. However, we emphasize that our interest focuses on the relationship $\lambda(t|x)$, specifically, the parameter β, which, in general, may not necessarily equal $\lambda(t|x,a)$. The censoring time C_i was generated from the right-truncated exponential distribution with hazard rate $\lambda_C = 0.01$ and truncating time $L = 5$, independently of all other random variables. With such a choice of parameter values, we will have, on average, 55% failures from the cause of interest, 30% failures from other causes, and 15% censored observations. The missing data mechanism was determined by logit $P(R_i = 0|\Delta_i > 0, W_i, \psi) = \psi_1 + \psi_2 T_i + \psi_3 X_i + \psi_4 A_i$, with different choices of ψ corresponding to different scenarios of missingness.

For sample sizes $n = 200, 500$, we carried out 10,000 simulations to compare the multiple imputation methods with $m = 1$ and $m = 10$ imputation and the complete case analysis. The results are summarized in Tables 1.1 and 1.2, where SEE denotes the empirical Monte Carlo average of our standard error estimates, SSE denotes the Monte Carlo standard error of the parameter estimates, and CP denotes the empirical coverage probability of the 95% confidence interval defined as $\hat{\beta} \pm 1.96\mathrm{SE}(\hat{\beta})$.

The scenario where $\psi = (-1, 0, 0, 0)$ corresponds to the case where the cause of death is missing completely at random. For this scenario, all analyses gave similar results, although the imputation methods were more efficient. When $\psi = (-1, 1, -3, 2)$, approximately the same proportion of missing observations were produced, but now the complete case analysis yielded large bias and poor coverage which becomes worse as the sample size increases. When $\psi = (-1, 2, -3, 2)$, the proportion of missing observations increased from 23% to 28% and the complete case analysis performed more poorly since it produced even larger biases and lower coverage probabilities. In all cases, imputation estimators were asymptotically unbiased, had the smallest standard errors, and achieved the nominal 95% coverage probability, with multiple imputation performing slightly better than single imputation. Also, the average of standard errors was very close to the Monte Carlo standard error, justifying our estimator of the asymptotic variance.

As pointed out in Satten et al. (1998), it is important to study robustness of parameter estimates from a semi- or non-parametric procedure when it uses data that were imputed using a parametric model. To investigate the robustness of the imputation procedure against misspecification of the parametric model for ϱ, we generated the survival times, T_{1i}, due to the competing causes, from gamma, log normal, log logistic as well as Weibull distributions. None of these distributions will induce a simple linear logistic regression model for ϱ. We report on the case where we generated T_{1i} from a Weibull distribution with shape parameter 0.5 and scale parameter $\exp[2\{\log(0.5) - \log(\phi) + \gamma_1 + \gamma_2 + (\gamma_3 - \beta)X + \gamma_4 A\}]$. In this case, the true model for ϱ is logit $\varrho(W_i) = \gamma_1 + \gamma_2 - (a-1)\log T_i + \gamma_3 X_i + \gamma_4 A_i$, yet we imputed missing cause of death fitting a simple linear logistic model. The results are included in Table 1.3, where we considered the missingness scenario $\psi = (-1, 2, -3, 2)$ for sample sizes $n = 200, 500$. As can be seen from the table, there is no substantial bias or loss of efficiency resulting from the use of the misspecified model. Although, not presented here, for all other distributions considered for T_{1i} mentioned above, the estimates for β never showed any appreciable bias and achieved the nominal coverage probability.

1.6 Breast Cancer Example

The data from a clinical trial in stage II breast cancer were analyzed to identify covariates that were significantly associated with death due to breast cancer. There were 169 patients enrolled in this study, among which 90 patients had censored death times. Among the 79 patients who died, 18 patients had cause of death unknown. For the remaining patients with known cause of death, 44 died from breast cancer and the other 17 died of other causes. Cummings et al. (1986) reported two covariates, presence of more than four positive nodes and having an ER-negative primary, as being significantly associated with overall survival. Goetghebeur and Ryan (1995) conducted a cause-specific survival analysis based on a standard proportional hazards structure for both failure types. We summarize their results in Table 1.4. Using the same data, we also derived the complete case estimator and the multiple imputation estimator as we will now describe. First, we had to establish a model for $\varrho(W)$, i.e., the probability that a cause of death is breast cancer, as a function of observed covariates W. For the covariates W, we considered ER-status, number of positive nodes, tumor size, treatment assignment, and time of death. One complication that we came across in analyzing the data is that, of the five patients who died and had ER-negative status, all died from breast cancer. Therefore, we could not use a logistic regression model that included ER-status as a covariate because the estimators would diverge. Since none of the patients with unknown cause of death were ER-negative, we used a logistic regression model for the subset of patients who were ER-positive. The logistic regression model considered all the covariates except ER-status and was derived using the subset of patients who were ER-positive with known cause of death. In conducting such an analysis, we found that, among the subset of ER-positive patients, none of the covariates except the intercept term was significant. Some minor adjustments to the theory developed in Section 1.4 were made to obtain correct asymptotic variance of the imputation estimator to account for the subsetting. The changes are minimal, basically inserting an ER-status indicator wherever necessary. It can be seen from Table 1.4 that the complete case estimates were biased and had large standard errors while the multiple imputation estimates were very close to those using the method of Goetghebeur and Ryan (1995). Similar analyses for noncancer causes were carried out and it turned out that none of the covariates were significantly associated with noncancer death.

1.7 Discussion

We have investigated a multiple imputation estimator for estimating regression coefficients in the competing risks model when the classification of cause of failure is missing for some individuals. The estimator and its asymptotic variance are easy to compute, lead to reliable inferences, and offer the data analyst flexibility. Based on the multiple imputation estimator, we can easily construct an estimator for the cumulative hazard function for time to failure from the cause of interest.

Table 1.1: Monte Carlo comparison of complete cases and imputation with sample size of 200

			MI		
ψ	% miss.		$m = 1$	$m = 10$	CC
$(-1, 0, 0, 0)$	22.86	Bias	-0.0026	-0.0020	-0.0174
		SEE	0.2084	0.2037	0.2295
		SSE	0.2080	0.2040	0.2301
		CP	0.9517	0.9519	0.9501
$(-1, 1, -3, 2)$	22.84	Bias	-0.0007	-0.0009	0.1257
		SEE	0.2066	0.2029	0.2603
		SSE	0.2087	0.2056	0.2690
		CP	0.9511	0.9504	0.9231
$(-1, 2, -3, 2)$	28.53	Bias	0.0028	0.0021	0.1662
		SEE	0.2116	0.2070	0.2812
		SSE	0.2144	0.2096	0.2944
		CP	0.9516	0.9493	0.9104

Table 1.2: Monte Carlo comparison of complete cases and imputation with sample size of 500

			MI		
ψ	% miss.		$m = 1$	$m = 10$	CC
$(-1, 0, 0, 0)$	22.88	Bias	-0.0001	-0.0006	-0.0161
		SEE	0.1306	0.1275	0.1433
		SSE	0.1310	0.1279	0.1443
		CP	0.9482	0.9484	0.9468
$(-1, 1, -3, 2)$	22.76	Bias	0.0019	0.0020	0.1349
		SEE	0.1291	0.1269	0.1623
		SSE	0.1300	0.1278	0.1693
		CP	0.9478	0.9482	0.8596
$(-1, 2, -3, 2)$	28.44	Bias	0.0001	0.0001	0.1744
		SEE	0.1321	0.1292	0.1752
		SSE	0.1319	0.1287	0.1820
		CP	0.9561	0.9544	0.8291

Table 1.3: Robustness of imputation against misspecification of the ϱ model

		MI		
n		$m = 1$	$m = 10$	CC
200	Bias	0.0016	0.0022	0.2910
	SEE	0.2450	0.2389	0.3606
	SSE	0.2456	0.2394	0.3842
	CP	0.9509	0.9512	0.8812
500	Bias	0.0008	0.0009	0.2984
	SEE	0.1518	0.1481	0.2225
	SSE	0.1512	0.1473	0.2353
	CP	0.9521	0.9529	0.7346

Table 1.4: Comparison of complete cases, Goetghebeur and Ryan, and imputation using the breast cancer data

	CC	GR	MI[a]
4+ nodes	0.71[0.3065]	0.57[0.2803]	0.60[0.2618]
ER-neg.	1.70[0.4861]	1.59[0.4822]	1.61[0.4794]

[a] $m = 10$

Chapter 2

Efficient Partial Likelihood Approach

2.1 Introduction

In a typical survival data analysis, a group of individuals are observed from some entry time until the occurrence of some particular event such as death. Often the observation of time to occurrence of the event is right-censored for some individuals as a result of staggered entry, finite study duration, withdrawal from the study, or loss to follow-up. Sometimes, the event can be classified into one of several categories, typically causes of death or other failures. For example, in a clinical trial that compares different therapies for breast cancer, interest may focus on death from breast cancer even though patients may die from other causes. In such cases, the theory of competing risks can be applied to assess the effects of covariates on cause-specific hazards, e.g., perform a standard proportional hazards analysis treating failure types which are not of interest as censored observations (Prentice and Kalbfleisch, 1978; Cox and Oakes, 1984; Goetghebeur and Ryan, 1995). In some circumstances, patients are known to die but the cause of death is unavailable, e.g., whether death is attributable to the cause of interest or other causes may require documentation with information that is not collected or lost or cause may be difficult for investigators to determine for some patients (Anderson, Goetghebeur, and Ryan, 1996). In such cases, excluding the missing observations from the analysis or treating them as censored may yield biased estimates and erroneous inferences. Under the assumption that the probability of having a missing cause of death may depend on time but not on covariates and that the baseline cause-specific hazards are proportional, Goetghebeur and Ryan (1995) proposed an approach that utilizes two types of partial likelihood (Cox, 1972, 1975). One is based on a full partial likelihood described in details in Section 2.3 (c.f., Holt, 1978; Kalbfleisch and Prentice, 1980; Dewanji, 1992). The other is a modified partial likelihood.

We extend their ideas to the more general settings where the probability of having a missing cause of death may depend on the covariates as well as time and where the ratio of the two baseline cause-specific hazards may also depend on time. This is achieved through the construction of an estimator using the full partial likelihood above. We show that the resulting estimator is consistent, asymptotically normal, and semiparametric efficient, under the more general missingness assumptions.

We introduce our notation and assumptions in Section 2.2. In Section 2.3, we propose the estimator which arises as the solution to the estimating equation based on the informative partial likelihood. Consistency and asymptotic normality of the resulting estimator will then follow from the martingale theory. Semiparametric efficiency can be established using semiparametric theory. Simulation results are also presented to compare the performance of our estimator with that of the complete-case estimator and that of the Goetghebeur and Ryan estimator. We conclude with an application followed by a brief discussion.

2.2 Notation and Assumptions

In this article, we consider a sample of n independent individuals, each of whom can die of fail from one of two possible causes which we refer to as causes two and one, respectively, or can be subject to a noninformative censoring mechanism. Typically, the data for individual i are $\{T_i, \Delta_i, X_i\}$, where T_i is the time to failure or censoring; Δ_i is an indicator taking values zero, one, or two, as the ith individual was censored, died from cause one, or died from cause two, respectively; X_i denotes a vector of covariates. Let $\lambda_\delta(t|x), \delta = 2, 1, 0$ be the cause-specific hazards for failure from cause two, failure from cause one, or censoring, respectively. Suppose that the cause-specific hazards for causes two and one follow proportional hazards relationships, namely,

$$\lambda_\delta(t|x) = \lambda(t) r_\delta(t, x, \beta),\ \delta = 1, 2, \tag{2.1}$$

where β is an unknown q-dimensional vector of parameters and $\lambda(t)$ is the common unspecified baseline cause-specific hazard. No assumptions are made on the cause-specific hazard of censoring, $\lambda_0(t|x)$, or the marginal distribution of X, $p_X(x)$.

Note that we allow the functions r_1 and r_2 to depend on time and the covariates through a finite set of parameters. This is a generalization of the case where the ratio of the two baseline cause-specific hazards is constant or piecewise constant over time. For example, if given $X = x$, T_2 follows an exponential distribution with constant hazard $\lambda_2(t|x) = \phi e^{\beta x}$ and T_1 follows a Gompertz distribution with hazard function $\lambda_1(t|x) = e^{\gamma_1 t + \gamma_2 x}$, then the ratio between the two baseline cause-specific hazards is equal to $\lambda_2(t)/\lambda_1(t) = \phi e^{-\gamma_1 t}$, which is not constant unless $\gamma_1 = 0$. Note that, however, only parameters associated with the covariates are of inherent interest. To avoid nonidentifiability problems, we also assume that all information about time that is common to the two cause-specific hazards has been incorporated into the common baseline cause-specific hazard.

Also note that we could have formulated the model with separate regression parameter vectors $\{\beta_1, \beta_2\}$ for the two failure causes. There may be examples, however, where some parameters are common to the two failure causes. Therefore, it is convenient to formulate the model with one vector of parameters β which contains all the different parameters in $\{\beta_1, \beta_2\}$ (c.f., Andersen, Borgan, Gill, and Keiding, 1997, p. 478).

In some circumstances, cause of failure might be missing for some individuals, in which case, we use R_i as the missingness indicator, taking values one or zero as the cause of failure $\Delta_i(> 0)$ is observed or missing. We assume that cause of failure is missing at random (Rubin, 1976), in the sense that the probability of having a missing cause of failure does not depend on the latent cause of failure, i.e.,

$$P(R_i = 1|Z_i, \Delta_i > 0) \quad = \quad \pi(T_i, X_i), \tag{2.2}$$

where π is an unknown function of time and covariates, taking values in the unit interval. Note that we allow the missingness probability to depend on both time and covariates, whereas Goetghebeur and Ryan (1995) allows the missingness probability to depend on time only. In the presence of missing cause of failure, the observed data for the ith individual can be summarized as $O_i = \{R_i, T_i, I(\Delta_i = 0), R_i I(\Delta_i = 1), R_i I(\Delta_i = 2), X_i\}$.

2.3 Parameter Estimation

For an uncensored individual, one of the following three types of events can occur at the time of failure, i.e., failure from cause one, failure from cause two, or failure with unknown cause. Let $N_i(t) = \{N_{i1}(t), N_{i2}(t), N_{iu}(t)\}$ be a multivariate counting process indicating the failure type. Based on the assumptions (2.1) and (2.2), the corresponding intensity processes are given by

$$\begin{aligned}
\lambda^*_{i1}(t, X_i) &= Y_i(t)\pi(t, X_i)r_1(t, X_i, \beta_0)\lambda(t), \\
\lambda^*_{i2}(t, X_i) &= Y_i(t)\pi(t, X_i)r_2(t, X_i, \beta_0)\lambda(t), \\
\lambda^*_{iu}(t, X_i) &= Y_i(t)\{1 - \pi(t, X_i)\}r_{\cdot}(t, X_i, \beta_0)\lambda(t),
\end{aligned}$$

respectively, where $Y_i(t) = I(T_i \geq t)$ denotes whether individual i is at risk at time t, $r_{\cdot}(t, x, \beta) = r_1(t, x, \beta) + r_2(t, x, \beta)$, and β_0 denotes the true value of β.

Under the missing-at-random assumption, we propose to use the full partial likelihood, which is based on the conditional probabilities of an event of specified type, given that one event occurs, but without conditioning on the type of event, i.e.,

$$
\begin{aligned}
L(\beta) &= \prod_{t\geq 0}\prod_{i=1}^{n}\left[\frac{r_1(t,X_i,\beta)}{\sum_{j=1}^{n} r_{\cdot}(t,X_j,\beta)Y_j(t)}\right]^{dN_{i1}(t)} \\
&\quad\times\left[\frac{r_2(t,X_i,\beta)}{\sum_{j=1}^{n} r_{\cdot}(t,X_j,\beta)Y_j(t)}\right]^{dN_{i2}(t)} \\
&\quad\times\left[\frac{r_{\cdot}(t,X_i,\beta)}{\sum_{j=1}^{n} r_{\cdot}(t,X_j,\beta)Y_j(t)}\right]^{dN_{iu}(t)},
\end{aligned}
$$

where $\prod_{t\geq 0}$ denotes product-integration (c.f., Gill and Johansen, 1990).

Let $\{r_d'(t,X_i,\beta), r_d''(t,X_i,\beta)\}$ denote the first two partial derivatives of $r_d(t,X_i,\beta)$ with respect to β for $d=1,2$, and let

$$
m(t,\beta) = \frac{\sum_{j=1}^{n} r_{\cdot}'(t,X_j,\beta)Y_j(t)}{\sum_{j=1}^{n} r_{\cdot}(t,X_j,\beta)Y_j(t)},
$$

$$
v(t,\beta) = \frac{\sum_{j=1}^{n} r_{\cdot}''(t,X_j,\beta)Y_j(t)}{\sum_{j=1}^{n} r_{\cdot}(t,X_j,\beta)Y_j(t)},
$$

then the corresponding score equation is $U(\beta)=0$, where

$$
\begin{aligned}
U(\beta) &= \sum_{i=1}^{n}\Bigg[\int\frac{r_1'(t,X_i,\beta)}{r_1(t,X_i,\beta)}dN_{i1}(t) \\
&\quad+\int\frac{r_2'(t,X_i,\beta)}{r_2(t,X_i,\beta)}dN_{i2}(t) \\
&\quad+\int\frac{r_{\cdot}'(t,X_i,\beta)}{r_{\cdot}(t,X_i,\beta)}dN_{iu}(t) \\
&\quad-\int m(t,\beta)dN_{i\cdot}(t)\Bigg],
\end{aligned}
$$

and the observed information is

$$
\begin{aligned}
I(\beta) &= -\sum_{i=1}^{n}\Bigg\{\int\left[\frac{r_1''(t,X_i,\beta)}{r_1(t,X_i,\beta)}-\left\{\frac{r_1'(t,X_i,\beta)}{r_1(t,X_i,\beta)}\right\}^{\otimes 2}\right]dN_{i1}(t) \\
&\quad+\int\left[\frac{r_2''(t,X_i,\beta)}{r_2(t,X_i,\beta)}-\left\{\frac{r_2'(t,X_i,\beta)}{r_2(t,X_i,\beta)}\right\}^{\otimes 2}\right]dN_{i2}(t) \\
&\quad+\int\left[\frac{r_{\cdot}''(t,X_i,\beta)}{r_{\cdot}(t,X_i,\beta)}-\left\{\frac{r_{\cdot}'(t,X_i,\beta)}{r_{\cdot}(t,X_i,\beta)}\right\}^{\otimes 2}\right]dN_{iu}(t) \\
&\quad-\int\left[v(t,\beta)-\{m(t,\beta)\}^{\otimes 2}\right]dN_{i\cdot}(t)\Bigg\},
\end{aligned}
$$

where $N_{i\cdot}=N_{i1}+N_{i2}+N_{iu}$ is the counting process of failure for the ith individual.

Note that $U(\beta_0)$ is a martingale and hence the score equation can be used to obtain a consistent estimator of β, say $\hat{\beta}_n$. In addition, it is straightforward to show that, when evaluated at the truth, the observed information matrix $I(\beta_0) = -\partial U(\beta)/\partial\beta|_{\beta=\beta_0}$ has the same expectation as the covariation process of the score vector $U(\beta_0)$. Consistency and asymptotic normality of $\hat{\beta}_n$ follow from arguments similar to those used by Andersen and Gill (1982) and the asymptotic variance can be consistently estimated by $n^{-1}\{I(\hat{\beta}_n)\}^{-1}$.

Using semiparametric theory (e.g., Newey, 1990; Bickel, Klaasen, Ritov, and Wellner, 1993; Robins, Rotnitzky, and Zhao, 1994), we show that the influence function of our proposed estimator is the efficient influence function. The proof is outlined in Appendix B.

2.4 Simulation Study

Several simulations were carried out to evaluate the performance of different estimators. We considered the situation where the treatment assignment X_i was the only covariate and $X_i \sim$ Bernoulli(0.5). For each subject i, we took $T_i = \min(T_{2i}, T_{1i}, C_i)$, where T_{2i}, T_{1i}, and C_i were generated independently, conditional on X_i, as described below; the resulting hazards for T_2, T_1, and C_0 were thus the same as the cause-specific hazards $\lambda_\delta(t|x)$ for $\delta = 2, 1, 0$, respectively. Conditional on $X_i = x$, T_{2i} was generated from an exponential distribution with constant hazard $\lambda_2(t|x) = \phi e^{\beta x}$, where $\phi = 0.8$ and $\beta = 0.5$. Comparison of different approaches was based on the estimation of β. In addition, T_{1i} was generated from a Gompertz distribution with hazard function $\lambda_1(t|x) = e^{\gamma_1 t + \gamma_2 x}$. Consequently, the ratio between the two baseline cause-specific hazards was $\lambda_2(t)/\lambda_1(t) = \phi\exp(-\gamma_1 t)$, which is not constant over time unless $\gamma_1 = 0$. However, as pointed out by Goetghebeur and Ryan (1995), the efficient partial likelihood estimator is more sensitive than the Goetghebeur and Ryan partial likelihood estimator to violations of the proportionality assumption relating the two baseline cause-specific hazards. Therefore, we only need to consider the case when $\gamma_1 = 0$. Furthermore, the censoring time C_i was generated from an exponential distribution with constant hazard $\lambda_C = 0.4$. Finally, the missingness indicator R_i was generated from a Bernoulli distribution with success probability depending only on (T_i, X_i) to comply with the MAR assumption. In particular, we let logit $\pi(T_i, X_i) = \psi_0 + \psi_1 T_i + \psi_2 X_i$, with different values of ψ corresponding to different scenarios of missingness.

For sample sizes $n = 200, 500$, we carried out 1000 simulations to compare different approaches. With such parameter values, we will have, on average, 34% to 46% failures from cause two ($\Delta_i = 2$), 36% to 52% failures from cause one ($\Delta_i = 1$), and 14% to 18% censored observations ($\Delta_i = 0$). For all cases of missing data mechanism we considered, the proportion of missing observations ($R_i = 0$) ranged between 17% and 29%. The results of the comparison among the naive complete case analysis (CC), the Goetghebeur and Ryan partial likelihood approach (GR), and the efficient partial likelihood approach (EPL) are summarized in terms of the sampling bias, the sampling standard error (SSE), the sampling average of the standard error estimates (SEE), and the empirical coverage probability (CP) of the asymptotic 95% confidence interval in Tables 2.1 and 2.2.

The scenario where $\psi = (1, 0, 0)$ corresponds to the case where cause of failure is missing completely at random. For this scenario, all analyses give similar results although our estimates are the most efficient. When $\psi = (5, -8, 0)$, which corresponds to the case where the probability of having a missing cause of failure depends on time only, the naive complete case estimator is biased and has a coverage probability substantially lower than the nominal level, but the Goetghebeur and Ryan estimator still performs well because their missingness assumption is still met. For $\psi = (1, 1, -1.5)$, in which case, the probability of having a missing cause of failure depends on both time and covariate, the naive complete case estimator is again biased as expected, so is the Goetghebeur and Ryan estimator. Furthermore, the Goetghebeur and Ryan variance estimator underestimates the true sampling variation, resulting in a further reduced coverage probability. In all cases, our efficient likelihood approach performs well.

2.5 Breast Cancer Example

The data from a clinical trial in elderly women with stage II breast cancer were analyzed to identify covariates that were significantly associated with death due to breast cancer. There were 169 eligible patients enrolled in this study, among which 90 patients had censored death times. Among the 79 patients who died, 18 patients had cause of death unknown. For the remaining patients with known cause of death, 44 died from breast cancer and the other 17 died of other causes. Cummings et al. (1986) reported two covariates, presence of 4-10 positive axillary lymph nodes and having an estrogen receptor (ER) negative primary tumor, as being significantly associated with overall survival. Goetghebeur and Ryan (1995) conducted a cause-specific survival analysis based on the standard proportional hazards structure for both failure types. We summarize their results along with our efficient estimates in Table 2.3, where the numbers inside the brackets denote the standard errors associated with the parameter estimates. It can be seen from the table that the hazard of death from breast cancer is significantly associated with the ER-status, but no firm conclusion can be drawn for the association of the hazard of death from breast cancer with the number of positive axillary lymph nodes. In addition, our efficient estimates are closer to the Goetghebeur and Ryan estimates than to the naive complete case estimates.

2.6 Discussion

We have proposed an approach for estimating regression coefficients in the competing risks model when the classification of cause of failure is missing for some individuals. The procedure is applicable in many situations and the resulting estimator is semiparametric efficient.

Table 2.1: Monte Carlo comparison of complete cases, Goetghebeur and Ryan, and efficient likelihood approach with sample size of 200

ψ	γ_2		CC	GR	EPL
$(1, 0, 0)$	0.9	Bias	0.0162	-0.0123	-0.0157
		SSE	0.3163	0.2986	0.2873
		SEE	0.3055	0.2888	0.2792
		CP	0.939	0.940	0.941
$(5, -8, 0)$	0.9	Bias	-0.1666	-0.0094	-0.0115
		SSE	0.2842	0.2819	0.2786
		SEE	0.2839	0.2733	0.2701
		CP	0.904	0.946	0.947
$(1, 1, -1.5)$	-0.5	Bias	-0.2115	0.1793	0.0097
		SSE	0.2707	0.3070	0.2378
		SEE	0.2738	0.2656	0.2381
		CP	0.880	0.891	0.956

Table 2.2: Monte Carlo comparison of complete cases, Goetghebeur and Ryan, and efficient likelihood approach with sample size of 500

ψ	γ_2		CC	GR	EPL
$(1, 0, 0)$	0.9	Bias	0.0249	-0.0044	-0.0050
		SSE	0.1893	0.1829	0.1761
		SEE	0.1900	0.1799	0.1743
		CP	0.951	0.946	0.955
$(5, -8, 0)$	0.9	Bias	-0.1637	-0.0028	-0.0035
		SSE	0.1839	0.1751	0.1738
		SEE	0.1772	0.1708	0.1689
		CP	0.835	0.951	0.953
$(1, 1, -1.5)$	-0.5	Bias	-0.2133	0.1694	0.0027
		SSE	0.1667	0.1893	0.1478
		SEE	0.1708	0.1640	0.1494
		CP	0.767	0.793	0.962

Table 2.3: Comparison of complete cases, Goetghebeur and Ryan, and efficient likelihood approach using the breast cancer data

	CC	GR	EPL
4+ nodes	0.66[0.3090]	0.57[0.2803]	0.57[0.2815]
ER-neg.	1.71[0.4865]	1.59[0.4822]	1.56[0.4770]

Chapter 3

Inverse Probability Weighting Approach

3.1 Introduction

In a typical clinical trial, researchers are interested in the effects of a set of prognostic factors on the hazard of death or relapse from some specific disease of interest, even though patients may die from other competing causes. For example, in a clinical trial that compares different therapies for breast cancer in the population of elderly women, interest focuses on death from breast cancer even though patients may die from cardiac or vascular disease. Often the observation of time to failure is right-censored for some individuals due to incomplete follow-up. In some circumstances, it may also be the case that patients are known to die but the cause of death is unavailable, e.g., whether death is attributable to the cause of interest or other causes may require documentation with information that is not collected or lost or cause may be difficult for investigators to determine for some patients (Andersen, Goetghebeur, and Ryan, 1996). If there were no missing cause of failure, the standard proportional hazards model can be used to model the relationship of the cause-specific hazard of interest with respect to the prognostic factors by treating deaths from competing causes as censored observations, and the regression coefficients can be estimated by maximizing the partial likelihood (Cox, 1972, 1975). However, when cause of failure is missing, excluding the observations with missing cause of failure from the analysis may yield inefficient or even biased estimates if cause of failure is not missing completely at random. In addition, treating missing observations as censored is certain to yield biased estimates if some of the missing cause of death is the cause of interest. With missing cause of failure, Goetghebeur and Ryan (1995) proposed an approach by making assumptions directly on the relationship between the cause-specific hazard of interest and that of competing causes, i.e., the two baseline cause-specific hazards are assumed proportional, although this may be relaxed.

In some circumstances, it may also be the case that auxiliary covariates ,which are not of inherent interest for modeling the cause-specific hazard of interest, but which may be related to the missingness mechanism, are available. For example, we may be able to identify some post-treatment variable which is related to the reason why the cause of death information was not collected, but we would not include it in the proportional hazards model because it may affect the causal interpretation associated with the parameters for treatment effects. When the auxiliary covariates are included in the model of missingness mechanism, it might be reasonable to assume that cause of failure is missing at random, whereas the MAR assumption might not hold if the auxiliary covariates are not included.

In this article, we take a different approach by using two parametric models to model the missingness probability and the probability that a missing cause is the cause of interest, respectively, while allowing the inclusion of additional auxiliary covariates. Using semiparametric theory (e.g., Newey, 1990; Bickel, Klaasen, Ritov, and Wellner, 1993; Robins, Rotnitzky, and Zhao, 1994), we identify various classes of inverse probability weighted (IPW) semiparametric estimators. In Section 3.3, we obtain the space of all full data influence functions and the full data efficient score. In Section 3.4, we derive the space of all observed data influence functions. In Section 3.5, we introduce a class of estimating equations whose solutions correspond to all semiparametric estimators when the missingness mechanism is not known but can be correctly specified through a parametric model. In Section 3.6, we construct a class of estimating equations whose solutions define all doubly robust semiparametric estimators when either of the two parametric models is correctly specified. In Section 3.7, we identify the observed data efficient score and construct an estimating equation based on the observed data efficient score. The solution to the estimating equation is the locally semiparametric efficient estimator, which will be fully efficient if all parametric models are correctly specified. Simulation results are then presented to compare three IPW semiparametric estimators with the complete case estimator and the imputation estimator, followed by a revisit of the breast cancer example using the doubly robust IPW semiparametric estimator. A brief discussion is also provided to conclude this article.

3.2 Notation and Assumptions

We are going to use the theory for estimation in arbitrary semiparametric models with missing data as developed by Robins, et al. (1994). Define a semiparametric estimator to be one that is consistent and asymptotically normal under the restrictions imposed by the model. To avoid super-efficiency, we will only consider regular estimators, for which the convergence to their limiting distributions is locally uniform. Also, an estimator $\hat{\beta}_n$ of β_0 is asymptotically linear with influence function φ if

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2}\sum_{i=1}^{n}\varphi_i + o_p(1).$$

Consider the generic semiparametric model indexed by some finite, say q-dimensional parameter of interest β and the infinite dimensional nuisance parameter η. The Hilbert space $\mathcal{H}$ consists of all $q \times 1$ random vectors of mean zero and square integrable measurable functions of Z equipped with covariance inner product. The nuisance tangent space Λ is defined to be the mean square closure of the set of all random vectors BS_γ, where S_γ is the score for γ in some regular parametric submodel and B is a conformable constant matrix with q rows. Also, $\Pi[h|\Lambda]$ denotes the projection of any vector $h \in \mathcal{H}$ on a closed linear space such as Λ. The semiparametric variance bound equals the inverse of the variance of $S_{\mathit{eff}} = \Pi[S_\beta|\Lambda^\perp]$, where S_β is the score for β and S_{eff} is called the efficient score. In addition, we will use the superscript "F" to distinguish the full data model from the observed data model. For example, we let S_β^F, Λ^F, S_{eff}^F, and $\Lambda_*^{F\perp}$ be the full data score for β, the full data nuisance tangent space, the full data efficient score, and the space of full data influence functions, respectively.

The complete data for a single observation can be represented as $Z = (T, \Delta, X, A)$, where T is the observed time to failure or censoring, Δ is an indicator taking values two, one or zero as the individual failed from cause two, failed from cause one, or was censored, respectively. Without loss of generality, assume cause two is the cause of interest and cause one is the competing cause. In addition, X denotes the vector of covariates of interest, which is assumed to be related to the cause-specific hazard of interest through the proportional hazards model,

$$\lambda(t|X) = \lambda(t)e^{\beta^T X}, \tag{3.1}$$

where β is the q-dimensional vector of regression coefficients and $\lambda(t)$ is the unspecified baseline cause-specific hazard. Also, A denotes auxiliary covariates which are not of inherent interest for modeling the cause-specific of interest but which may be related to the missingness mechanism.

In certain instances, patients are known to die but the cause of death information is not available, in which case, we use R as the complete case indicator taking values one or zero as the cause of death is known or missing, so that the observed data for a typical observation can be summarized as $O = \{R, G_R(Z)\} = \{R, T, X, A, I(\Delta = 0), RI(\Delta = 1), RI(\Delta = 2)\}$. Write $W = (T, X, A)$, then $G_1(Z) = Z = (W, \Delta)$. Also let $Q = \{W, I(\Delta > 0)\}$, which denotes variables that are always observed, then $G_0(Z) = Q$. Furthermore, assume that

$$P(R = 1|W, \Delta, \Delta > 0) = P(R = 1|W, \Delta > 0), \tag{3.2}$$

then $\{R \perp\!\!\!\perp \Delta | Q\}$, so that (3.2) implies that cause of failure is missing at random (Rubin, 1976). Write $\pi(W) = P(R = 1|W, \Delta > 0)$ and assume $\pi(W) > \epsilon > 0$ with probability one so that the probability of observing complete data is bounded away from zero.

3.3 Full Data Influence Functions

In the absence of missing data, the density for a typical observation can be factorized as

$$\begin{aligned} & P(T = t, \Delta = \delta, X = x, A = a) \\ = \; & p_{A|T,\Delta,X}(a|t, \delta, x) \\ & \times \exp[-\{\Lambda(t|x) + \Lambda_1(t|x) + \Lambda_0(t|x)\}] \\ & \times \lambda(t|x)^{I(\delta=2)} \lambda_1(t|x)^{I(\delta=1)} \lambda_0(t|x)^{I(\delta=0)} \\ & \times p_X(x), \end{aligned}$$

where $p_{A|T,\Delta,X}$ is the conditional density of A given (T, Δ, X), $\{\lambda_1(t|x), \lambda_0(t|x)\}$ are the conditional cause-specific hazard for failure from the competing cause and the conditional cause-specific hazard for censoring, given $X = x$, respectively, $\{\Lambda(t|x), \Lambda_1(t|x), \Lambda_0(t|x)\}$ are the corresponding cumulative cause-specific hazards, and p_X is the marginal density of X.

Therefore, the log-likelihood for a typical observation can be written as

$$\begin{aligned} \ell^F(Z) \; = \; & -\Lambda(T|X) + I(\Delta = 2) \log \lambda(T|X) \\ & -\Lambda_1(T|X) + I(\Delta = 1) \log \lambda_1(T|X) \\ & -\Lambda_0(T|X) + I(\Delta = 0) \log \lambda_0(T|X) \\ & + \log p_X(X) \\ & + \log p_{A|T,\Delta,X}(A|T, \Delta, X). \end{aligned} \tag{3.3}$$

Write $\Lambda(t) = \int_0^t \lambda(s) ds$, then, under assumption (3.1), (3.3) reduces to

$$\begin{aligned} \ell^F(\beta, Z) \; = \; & -\Lambda(T) e^{\beta^T X} + I(\Delta = 2)\{\log \lambda(T) + \beta^T X\} \\ & -\Lambda_1(T|X) + I(\Delta = 1) \log \lambda_1(T|X) \\ & -\Lambda_0(T|X) + I(\Delta = 0) \log \lambda_0(T|X) \\ & + \log p_X(X) \\ & + \log p_{A|T,\Delta,X}(A|T, \Delta, X). \end{aligned} \tag{3.4}$$

Since the nuisance parameters $\{\lambda(t), \lambda_1(t|x), \lambda_0(t|x), p_X(x), p_{A|T,\Delta,X}(a|t, \delta, x)\}$ are functionally independent and separate from each other in the log-likelihood (3.4), the full data nuisance tangent space can be written as a direct sum of five orthogonal spaces,

$$\Lambda^F = \Lambda_{1s} + \Lambda_{2s} + \cdots + \Lambda_{5s},$$

where Λ_{1s} is associated with $\lambda(t)$, Λ_{2s} is associated with $\lambda_1(t|x)$, Λ_{3s} is associated with $\lambda_0(t|x)$, Λ_{4s} is associated with p_X, and Λ_{5s} is associated with $p_{A|T,\Delta,X}$, respectively.

It is straightforward to show that

$$\Lambda_{1s} = \left\{ \int \alpha(t) dM(t) : \forall \alpha^{q\times 1}(t) \right\},$$

where $dM(t) = dN(t) - \lambda(t) e^{\beta_0^T X} I(T \geq t)dt, N(t) = I(T \leq t, \Delta = 2)$.

On the other hand, had no restrictions been put on the form of the cause-specific hazard of interest, the log-likelihood (3.3) would correspond to a saturated model, so that the entire full data Hilbert space can be written as the direct sum of five orthogonal spaces,

$$\mathcal{H}^F = \Lambda_{1s}^* + \Lambda_{2s} + \cdots + \Lambda_{5s},$$

where Λ_{1s}^* is associated with $\lambda(t|x)$.

It is straightforward to show that

$$\Lambda_{1s}^* = \left\{ \int a(t,X) dM(t) : \forall a^{q\times 1}(t,X) \right\}.$$

Therefore, the space orthogonal to the full data nuisance tangent space, i.e. $\Lambda^{F\perp}$, is the subspace of Λ_{1s}^* that is orthogonal to Λ_{1s}. By the projection theorem, it is straightforward to show that

$$\Lambda^{F\perp} = \left\{ \int \{a(t,X) - \mu_a(t)\} dM(t) : \forall a^{q\times 1}(t,X) \right\}, \tag{3.5}$$

where

$$\mu_a(t) = \frac{E\{a(t,X) e^{\beta_0^T X} I(T \geq t)\}}{E\{e^{\beta_0^T X} I(T \geq t)\}}.$$

For an element of $\Lambda^{F\perp}$, say $\varphi^F(Z)$, to be an influence function for a semiparametric estimator for β, we must also have $E\{\varphi^F(Z) S_\beta^{FT}(Z)\} = I_q$, where $S_\beta^F(Z)$ is the full data score for β and I_q is the $q \times q$ identity matrix.

From (3.4),

$$S_\beta^F(Z) = \int X dM(t). \tag{3.6}$$

By standard properties for martingales (e.g., Fleming and Harrington, 1991),

$$\begin{aligned} & E\left[\int \{a(t,X) - \mu_a(t)\} dM(t)\ S_\beta^F(Z) \right] \\ = & \int E[\{a(t,X) - \mu_a(t)\}\{X - \mu_X(t)\}^T e^{\beta_0^T X} I(T \geq t)] \lambda(t) dt \\ \equiv & V(a,X), \end{aligned}$$

where "≡" means "denoted as" and

$$\mu_X(t) = \frac{E\{X e^{\beta_0^T X} I(T \geq t)\}}{E\{e^{\beta_0^T X} I(T \geq t)\}}.$$

Therefore, the space of full data influence functions is given by

$$\Lambda_*^{F\perp} = \left\{ V^{-1}(a, X) \int \{a(t, X) - \mu_a(t)\} dM(t) : \forall a^{q\times 1}(t, X) \right\}. \tag{3.7}$$

In addition, by (3.5) and (3.6), the full data efficient score is given by

$$S_{\mathit{eff}}^F(Z) = \int \{X - \mu_X(t)\} dM(t). \tag{3.8}$$

3.4 Observed Data Influence Functions

First suppose that $\pi(W)$ is completely known as in a designed study, then

$$P(R = 1|Z) = \pi(W)I(\Delta > 0) + I(\Delta = 0) \equiv \pi(Q).$$

By Proposition 8.1 of Robins et al. (1994), the space of all observed data influence functions is given by

$$\Lambda_{0*}^{\perp} = \frac{R}{\pi(Q)} \Lambda_*^{F\perp} + \Lambda_2, \tag{3.9}$$

where $\Lambda_2 = \{L_2(O) \in \mathcal{H} : E\{L_2(O)|Z\} = 0\}$.

Write

$$L_2(O) = RL_{21}(Z) + (1 - R)L_{20}(Q), \tag{3.10}$$

then

$$E\{L_2(O)|Z\} = \pi(Q)L_{21}(Z) + \{1 - \pi(Q)\}L_{20}(Q).$$

Set $E\{L_2(O)|Z)\} = 0$, we have $L_{21}(Z) = -\frac{\{1-\pi(Q)\}}{\pi(Q)} L_{20}(Q)$. Substituting it into (3.10), a typical element of Λ_2 is given by

$$L_2(O) = -\frac{\{R - \pi(Q)\}}{\pi(Q)} L_{20}(Q), \tag{3.11}$$

where $L_{20}(Q)$ is an arbitrary $q \times 1$ function of Q satisfying $E\{L_{20}^T(Q)L_{20}(Q)\} < \infty$.

By (3.7), (3.9), and (3.11), a typical element of $\Lambda_{0*}^{\perp}$ is given by

$$\begin{aligned} \varphi_0(O) &= \frac{R}{\pi(Q)} V^{-1}(a, X) \int \{a(t, X) - \mu_a(t)\} dM(t) \\ &\quad - \frac{\{R - \pi(Q)\}}{\pi(Q)} L_{20}(Q). \end{aligned}$$

Now suppose that the missingness mechanism $\pi(W)$ is not known but we can correctly specify a parametric model, say $\pi(W) = \pi(W, \psi)$, then

$$\pi(Q, \psi) = \pi(W, \psi)I(\Delta > 0) + I(\Delta = 0).$$

By Proposition 8.1 of Robins et al. (1994), the space of all observed data influence functions is given by

$$\Lambda_*^\perp = \Lambda_{0*}^\perp - \Pi[\Lambda_{0*}^\perp | \Lambda_\psi],$$

where Λ_ψ is the nuisance tangent space for ψ. Therefore, a typical element of $\Lambda_*^\perp$ is given by

$$\varphi(O) = \varphi_0(O) - \Pi[\varphi_0(O) | \Lambda_\psi].$$

Note that the observed data likelihood for ψ is

$$\prod_{i=1}^{n} \{\pi(Q_i, \psi)\}^{R_i} \{1 - \pi(Q_i, \psi)\}^{1-R_i}.$$

Hence, the log-likelihood for ψ for a typical observation is

$$\ell(\psi, O) = R \log \pi(Q, \psi) + (1 - R) \log\{1 - \pi(Q, \psi)\}.$$

Consequently, the score vector for ψ is

$$S_\psi(O) = \frac{\{R - \pi(Q)\}\pi_\psi(Q)}{\pi(Q)\{1 - \pi(Q)\}}, \tag{3.12}$$

where $\pi_\psi(Q)$ denotes the partial derivative of $\pi(Q, \psi)$ with respect to ψ and evaluated at $\psi = \psi_0$. A typical element of Λ_ψ is given by BS_ψ for some arbitrary conformable matrix B with q rows. By (3.12) and (3.11), $\Lambda_\psi \subset \Lambda_2$.

3.5 Semiparametric Estimators

Assume that the parametric model for the missingness mechanism, $\pi(W) = \pi(W, \psi)$, is correctly specified. Let $\hat{\psi}_n$ be the MLE of ψ, and ψ_0 be the true value of ψ, then $\hat{\psi}_n \xrightarrow{p} \psi_0$.

It is shown in Section 3.4 that a typical element of $\Lambda_{0*}^\perp$ is given by

$$\varphi_0(O) = V^{-1}(a, X)\varphi_0^*(O),$$

where

$$\begin{aligned}
\varphi_0^*(O) &= \frac{R}{\pi(Q, \psi_0)} \int \{a(t, X) - \mu_a(t)\} dM(t) \\
&\quad - \frac{\{R - \pi(Q, \psi_0)\}}{\pi(Q, \psi_0)} L(Q) \\
&= \frac{R}{\pi(Q, \psi_0)} \int \{a(t, X) - \mu_a(t)\} dN(t) \\
&\quad - \frac{\{R - \pi(Q, \psi_0)\}}{\pi(Q, \psi_0)} L(Q) \\
&\quad - \frac{R}{\pi(Q, \psi_0)} \int \{a(t, X) - \mu_a(t)\} \lambda(t) e^{\beta_0^T X} I(T \geq t) dt.
\end{aligned}$$

Denote

$$\bar{a}(t,\beta,\psi) = \frac{\sum_{j=1}^n \frac{R_j}{\pi(Q_j,\psi)} a(t,X_j) e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^n \frac{R_j}{\pi(Q_j,\psi)} e^{\beta^T X_j} I(T_j \geq t)}.$$

By the WLLN and the LIE by conditioning on Z, we have that

$$\begin{aligned} & n^{-1} \sum_{j=1}^n \frac{R_j}{\pi(Q_j,\hat{\psi}_n)} a(t,X_j) e^{\beta_0^T X_j} I(T_j \geq t) \\ \xrightarrow{p} \; & E\left[\frac{R}{\pi(Q,\psi_0)} a(t,X) e^{\beta_0^T X} I(T \geq t)\right] \\ = \; & E\{a(t,X) e^{\beta_0^T X} I(T \geq t)\}. \end{aligned}$$

Similarly,

$$n^{-1} \sum_{j=1}^n \frac{R_j}{\pi(Q_j,\hat{\psi}_n)} e^{\beta_0^T X_j} I(T_j \geq t) \xrightarrow{p} E\{e^{\beta_0^T X} I(T \geq t)\}.$$

Therefore,

$$\bar{a}(t,\beta_0,\hat{\psi}_n) \xrightarrow{p} \mu_a(t).$$

On the other hand, it is straightforward to show that

$$\sum_{i=1}^n \frac{R_i}{\pi(Q_i,\psi)} \int \{a(t,X_i) - \bar{a}(t,\beta,\psi)\} \lambda(t) e^{\beta^T X_i} I(T_i \geq t) dt = 0, \; \forall \beta, \forall \psi. \tag{3.13}$$

Consequently, φ_0^* suggests the following estimating equations for β,

$$\begin{aligned} 0 = & \sum_{i=1}^n \left[\frac{R_i}{\pi(Q_i,\hat{\psi}_n)} \int \{a(t,X_i) - \bar{a}(t,\beta,\hat{\psi}_n)\} dN_i(t) \right. \\ & \left. - \frac{\{R_i - \pi(Q_i,\hat{\psi}_n)\}}{\pi(Q_i,\hat{\psi}_n)} L(Q_i)\right], \; \forall a(t,X), \forall L(Q). \end{aligned} \tag{3.14}$$

Alternatively, one can solve the following two sets of estimating equations jointly for β and $d\Lambda(t)$,

$$\begin{aligned} 0 &= \sum_{i=1}^n \left[\frac{R_i}{\pi(Q_i,\hat{\psi}_n)} \int a(t,X_i) dM_i(t,\beta) - \frac{\{R_i - \pi(Q_i,\hat{\psi}_n)\}}{\pi(Q_i,\hat{\psi}_n)} L(Q_i)\right], \\ 0 &= \sum_{i=1}^n \frac{R_i}{\pi(Q_i,\hat{\psi}_n)} dM_i(t,\beta), \end{aligned}$$

where $dM(t,\beta) = dN(t) - \lambda(t) e^{\beta^T X} I(T \geq t) dt$, so that $dM(t) = dM(t,\beta_0)$.

In addition to yielding (3.14), this also motivates an estimator for $d\Lambda(t)$, i.e.,

$$d\hat{\Lambda}(t) = \frac{\sum_{j=1}^n \frac{R_j}{\pi(Q_j,\hat{\psi}_n)} dN_j(t)}{\sum_{j=1}^n \frac{R_j}{\pi(Q_j,\hat{\psi}_n)} e^{\hat{\beta}_n^T X_j} I(T_j \geq t)}. \tag{3.15}$$

By (3.13), (3.14) is identical to

$$0 = n^{-1}\sum_{i=1}^{n}\left[\frac{R_i}{\pi(Q_i,\hat\psi_n)}\int\{a(t,X_i)-\bar a(t,\beta,\hat\psi_n)\}dM_i(t,\beta) - \frac{\{R_i-\pi(Q_i,\hat\psi_n)\}}{\pi(Q_i,\hat\psi_n)}L(Q_i)\right]. \quad (3.16)$$

When evaluated at β_0, a typical summand of (3.16) is asymptotically equivalent to φ^*_{0i} as expected. By the LIE and martingale properties, $E(\varphi^*_0)=0$. Therefore, (3.14) is an asymptotically unbiased estimating equation for β. Consequently, under certain regularity conditions, the resulting estimator is consistent.

Denote

$$\bar X(t,\beta,\psi)=\frac{\sum_{j=1}^{n}\frac{R_j}{\pi(Q_j,\psi)}X_je^{\beta^TX_j}I(T_j\geq t)}{\sum_{j=1}^{n}\frac{R_j}{\pi(Q_j,\psi)}e^{\beta^TX_j}I(T_j\geq t)}.$$

Then it is straightforward to show that

$$\begin{aligned}\frac{\partial\bar a(t,\beta,\psi)}{\partial\beta^T} &= \frac{\sum_{j=1}^{n}\frac{R_j}{\pi(Q_j,\psi)}\{a(t,X_j)-\bar a(t,\beta,\psi)\}\{X_j-\bar X(t,\beta,\psi)\}^Te^{\beta^TX_j}I(T_j\geq t)}{\sum_{j=1}^{n}\frac{R_j}{\pi(Q_j,\psi)}e^{\beta^TX_j}I(T_j\geq t)}\\ &\equiv C_n(a,X;t,\beta,\psi).\end{aligned}$$

Expanding (3.14) about β_0, while keeping $\hat\psi_n$ fixed, yields

$$\begin{aligned}&n^{1/2}(\hat\beta_n-\beta_0)\\ &= \left\{n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi(Q_i,\hat\psi_n)}\int C_n(a,X;t,\beta_n^*,\hat\psi_n)dN_i(t)\right\}^{-1}\\ &\quad\times n^{-1/2}\sum_{i=1}^{n}\left[\frac{R_i}{\pi(Q_i,\hat\psi_n)}\int\{a(t,X_i)-\bar a(t,\beta_0,\hat\psi_n)\}dN_i(t) - \frac{\{R_i-\pi(Q_i,\hat\psi_n)\}}{\pi(Q_i,\hat\psi_n)}L(Q_i)\right],\end{aligned} \quad (3.17)$$

where β_n^* lies between $\hat\beta_n$ and β_0.

Since $\hat\beta_n\xrightarrow{p}\beta_0$, $\beta_n^*\xrightarrow{p}\beta_0$. By the WLLN, $\bar a(t,\beta_n^*,\hat\psi_n)\xrightarrow{p}\mu_a(t)$, $\bar X(t,\beta_n^*,\hat\psi_n)\xrightarrow{p}\mu_X(t)$. In addition, by the LIE, $C_n(a,X;t,\beta_n^*,\hat\psi_n)\xrightarrow{p}\sigma(a,X;t)$, where

$$\sigma(a,X;t)=\frac{E[\{a(t,X)-\mu_a(t)\}\{X-\mu_X(t)\}^Te^{\beta_0^TX}I(T\geq t)]}{E\{e^{\beta_0^TX}I(T\geq t)\}}.$$

Therefore, by the LIE and martingale properties, the leading $q\times q$ matrix inside the bracket on the RHS of (3.17) converges in probability to

$$E\left[\frac{R}{\pi(Q,\psi_0)}\int\sigma(a,X;t)dN(t)\right]=V(a,X). \quad (3.18)$$

This suggests that we can estimate $V(a, X)$ by

$$\hat{V}(a, X) = n^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi(Q_i, \hat{\psi}_n)} \int C_n(a, X; t, \hat{\beta}_n, \hat{\psi}_n) dN_i(t).$$

On the other hand, by martingale properties and the LIE, it can be shown that

$$V(a, X) = E\left[\frac{R}{\pi(Q, \psi_0)} \int \{a(t, X) - \mu_a(t)\}\{X - \mu_X(t)\}^T dN(t)\right].$$

Therefore, an alternative estimator for $V(a, X)$ is provided by

$$\tilde{V}(a, X) = n^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi(Q_i, \hat{\psi}_n)} \int \{a(t, X_i) - \bar{a}(t, \hat{\beta}_n)\}\{X_i - \bar{X}(t, \hat{\beta}_n)\}^T dN_i(t).$$

Note that

$$\begin{aligned} \frac{\partial \bar{a}(t, \beta, \psi)}{\partial \psi^T} &= \frac{\sum_{j=1}^{n} \{a(t, X_j) - \bar{a}(t, \beta, \psi)\} \frac{-R_j \pi_\psi^T(Q_j, \psi)}{\pi^2(Q_j, \psi)} e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^{n} \frac{R_j}{\pi(Q_j, \psi)} e^{\beta^T X_j} I(T_j \geq t)} \\ &\equiv \xi_n(t, \beta, \psi). \end{aligned}$$

Therefore, by (3.13) and by expanding about ψ_0, the $q \times 1$ vector on the RHS of (3.17) is equal to

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^{n} \left[\frac{R_i}{\pi(Q_i, \hat{\psi}_n)} \int \{a(t, X_i) - \bar{a}(t, \beta_0, \hat{\psi}_n)\} dM_i(t) \right. \\ & \qquad \left. - \frac{\{R_i - \pi(Q_i, \hat{\psi}_n)\}}{\pi(Q_i, \hat{\psi}_n)} L(Q_i)\right] \\ = \; & n^{-1/2} \sum_{i=1}^{n} \left[\frac{R_i}{\pi(Q_i, \psi_0)} \int \{a(t, X_i) - \bar{a}(t, \beta_0, \psi_0)\} dM_i(t) \right. \\ & \qquad \left. - \frac{\{R_i - \pi(Q_i, \psi_0)\}}{\pi(Q_i, \psi_0)} L(Q_i)\right] \\ & + \left\{ n^{-1} \sum_{i=1}^{n} \left[-\frac{R_i}{\pi(Q_i, \psi_n^*)} \int \xi_n(t, \beta_0, \psi_n^*) dM_i(t) \right. \right. \\ & \qquad - \int \{a(t, X_i) - \bar{a}(t, \beta_0, \psi_n^*)\} dM_i(t) \frac{R_i \pi_\psi^T(Q_i, \psi_n^*)}{\pi^2(Q_i, \psi_n^*)} \\ & \qquad \left. \left. + \; L(Q_i) \frac{R_i \pi_\psi^T(Q_i, \psi_n^*)}{\pi^2(Q_i, \psi_n^*)} \right] \right\} n^{1/2}(\hat{\psi}_n - \psi_0), \end{aligned} \tag{3.19}$$

where ψ_n^* lies between $\hat{\psi}_n$ and ψ_0.

Similar to Tsiatis (1981), it can be shown that

$$n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi(Q_i, \psi_0)} \int \{\bar{a}(t, \beta_0, \psi_0) - \mu_a(t)\} dM_i(t) = o_p(1).$$

Therefore,

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi(Q_i,\psi_0)}\int\{a(t,X_i)-\bar{a}(t,\beta_0,\psi_0)\}dM_i(t) \\
= \; & n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi(Q_i,\psi_0)}\int\{a(t,X_i)-\mu_a(t)\}dM_i(t)+o_p(1).
\end{aligned}
$$

Consequently, a typical summand of the first term on the RHS of (3.19) is asymptotically equivalent to φ_{0i}^*.

Let us now consider the three matrices inside the bracket of the second term on the RHS of (3.19). Since $\hat{\psi}_n \xrightarrow{p} \psi_0$, $\psi_n^* \xrightarrow{p} \psi_0$. By the WLLN and the LIE, it is straightforward to show that $\bar{a}(t,\beta_0,\psi_n^*) \xrightarrow{p} \mu_a(t)$, so that $\xi_n(t,\beta_0,\psi_n^*) \xrightarrow{p} \xi(t)$, where

$$
\xi(t)=\frac{E\left[\{a(t,X)-\mu_a(t)\}\frac{-\pi_\psi^T(Q,\psi_0)}{\pi(Q,\psi_0)}e^{\beta_0^TX}I(T\geq t)\right]}{E\{e^{\beta_0^TX}I(T\geq t)\}}.
$$

Therefore, by the LIE and martingale properties, the first matrix converges in probability to zero.

By the LIE and (3.12), the second matrix converges in probability to

$$
\begin{aligned}
& -E\left[\int\{a(t,X)-\mu_a(t)\}\,dM(t)\frac{R\pi_\psi^T(Q,\psi_0)}{\pi^2(Q,\psi_0)}\right] \\
= \; & -E\left[\int\{a(t,X)-\mu_a(t)\}\,dM(t)\frac{\pi_\psi^T(Q,\psi_0)}{\pi(Q,\psi_0)}\right] \\
= \; & -E\left[\frac{R}{\pi(Q,\psi_0)}\int\{a(t,X)-\mu_a(t)\}\,dM(t)S_\psi^T\right].
\end{aligned}
$$

Similarly, the third matrix converges in probability to

$$
-E\left[-\frac{\{R-\pi(Q,\psi_0)\}}{\pi(Q,\psi_0)}L(Q)S_\psi^T\right].
$$

Therefore, the matrix as sum of three matrices inside the bracket of the second term on the RHS of (3.19) converges in probability to $-E(\varphi_0^*S_\psi^T)$.

On the other hand, since $\hat{\psi}_n$ is the MLE of ψ, we have that

$$
n^{1/2}(\hat{\psi}_n-\psi_0)=n^{-1/2}\sum_{i=1}^{n}I_\psi^{-1}S_{\psi i}+o_p(1), \tag{3.20}
$$

where

$$
I_\psi=E(S_\psi S_\psi^T)=E\left[\frac{\pi_\psi(Q)\pi_\psi^T(Q)}{\pi(Q)\{1-\pi(Q)\}}\right].
$$

Consequently, (3.19) is equal to

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\{\varphi_{0i}^{*} - E(\varphi_0^{*}S_\psi^T)I_\psi^{-1}S_{\psi i}\} + o_p(1) \\
= \; & n^{-1/2}\sum_{i=1}^{n}\{\varphi_{0i}^{*} - \Pi[\varphi_{0i}^{*}|\Lambda_\psi]\} + o_p(1).
\end{aligned}
$$

Substituting into (3.17), we have that

$$
n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2}\sum_{i=1}^{n}\{\varphi_{0i} - \Pi[\varphi_{0i}|\Lambda_\psi]\} + o_p(1).
$$

Therefore, the influence function for $\hat{\beta}_n$ is given by $\varphi = \varphi_0 - \Pi[\varphi_0|\Lambda_\psi]$.

By the CLT, $n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Sigma)$, where $\Sigma = E(\varphi\varphi^T)$.

By the Pythagorean theorem,

$$
\Sigma = E(\varphi_0\varphi_0^T) - E(\varphi_0 S_\psi^T)I_\psi^{-1}\{E(\varphi_0 S_\psi^T)\}^T.
$$

Since $\varphi_0 = V^{-1}(a, X)\varphi_0^{*}$, we have that

$$
\Sigma = V^{-1}(a, X)[E(\varphi_0^{*}\varphi_0^{*T}) - E(\varphi_0^{*}S_\psi^T)I_\psi^{-1}\{E(\varphi_0^{*}S_\psi^T)\}^T]V^{-T}(a, X).
$$

To construct an estimator for the asymptotic variance, we might first estimate the ith influence function by plugging in all parameter estimates. For example, we might consider

$$
\begin{aligned}
\hat{\varphi}_{0i}^{*} \;=\; & \frac{R_i}{\pi(Q_i, \hat{\psi}_n)}\int\{a(t, X_i) - \bar{a}(t, \hat{\beta}_n, \hat{\psi}_n)\}\{dN_i(t) - e^{\hat{\beta}_n^T X_i}I(T_i \geq t)d\hat{\Lambda}(t)\} \\
& - \frac{\{R_i - \pi(Q_i, \hat{\psi}_n)\}}{\pi(Q_i, \hat{\psi}_n)}L(Q_i),
\end{aligned}
$$

where $d\hat{\Lambda}(t)$ is given by (3.15).

Then substitute the estimate for the ith influence function into the asymptotic variance. For example, we might consider

$$
\hat{E}(\varphi_0^{*}\varphi_0^{*T}) = n^{-1}\sum_{i=1}^{n}\hat{\varphi}_{0i}^{*}\hat{\varphi}_{0i}^{*T},
$$

$$
\hat{E}(\varphi_0^{*}S_\psi^T) = n^{-1}\sum_{i=1}^{n}\hat{\varphi}_{0i}^{*}\frac{\pi_\psi^T(Q_i, \hat{\psi}_n)}{\pi(Q_i, \hat{\psi}_n)}.
$$

In addition, let $\hat{I}_\psi = \hat{A}_\psi^{-1}$, where

$$
\hat{A}_\psi = n^{-1}\sum_{i=1}^{n}\frac{\pi_\psi(Q_i, \hat{\psi}_n)\pi_\psi^T(Q_i, \hat{\psi}_n)}{\pi(Q_i, \hat{\psi}_n)\{1 - \pi(Q_i, \hat{\psi}_n)\}}.
$$

Therefore, when the π model is correctly specified, we have

$$
\hat{\text{Var}}(\hat{\beta}_n) = n^{-1}\hat{V}^{-1}(a, X)[\hat{E}(\varphi_0^{*}\varphi_0^{*T}) - \hat{E}(\varphi_0^{*}S_\psi^T)\hat{I}_\psi^{-1}\{\hat{E}(\varphi_0^{*}S_\psi^T)\}^T]\hat{V}^{-T}(a, X).
$$

3.6 Doubly Robust Semiparametric Estimators

In this section, we will fix some arbitrary full data influence function and search for the element of Λ_2 that gives rise to the most efficient observed data influence function associated with the full data influence function. It is shown by Robins, Rotnitzky, and Scharfstein (1999) that estimators with such influence functions are doubly robust. In a missing data model, an estimator is said to be doubly robust if it remains consistent when either the model for the missing data mechanism or the model for the distribution of the complete data is correctly specified.

Let $\varphi^F(Z)$ be the arbitrary full data influence function to be fixed throughout this section. When the missingness mechanism is known, i.e., ψ_0 fixed, the class of observed data influence functions associated with φ^F is given by

$$_{\varphi^F}\Lambda_{0*}^{\perp} \quad = \quad \left\{ \frac{R}{\pi(Q)}\varphi^F(Z) + L_2(O) : \forall L_2 \in \Lambda_2 \right\}.$$

When the missingness mechanism is unknown, the space of observed data influence functions associated with φ^F is given by

$$_{\varphi^F}\Lambda_{*}^{\perp} = _{\varphi^F}\Lambda_{0*}^{\perp} - \Pi\left[_{\varphi^F}\Lambda_{0*}^{\perp} \middle| \Lambda_\psi \right].$$

Define

$$\varphi(O) = \frac{R}{\pi(Q)}\varphi^F(Z) - \Pi\left[\frac{R}{\pi(Q)}\varphi^F(Z) \middle| \Lambda_2 \right].$$

Then, by the projection theorem, we have that

$$\varphi = \text{argmin}_{h \in _{\varphi^F}\Lambda_{0*}^{\perp}} \|h\|,$$

where $\|h\|^2 = E\{h^T(O)h(O)\}$. Recall $\Lambda_\psi \subset \Lambda_2$, hence $\varphi \in _{\varphi^F} \Lambda_*^{\perp} \subset _{\varphi^F} \Lambda_{0*}^{\perp}$, so that

$$\varphi = \text{argmin}_{h \in _{\varphi^F}\Lambda_{*}^{\perp}} \|h\|.$$

Therefore, φ is the most efficient observed data influence function associated with the full data influence function φ^F in the sense that it has the smallest variance.

By (3.11), we have that

$$\Pi\left[\frac{R}{\pi(Q)}\varphi^F(Z) \middle| \Lambda_2 \right] = -\frac{\{R - \pi(Q)\}}{\pi(Q)} L^*(Q), \tag{3.21}$$

for some $q \times 1$ function of Q, $L^*(Q)$ satisfying $E\{L^{*T}(Q)L^*(Q)\} < \infty$.

By the projection theorem,

$$\begin{aligned} 0 \quad = \quad & E\left\{ \left[\frac{R}{\pi(Q)}\varphi^F(Z) + \frac{\{R - \pi(Q)\}}{\pi(Q)} L^*(Q) \right]^T \right. \\ & \left. \times \left[-\frac{\{R - \pi(Q)\}}{\pi(Q)} L(Q) \right] \right\}, \ \forall L(Q). \end{aligned}$$

By the LIE, this is equivalent to

$$
\begin{aligned}
0 &= E\left[\frac{R\{R-\pi(Q)\}}{\pi^2(Q)}\varphi^{FT}(Z)L(Q)\right] \\
&\quad +E\left[\frac{\{R-\pi(Q)\}^2}{\pi^2(Q)}L^{*T}(Q)L(Q)\right] \\
&= E\left[\frac{\{1-\pi(Q)\}}{\pi(Q)}\varphi^{FT}(Z)L(Q)\right] \\
&\quad +E\left[\frac{\{1-\pi(Q)\}}{\pi(Q)}L^{*T}(Q)L(Q)\right] \\
&= E\left[\frac{\{1-\pi(Q)\}}{\pi(Q)}\{\varphi^{F}(Z)+L^{*}(Q)\}^T L(Q)\right] \\
&= E\left[\frac{\{1-\pi(Q)\}}{\pi(Q)}[E\{\varphi^{F}(Z)|Q\}+L^{*}(Q)]^T L(Q)\right],\ \forall L(Q). \qquad (3.22)
\end{aligned}
$$

Let $L(Q) = E\{\varphi^F(Z)|Q\} + L^*(Q)$, then (3.22) implies that

$$\frac{\{1-\pi(Q)\}}{\pi(Q)}[E\{\varphi^F(Z)|Q\} + L^*(Q)] = 0.$$

Equivalently, we have that

$$-\frac{\{1-\pi(Q)\}}{\pi(Q)}L^*(Q) = \frac{\{1-\pi(Q)\}}{\pi(Q)}E\{\varphi^F(Z)|Q\}.$$

If $\{1-\pi(Q)\} > 0$, then multiplying both sides by $\frac{\{R-\pi(Q)\}}{\{1-\pi(Q)\}}$ yields

$$-\frac{\{R-\pi(Q)\}}{\pi(Q)}L^*(Q) = \frac{\{R-\pi(Q)\}}{\pi(Q)}E\{\varphi^F(Z)|Q\}. \qquad (3.23)$$

If $\{1-\pi(Q)\} = 0$, then $R = \pi(Q) = 1$, which would trivially imply (3.23). Therefore, (3.23) is satisfied in all cases.
Substituting (3.23) into (3.21), we have that

$$\Pi\left[\frac{R}{\pi(Q)}\varphi^F(Z)\middle|\Lambda_2\right] = \frac{\{R-\pi(Q)\}}{\pi(Q)}E\{\varphi^F(Z)|Q\}.$$

Consequently,

$$\varphi(O) = \frac{R}{\pi(Q)}\varphi^F(Z) - \frac{\{R-\pi(Q)\}}{\pi(Q)}E\{\varphi^F(Z)|Q\}. \qquad (3.24)$$

On the other hand, by (3.7), we have that

$$\varphi^F(Z) = V^{-1}(a,X)\int\{a(t,X) - \mu_a(t)\}dM(t), \qquad (3.25)$$

for some $q\times 1$ function $a(t,X)$.

Denote $N^*(t) = I(T \le t)$, then $N(t) = I(\Delta = 2)N^*(t)$. Therefore,

$$\begin{aligned} & E\left[\int \{a(t,X) - \mu_a(t)\}dM(t)\bigg| Q\right] \\ = & \int \{a(t,X) - \mu_a(t)\}E\{dM(t)|Q\} \\ = & \int \{a(t,X) - \mu_a(t)\}\{\varrho(Q)dN^*(t) - \lambda(t)e^{\beta_0^T X}I(T \ge t)dt\}, \end{aligned} \tag{3.26}$$

where

$$\varrho(Q) = P(\Delta = 2|Q).$$

Write $\varrho(W) = P(\Delta = 2|W, \Delta > 0)$, then $\varrho(Q) = \varrho(W)I(\Delta > 0)$. Suppose that we posit a parametric model for ϱ, say $\varrho(W) = \varrho(W, \gamma)$, then $\varrho(Q, \gamma) = \varrho(W, \gamma)I(\Delta > 0)$. Let $\hat{\gamma}_n$ be the MLE of γ, then $\hat{\gamma}_n \xrightarrow{p} \gamma^*$ for some γ^*. Similarly, assume that $\hat{\psi}_n \xrightarrow{p} \psi^*$ for some ψ^*.

Note that the π model describes the missingness mechanism and the ϱ model describes the distribution of the complete data. To gain double robustness, we further assume that either the π model or the ϱ model is correctly specified. Therefore, either $\psi^* = \psi_0$ or $\gamma^* = \gamma_0$.

To simplify notation, write

$$\begin{aligned} \Phi(R, Z; \psi, \gamma) &= \frac{R}{\pi(Q,\psi)}I(\Delta = 2) - \frac{\{R - \pi(Q,\psi)\}}{\pi(Q,\psi)}\varrho(Q,\gamma), \\ \Omega(R, Z; \psi, \gamma) &= \frac{\{R - \pi(Q,\psi)\}}{\pi(Q,\psi)}\{I(\Delta = 2) - \varrho(Q,\gamma)\}. \end{aligned}$$

Then

$$\Phi(R, Z; \psi, \gamma) = I(\Delta = 2) + \Omega(R, Z; \psi, \gamma). \tag{3.27}$$

Notice, however, for fixed (ψ, γ), $\Phi(R, Z; \psi, \gamma)$ is a function of the observed data, while $\Omega(R, Z; \psi, \gamma)$ involves not only the observed data, but also the cause of failure indicator, $I(\Delta = 2)$, which might be missing for some individuals. Consequently, $\Phi(R, Z; \psi, \gamma)$ is calculable on all subjects, while $\Omega(R, Z; \psi, \gamma)$ is not.

Now substituting (3.25) and (3.26) into (3.24), we have that

$$\varphi(O) = V^{-1}(a, X)\varphi^*(O),$$

where

$$\begin{aligned} \varphi^*(O) &= \frac{R}{\pi(Q)}\int \{a(t,X) - \mu_a(t)\}dM(t) \\ & \quad -\frac{\{R - \pi(Q)\}}{\pi(Q)}\int \{a(t,X) - \mu_a(t)\}\{\varrho(Q)dN^*(t) - \lambda(t)e^{\beta_0^T X}I(T \ge t)dt\} \\ &= \Phi(R, Z)\int \{a(t,X) - \mu_a(t)\}dN^*(t) \\ & \quad -\int \{a(t,X) - \mu_a(t)\}\lambda(t)e^{\beta_0^T X}I(T \ge t)dt. \end{aligned}$$

Denote

$$\bar{a}(t,\beta) = \frac{\sum_{j=1}^n a(t,X_j)e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^n e^{\beta^T X_j} I(T_j \geq t)}.$$

By the WLLN, $\bar{a}(t,\beta_0) \xrightarrow{p} \mu_a(t)$.

On the other hand, it is straightforward to show that

$$\sum_{i=1}^n \int \{a(t,X_i) - \bar{a}(t,\beta)\}\lambda(t)e^{\beta^T X_i} I(T_i \geq t)dt = 0, \ \forall\beta. \tag{3.28}$$

Consequently, φ^* suggests the following estimating equations for β,

$$0 \ = \ \sum_{i=1}^n \Phi(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int \{a(t,X_i) - \bar{a}(t,\beta)\} dN_i^*(t), \tag{3.29}$$

where $a(t,X)$ is some arbitrary $q \times 1$ function of t and X.

Alternatively, one can solve the following two sets of estimating equations jointly for β and $d\Lambda(t)$,

$$\begin{aligned} 0 \ = \ \sum_{i=1}^n \Bigg[& \frac{R_i}{\pi(Q_i,\hat{\psi}_n)} \int a(t,X_i) dM_i(t,\beta) \\ & - \frac{\{R_i - \pi(Q_i,\hat{\psi}_n)\}}{\pi(Q_i,\hat{\psi}_n)} \int a(t,X_i)\{\varrho(Q_i,\hat{\gamma}_n)dN_i^*(t) \\ & \qquad - \lambda(t)e^{\beta^T X_i} I(T_i \geq t)dt\} \Bigg], \end{aligned} \tag{3.30}$$

$$\begin{aligned} 0 \ = \ \sum_{i=1}^n \Bigg[& \frac{R_i}{\pi(Q_i,\hat{\psi}_n)} dM_i(t,\beta) \\ & - \frac{\{R_i - \pi(Q_i,\hat{\psi}_n)\}}{\pi(Q_i,\hat{\psi}_n)} \{\varrho(Q_i,\hat{\gamma}_n)dN_i^*(t) \\ & \qquad - \lambda(t)e^{\beta^T X_i} I(T_i \geq t)dt\} \Bigg]. \end{aligned} \tag{3.31}$$

In addition to yielding (3.29), (3.30) and (3.31) also motivates an estimator for $d\Lambda(t)$, i.e.,

$$d\hat{\Lambda}(t) = \frac{\sum_{j=1}^n \Phi(R_j, Z_j; \hat{\psi}_n, \hat{\gamma}_n) dN_j^*(t)}{\sum_{j=1}^n e^{\hat{\beta}_n^T X_j} I(T_j \geq t)}. \tag{3.32}$$

By (3.27) and (3.28), (3.29) is identical to

$$\begin{aligned} 0 \ = \ n^{-1} \sum_{i=1}^n \Bigg[& \int \{a(t,X_i) - \bar{a}(t,\beta)\} dM_i(t,\beta) \\ & + \Omega(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int \{a(t,X_i) - \bar{a}(t,\beta)\} dN_i^*(t) \Bigg]. \end{aligned} \tag{3.33}$$

When evaluated at β_0, a typical summand of (3.33) is asymptotically equivalent to $\varphi_i^* = T_{1i} + T_{2i}$, where

$$\begin{aligned} T_1 &= \int \{a(t,X) - \mu_a(t)\} dM(t), \\ T_2 &= \Omega(R, Z; \psi^*, \gamma^*) \int \{a(t,X) - \mu_a(t)\} dN^*(t). \end{aligned}$$

By martingale properties, $E(T_1) = 0$. To show $E(T_2) = 0$ when either of the two parametric models is correctly specified, we will use the double robustness argument as we will describe shortly. Similar arguments will be frequently used later on. We will generically refer to this class of arguments as the DR argument.

If the π model is correctly specified, $\psi^* = \psi_0$, $P(R = 1|Z) = \pi(Q, \psi_0)$, hence

$$\begin{aligned} E\{\Omega(R, Z; \psi_0, \gamma^*)|Z\} &= \frac{\{P(R = 1|Z) - \pi(Q, \psi_0)\}}{\pi(Q, \psi_0)} \{I(\Delta = 2) - \varrho(Q, \gamma^*)\} \\ &= 0. \end{aligned} \tag{3.34}$$

If the ϱ model is correctly specified, $\gamma^* = \gamma_0$, $P(\Delta = 2|Q) = \varrho(Q, \gamma_0)$, hence, by the MAR assumption, we have

$$\begin{aligned} E\{\Omega(R, Z; \psi^*, \gamma_0)|R, Q\} &= \frac{\{R - \pi(Q, \psi^*)\}}{\pi(Q, \psi^*)} \{P(\Delta = 2|R, Q) - \varrho(Q, \gamma_0)\} \\ &= \frac{\{R - \pi(Q, \psi^*)\}}{\pi(Q, \psi^*)} \{P(\Delta = 2|Q) - \varrho(Q, \gamma_0)\} \\ &= 0. \end{aligned} \tag{3.35}$$

Now if the π model is correctly specified, then by the LIE and (3.34),

$$\begin{aligned} E(T_2) &= E\{E(T_2|Z)\} \\ &= E\left[E\{\Omega(R, Z; \psi_0, \gamma^*)|Z\} \int \{a(t,X) - \mu_a(t)\} dN^*(t)\right] \\ &= 0. \end{aligned}$$

Similarly, if the ϱ model is correctly specified, then by the LIE and (3.35),

$$\begin{aligned} E(T_2) &= E\{E(T_2|R, Q)\} \\ &= E\left[E\{\Omega(R, Z; \psi^*, \gamma_0)|R, Q\} \int \{a(t,X) - \mu_a(t)\} dN^*(t)\right] \\ &= 0. \end{aligned}$$

In summary, when either of the two parametric models is correctly specified, $E(T_2) = 0$, hence $E(\varphi^*) = 0$, so that (3.29) is an asymptotically unbiased estimating equation for β. Consequently, under certain regularity conditions, the resulting estimator is consistent.

Denote

$$\bar{X}(t,\beta) = \frac{\sum_{j=1}^{n} X_j e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^{n} e^{\beta^T X_j} I(T_j \geq t)}.$$

Then

$$\begin{aligned}\frac{\partial \bar{a}(t,\beta)}{\partial \beta^T} &= \frac{\sum_{j=1}^{n}\{a(t,X_j) - \bar{a}(t,\beta)\}\{X_j - \bar{X}(t,\beta)\}^T e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^{n} e^{\beta^T X_j} I(T_j \geq t)} \\ &\equiv D_n(a,X;t,\beta).\end{aligned}$$

Expanding (3.29) about β_0 yields

$$\begin{aligned}& n^{1/2}(\hat{\beta}_n - \beta_0) \\ =\ & \left\{\Phi(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int D_n(a,X;t,\beta_n^*)dN_i^*(t)\right\}^{-1} \\ & \times n^{-1/2}\sum_{i=1}^{n}\Phi(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int\{a(t,X_i) - \bar{a}(t,\beta_0)\}dN_i^*(t), \end{aligned} \quad (3.36)$$

where β_n^* lies between $\hat{\beta}_n$ and β_0.

Since $\hat{\beta}_n \xrightarrow{p} \beta_0$, $\beta_n^* \xrightarrow{p} \beta_0$. By the WLLN, $\bar{a}(t,\beta_n^*) \xrightarrow{p} \mu_a(t)$, $\bar{X}(t,\beta_n^*) \xrightarrow{p} \mu_X(t)$, so that $D_n(a,X;t,\beta_n^*) \xrightarrow{p} \sigma(a,X;t)$.

Now suppose that

$$n^{1/2}(\hat{\psi}_n - \psi^*) = O_p(1), \quad n^{1/2}(\hat{\gamma}_n - \gamma^*) = O_p(1). \quad (3.37)$$

By the DR argument, the leading matrix inside the bracket on the RHS of (3.36) converges in probability to

$$E\left[\Phi(R,Z;\psi^*,\gamma^*)\int \sigma(a,X;t)dN^*(t)\right] = E\left[\int \sigma(a,X;t)dN(t)\right] = V(a,X).$$

This suggests the following doubly robust estimator for $V(a,X)$,

$$\hat{V}(a,X) = n^{-1}\sum_{i=1}^{n}\Phi(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int D_n(a,X;t,\hat{\beta}_n)dN_i^*(t).$$

On the other hand, by the DR argument,

$$V(a,X) = E\left[\Phi(R,Z;\psi^*,\gamma^*)\int\{a(t,X) - \mu_a(t)\}\{X - \mu_X(t)\}^T dN^*(t)\right].$$

Therefore, an alternative estimator for $V(a,X)$ is provided by

$$\tilde{V}(a,X) = n^{-1}\sum_{i=1}^{n}\Phi(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int\{a(t,X_i) - \bar{a}(t,\hat{\beta}_n)\}\{X_i - \bar{X}(t,\hat{\beta}_n)\}^T dN_i^*(t).$$

By (3.27), (3.28), and by expanding about (ψ^*, γ^*), the $q \times 1$ vector on the RHS of (3.36) is equal to

$$
\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} \left[\int \{a(t, X_i) - \bar{a}(t, \beta_0)\} dM_i(t) \right. \\
& \qquad \left. + \Omega(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int \{a(t, X_i) - \bar{a}(t, \beta_0)\} dN_i^*(t) \right] \\
= \; & n^{-1/2} \sum_{i=1}^{n} \left[\int \{a(t, X_i) - \bar{a}(t, \beta_0)\} dM_i(t) \right. \\
& \qquad \left. + \Omega(R_i, Z_i; \psi^*, \gamma^*) \int \{a(t, X_i) - \bar{a}(t, \beta_0)\} dN_i^*(t) \right] \\
& - \left[n^{-1} \sum_{i=1}^{n} \int \{a(t, X_i) - \bar{a}(t, \beta_0)\} dN_i^*(t) \right. \\
& \qquad \left. \times \{I(\Delta_i = 2) - \varrho(Q_i, \gamma_n^*)\} \frac{R_i \pi_\psi^T(Q_i, \psi_n^*)}{\pi^2(Q_i, \psi_n^*)} \right] n^{1/2}(\hat{\psi}_n - \psi^*) \\
& - \left[n^{-1} \sum_{i=1}^{n} \int \{a(t, X_i) - \bar{a}(t, \beta_0)\} dN_i^*(t) \right. \\
& \qquad \left. \times \frac{\{R_i - \pi(Q_i, \psi_n^*)\}}{\pi(Q_i, \psi_n^*)} \varrho_\gamma^T(Q_i, \gamma_n^*) \right] n^{1/2}(\hat{\gamma}_n - \gamma^*), \qquad (3.38)
\end{aligned}
$$

where ψ_n^* lies between $\hat{\psi}_n$ and ψ_0, and γ_n^* lies between $\hat{\gamma}_n$ and γ_0.

By martingale properties and the DR argument, when either of the two parametric models is correctly specified, a typical summand of the first term on the RHS of (3.38) is asymptotically equivalent to φ_i^*.

The leading matrix inside the bracket of the second term on the RHS of (3.38) converges in probability to

$$P_\psi = E\left[\int \{a(t, X) - \mu_a(t)\} dN^*(t) \{I(\Delta = 2) - \varrho(Q, \gamma^*)\} \frac{R \pi_\psi^T(Q, \psi^*)}{\pi^2(Q, \psi^*)} \right].$$

The leading matrix inside the bracket of the third term on the RHS of (3.38) converges in probability to

$$P_\gamma = E\left[\int \{a(t, X) - \mu_a(t)\} dN^*(t) \frac{\{R - \pi(Q, \psi^*)\}}{\pi(Q, \psi^*)} \varrho_\gamma^T(Q, \gamma^*) \right].$$

Note that if the π model is correctly specified, then $P_\gamma = 0$, hence, by (3.37), the third term on the RHS of (3.38) is negligible. Similarly, if the ϱ model is correctly specified, then $P_\psi = 0$, hence, by (3.37), the second term on the RHS of (3.38) is negligible.

Therefore, in the ideal case when both parametric models are correctly specified, a typical summand of the second vector on the RHS of (3.36) is asymptotically equivalent to

φ_i^*. Consequently, by (3.36),

$$n^{1/2}(\hat\beta_n-\beta_0)=n^{-1/2}\sum_{i=1}^n V^{-1}(a,X)\varphi_i^*+o_p(1).$$

Therefore, the influence function for $\hat\beta_n$ is given by $\varphi=V^{-1}(a,X)\varphi^*$.

By the CLT, $n^{1/2}(\hat\beta_n-\beta_0)\xrightarrow{d}N(0,\Sigma)$, where $\Sigma=E(\varphi\varphi^T)$.

By the LIE, when the π model is correctly specified, $E(T_1T_2^T)=0$. Therefore,

$$\Sigma=V^{-1}(a,X)\{E(T_1T_1^T)+E(T_2T_2^T)\}V^{-T}(a,X).$$

By martingale properties, we have that

$$\begin{aligned}E(T_1T_1^T)&=E\left[\int\{a(t,X)-\mu_a(t)\}^{\otimes 2}\lambda(t)e^{\beta_0^TX}I(T\ge t)dt\right]\\&=\int E[\{a(t,X)-\mu_a(t)\}^{\otimes 2}e^{\beta_0^TX}I(T\ge t)]\lambda(t)dt\\&\equiv V(a,a).\end{aligned}$$

Similar to the estimation of $V(a,X)$, we have that

$$\hat V(a,a)=n^{-1}\sum_{i=1}^n\Phi(R_i,Z_i;\hat\psi_n,\hat\gamma_n)\int D_n(a,a;t,\hat\beta_n)dN_i^*(t),$$

where

$$D_n(a,a;t,\beta)=\frac{\sum_{j=1}^n\{a(t,X_j)-\bar a(t,\beta)\}^{\otimes 2}e^{\beta^TX_j}I(T_j\ge t)}{\sum_{j=1}^n e^{\beta^TX_j}I(T_j\ge t)}.$$

And an alternative estimator for $V(a,a)$ is provided by

$$\tilde V(a,a)=n^{-1}\sum_{i=1}^n\Phi(R_i,Z_i;\hat\psi_n,\hat\gamma_n)\int\{a(t,X_i)-\bar a(t,\hat\beta_n)\}^{\otimes 2}dN_i^*(t).$$

By the DR argument, it can be shown that

$$\begin{aligned}E(T_2T_2^T)&=E\left[\frac{\{R-\pi(Q,\psi^*)\}^2}{\pi^2(Q,\psi^*)}\{I(\Delta=2)-\varrho(Q,\gamma^*)\}^2\right.\\&\qquad\left.\times\int\{a(t,X)-\mu_a(t)\}^{\otimes 2}dN^*(t)\right]\\&=E\left\{\left[\frac{R\{1-\pi(Q,\psi^*)\}}{\pi^2(Q,\psi^*)}\{I(\Delta=2)-\varrho(Q,\gamma^*)\}^2\right.\right.\\&\qquad\left.-\frac{\{R-\pi(Q,\psi^*)\}}{\pi(Q,\psi^*)}\varrho(Q,\gamma^*)\{1-\varrho(Q,\gamma^*)\}\right]\\&\qquad\left.\times\int\{a(t,X)-\mu_a(t)\}^{\otimes 2}dN^*(t)\right\}.\end{aligned}\tag{3.39}$$

In fact, if the π model is correctly specified, $\psi^* = \psi_0$, hence

$$\begin{aligned}
& E\left[\frac{R\{1-\pi(Q,\psi^*)\}}{\pi^2(Q,\psi^*)}\{I(\Delta=2)-\varrho(Q,\gamma^*)\}^2\right. \\
& \qquad \left. -\frac{\{R-\pi(Q,\psi^*)\}}{\pi(Q,\psi^*)}\varrho(Q,\gamma^*)\{1-\varrho(Q,\gamma^*)\}\right| Z\Bigg] \\
= & \frac{\{1-\pi(Q,\psi^*)\}}{\pi(Q,\psi^*)}\{I(\Delta=2)-\varrho(Q,\gamma^*)\}^2 \\
= & E\left[\frac{\{R-\pi(Q,\psi^*)\}^2}{\pi^2(Q,\psi^*)}\{I(\Delta=2)-\varrho(Q,\gamma^*)\}^2\right| Z\Bigg].
\end{aligned}$$

Therefore, (3.39) follows from the LIE by conditioning on Z.

Similarly, if the ϱ model is correctly specified, $\gamma^* = \gamma_0$, hence, by the MAR assumption,

$$\begin{aligned}
& E\left[\frac{R\{1-\pi(Q,\psi^*)\}}{\pi^2(Q,\psi^*)}\{I(\Delta=2)-\varrho(Q,\gamma^*)\}^2\right. \\
& \qquad \left. -\frac{\{R-\pi(Q,\psi^*)\}}{\pi(Q,\psi^*)}\varrho(Q,\gamma^*)\{1-\varrho(Q,\gamma^*)\}\right| R,Q\Bigg] \\
= & \frac{R\{1-\pi(Q,\psi^*)\}}{\pi^2(Q,\psi^*)}\varrho(Q,\gamma^*)\{1-\varrho(Q,\gamma^*)\} \\
& -\frac{\{R-\pi(Q,\psi^*)\}}{\pi(Q,\psi^*)}\varrho(Q,\gamma^*)\{1-\varrho(Q,\gamma^*)\} \\
= & \frac{\{R-\pi(Q,\psi^*)\}^2}{\pi^2(Q,\psi^*)}\varrho(Q,\gamma^*)\{1-\varrho(Q,\gamma^*)\} \\
= & E\left[\frac{\{R-\pi(Q,\psi^*)\}^2}{\pi^2(Q,\psi^*)}\{I(\Delta=2)-\varrho(Q,\gamma^*)\}^2\right| R,Q\Bigg].
\end{aligned}$$

Therefore, (3.39) follows from the LIE by conditioning on (R, Q).

This suggests the following doubly robust estimator for $E(T_2T_2^T)$,

$$\begin{aligned}
\hat{E}(T_2T_2^T) = & n^{-1}\sum_{i=1}^{n}\left[\frac{R_i\{1-\pi(Q_i,\hat{\psi}_n)\}}{\pi^2(Q_i,\hat{\psi}_n)}\{I(\Delta_i=2)-\varrho(Q_i,\hat{\gamma}_n)\}^2\right. \\
& \qquad \left. -\frac{\{R_i-\pi(Q_i,\hat{\psi}_n)\}}{\pi(Q_i,\hat{\psi}_n)}\varrho(Q_i,\hat{\gamma}_n)\{1-\varrho(Q_i,\hat{\gamma}_n)\}\right] \\
& \qquad \times\int\{a(t,X_i)-\bar{a}(t,\hat{\beta}_n)\}^{\otimes 2}dN_i^*(t).
\end{aligned}$$

Therefore, in the ideal case when both parametric models are correctly specified, we have

$$\hat{\text{Var}}(\hat{\beta}_n) = n^{-1}\hat{V}^{-1}(a,X)\{\hat{V}(a,a)+\hat{E}(T_2T_2^T)\}\hat{V}^{-T}(a,X).$$

Now suppose that the π model is correctly specified, but the ϱ model might be misspecified, in which case, $\psi^* = \psi_0$, $\gamma^* \neq \gamma_0$. It has been shown previously that the third term on the RHS of (3.38) is negligible and that the leading matrix inside the bracket of the second term on the RHS of (3.38) converges in probability to P_ψ. By the LIE and (3.12), P_ψ reduces to

$$P_\psi = E\left[\int \{a(t,X) - \mu_a(t)\}dN^*(t)\{I(\Delta = 2) - \varrho(Q,\gamma^*)\}\frac{\pi_\psi^T(Q)}{\pi(Q)}\right] = E(T_2 S_\psi^T).$$

On the other hand, by the LIE, it is straightforward to verify that

$$E(T_1 S_\psi^T) = E\left[\int \{a(t,X) - \mu_a(t)\}dM(t)\frac{\{R - \pi(Q)\}\pi_\psi(Q)}{\pi(Q)\{1-\pi(Q)\}}\right] = 0.$$

Therefore, the leading matrix inside the bracket of the second term on the RHS of (3.38) converges in probability to $P_\psi = E(\psi^* S_\psi^T)$.

Applying these results as well as (3.20) to (3.38), a typical summand of the $q \times 1$ vector on the RHS of (3.36) is asymptotically equivalent to $\{\varphi_i^* - \Pi[\varphi_i^*|\Lambda_\psi]\}$. Consequently, by (3.36),

$$n^{1/2}(\hat\beta_n - \beta_0) = n^{-1/2}\sum_{i=1}^n V^{-1}(a,X)\left\{\varphi_i^* - \Pi[\varphi_i^*|\Lambda_\psi]\right\} + o_p(1).$$

Therefore, the influence function for $\hat\beta_n$ is given by $\varphi = V^{-1}(a,X)\{\varphi^* - \Pi[\varphi^*|\Lambda_\psi]\}$.

By the CLT, $n^{1/2}(\hat\beta_n - \beta_0) \xrightarrow{d} N(0,\Sigma)$, where $\Sigma = E(\varphi\varphi^T)$.

Note that $\Pi[\varphi^*|\Lambda_\psi] = P_\psi I_\psi^{-1} S_\psi$. Also recall $E(T_1 T_2^T) = 0$, $E(T_1 T_1^T) = V(a,a)$. Therefore, by the Pythagorean theorem,

$$\Sigma = V^{-1}(a,X)\{V(a,a) + E(T_2T_2^T) - P_\psi I_\psi^{-1} P_\psi^T\}V^{-T}(a,X).$$

By definition, a doubly robust estimator for P_ψ is given by

$$\hat P_\psi = n^{-1}\sum_{i=1}^n \int \{a(t,X_i) - \bar a(t,\hat\beta_n)\}dN_i^*(t)\{I(\Delta_i = 2) - \varrho(Q_i,\hat\gamma_n)\}\frac{R_i\pi_\psi^T(Q_i,\hat\psi_n)}{\pi^2(Q_i,\hat\psi_n)}.$$

Finally, let $\hat I_\psi = \hat A_\psi^{-1}$. Therefore, when the π model is correctly specified, we have

$$\hat{\text{Var}}(\hat\beta_n) = n^{-1}\hat V^{-1}(a,X)\{\hat V(a,a) + \hat E(T_2T_2^T) - \hat P_\psi \hat I_\psi^{-1}\hat P_\psi^T\}\hat V^{-T}(a,X).$$

Now suppose that the ϱ model is correctly specified, but the π model may be misspecified, in which case, $\gamma^* = \gamma_0$, $\psi^* \neq \psi_0$. It has been shown previously that the second term on the RHS of (3.38) is negligible and that the leading matrix inside the bracket of the third term on the RHS of (3.38) converges in probability to P_γ.

Note that the observed data likelihood for γ is

$$\prod_{i=1}^n \{\varrho(Q_i,\gamma)\}^{I(\Delta_i=2,R_i=1)}\{1-\varrho(Q_i,\gamma)\}^{I(\Delta_i=1,R_i=1)}.$$

Therefore, the log-likelihood for γ for a typical observation is

$$\ell(\gamma, O) = R[I(\Delta = 2) \log \varrho(Q, \gamma) + I(\Delta = 1) \log\{1 - \varrho(Q, \gamma)\}].$$

Consequently, the score vector for γ is

$$S_\gamma(O) = \frac{R\{I(\Delta = 2) - \varrho(Q)\}\varrho_\gamma(Q)}{\varrho(Q)\{1 - \varrho(Q)\}}.$$

By the MAR assumption, the associated Fisher information matrix is

$$I_\gamma = E(S_\gamma S_\gamma^T) = E\left[\frac{R\varrho_\gamma(Q)\varrho_\gamma^T(Q)}{\varrho(Q)\{1 - \varrho(Q)\}}\right].$$

It follows that

$$n^{1/2}(\hat{\gamma}_n - \gamma_0) = n^{-1/2} \sum_{i=1}^n I_\gamma^{-1} S_{\gamma i} + o_p(1).$$

Now a typical summand of the vector on the RHS of (3.36) is asymptotically equivalent to $\{\varphi_i^* - P_\gamma I_\gamma^{-1} S_{\gamma i}\}$. Consequently, by (3.36),

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} \sum_{i=1}^n V^{-1}(a, X)\{\varphi_i^* - P_\gamma I_\gamma^{-1} S_{\gamma i}\} + o_p(1).$$

Therefore, the influence function for $\hat{\beta}_n$ is given by $\varphi = V^{-1}(a, X)\{\varphi^* - P_\gamma I_\gamma^{-1} S_\gamma\}$.

By the CLT, $n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Sigma)$, where $\Sigma = E(\varphi\varphi^T)$.

As we will see shortly, $P_\gamma \neq E(\varphi^* S_\gamma^T)$, hence $P_\gamma I_\gamma^{-1} S_\gamma \neq \Pi[\varphi^* | \Lambda_\gamma]$. Consequently, it might be reasonable to suspect that φ can not be written in the generic form of a typical element of $\Lambda_{0*}^\perp$ or $\Lambda_*^\perp$. In fact, when the ϱ model is correctly specified, the full data nuisance tangent space and the observed data nuisance tangent space will change to account for the finite-dimensional nuisance parameter γ, because there is an intrinsic relationship between $\lambda(t|x)$, $\lambda_1(t|x)$, $p_{A|T,X,\Delta}$ and $\varrho(W)$ given by (C.1). Nonetheless, we can still consistently estimate the asymptotic variance matrix for $\hat{\beta}_n$ from the observed data. As we will also see shortly, $E(T_1 T_2^T) = E(T_2 T_1^T)$. Recall $E(T_1 T_1^T) = V(a, a)$. Therefore,

$$\begin{aligned}\Sigma &= V^{-1}(a, X)[V(a, a) + E(T_2 T_2^T) + 2E(T_1 T_2^T) + P_\gamma I_\gamma^{-1} P_\gamma^T \\ &\quad - E(\varphi^* S_\gamma^T) I_\gamma^{-1} P_\gamma^T - P_\gamma I_\gamma^{-1}\{E(\varphi^* S_\gamma^T)\}^T] V^{-T}(a, X).\end{aligned}$$

By definition, a doubly robust estimator for P_γ is given by

$$\hat{P}_\gamma = n^{-1} \sum_{i=1}^n \int \{a(t, X_i) - \bar{a}(t, \hat{\beta}_n)\} dN_i^*(t) \frac{\{R_i - \pi(Q_i, \hat{\psi}_n)\}}{\pi(Q_i, \hat{\psi}_n)} \varrho_\gamma^T(Q_i, \hat{\gamma}_n).$$

Let $\hat{I}_\gamma = \hat{A}_\gamma^{-1}$, where

$$\hat{A}_\gamma = n^{-1}\sum_{i=1}^{n}\frac{R_i\varrho_\gamma(Q_i,\hat{\gamma}_n)\varrho_\gamma^T(Q_i,\hat{\gamma}_n)}{\varrho(Q_i,\hat{\gamma}_n)\{1-\varrho(Q_i,\hat{\gamma}_n)\}},$$

By the LIE, it is straightforward to show that

$$\begin{aligned} E(T_1T_2^T) &= E\left[\frac{\{R-\pi(Q,\psi^*)\}}{\pi(Q,\psi^*)}\varrho(Q,\gamma^*)\{1-\varrho(Q,\gamma^*)\}\right. \\ &\quad \left.\times\int\{a(t,X)-\mu_a(t)\}^{\otimes 2}dN^*(t)\right]. \end{aligned}$$

Note that $E(T_1T_2^T)$ is symmetric. It can be consistently estimated by

$$\begin{aligned} \hat{E}(T_1T_2^T) &= n^{-1}\sum_{i=1}^{n}\frac{\{R_i-\pi(Q_i,\hat{\psi}_n)\}}{\pi(Q_i,\hat{\psi}_n)}\varrho(Q_i,\hat{\gamma}_n)\{1-\varrho(Q_i,\hat{\gamma}_n)\} \\ &\quad \times\int\{a(t,X_i)-\bar{a}(t,\hat{\beta}_n)\}^{\otimes 2}dN_i^*(t). \end{aligned}$$

In addition,

$$E(\varphi^*S_\gamma^T) = E\left[\int\{a(t,X)-\mu_a(t)\}dN^*(t)\frac{R}{\pi(Q,\psi^*)}\varrho_\gamma^T(Q,\gamma^*)\right].$$

Note that $E(\varphi^*S_\gamma^T)\neq P_\gamma$. It can be estimated by

$$\hat{E}(\varphi^*S_\gamma^T) = n^{-1}\sum_{i=1}^{n}\int\{a(t,X_i)-\bar{a}(t,\hat{\beta}_n)\}dN_i^*(t)\frac{R_i}{\pi(Q_i,\hat{\psi}_n)}\varrho_\gamma^T(Q_i,\hat{\gamma}_n).$$

Note that $\hat{E}(\varphi^*S_\gamma^T)$ is a consistent estimator for $E(\varphi^*S_\gamma^T)$ only if the ϱ model is correctly specified. However, it does not matter because $E(\varphi^*S_\gamma^T)$ appears only in product terms along with P_γ, which vanishes when the π model is correctly specified, and because when the ϱ model is misspecified, the π model has to be correctly specified by assumption.

Therefore, when the ϱ model is correctly specified, we have

$$\begin{aligned} \hat{\text{Var}}(\hat{\beta}_n) &= n^{-1}\hat{V}^{-1}(a,X)\{\hat{V}(a,a)+\hat{E}(T_2T_2^T)+2\hat{E}(T_1T_2^T)+\hat{P}_\gamma\hat{I}_\gamma^{-1}\hat{P}_\gamma^T \\ &\quad -\hat{E}(\varphi^*S_\gamma^T)\hat{I}_\gamma^{-1}\hat{P}_\gamma^T-\hat{P}_\gamma\hat{I}_\gamma^{-1}[\hat{E}(\varphi^*S_\gamma^T)]^T\}\hat{V}^{-T}(a,X). \end{aligned}$$

In some circumstances, we know that one of the two parametric models is correctly specified, but do not know which one it is. By (3.36) and (3.38),

$$n^{1/2}(\hat{\beta}_n-\beta_0) = n^{-1/2}\sum_{i=1}^{n}V^{-1}(a,X)\{\varphi_i^*-P_\psi I_\psi^{-1}S_{\psi i}-P_\gamma I_\gamma^{-1}S_{\gamma i}\}+o_p(1).$$

Therefore, the influence function for $\hat{\beta}_n$ is given by $\varphi = V^{-1}(a,X)\{\varphi^*-P_\psi I_\psi^{-1}S_\psi - P_\gamma I_\gamma^{-1}S_\gamma\}$.

By the CLT, $n^{1/2}(\hat\beta_n - \beta_0) \xrightarrow{d} N(0,\Sigma)$, where $\Sigma = E(\varphi\varphi^T)$.

Note that $P_\gamma = 0$ if the π model is correctly specified, $P_\psi = 0$ if the ϱ model is correctly specified. In addition, when the π model is correctly specified, $E(\varphi^* S_\psi^T) = P_\psi$. Also recall $E(T_1T_1^T) = V(a,a)$, $E(T_1T_2^T) = E(T_2T_1^T)$. Consequently,

$$\begin{aligned}\Sigma \;=\; & V^{-1}(a,X)[V(a,a) + E(T_2T_2^T) + 2E(T_1T_2^T) - P_\psi I_\psi^{-1}P_\psi^T + P_\gamma I_\gamma^{-1}P_\gamma^T \\ & -E(\varphi^* S_\gamma^T)I_\gamma^{-1}P_\gamma^T - P_\gamma I_\gamma^{-1}\{E(\varphi^* S_\gamma^T)\}^T]V^{-T}(a,X).\end{aligned}$$

Therefore, when either of the two parametric models is correctly specified, we have

$$\begin{aligned}\hat{\text{Var}}(\hat\beta_n) \;=\; & n^{-1}\hat V^{-1}(a,X)[\hat V(a,a) + \hat E(T_2T_2^T) + 2\hat E(T_1T_2^T) - \hat P_\psi \hat I_\psi^{-1}\hat P_\psi^T + \hat P_\gamma \hat I_\gamma^{-1}\hat P_\gamma^T \\ & -\hat E(\varphi^* S_\gamma^T)\hat I_\gamma^{-1}\hat P_\gamma^T - \hat P_\gamma \hat I_\gamma^{-1}\{\hat E(\varphi^* S_\gamma^T)\}^T]\hat V^{-T}(a,X).\end{aligned}$$

3.7 Locally Semiparametric Efficient Estimator

By results from the previous section, the search for the observed data efficient influence function can be restricted to the following class of influence functions,

$$\left\{\frac{R\varphi^F(Z)}{\pi(Q)} - \Pi\left[\left.\frac{R\varphi^F(Z)}{\pi(Q)}\right|\Lambda_2\right] :\; \varphi^F \in \Lambda_*^{F\perp}\right\}.$$

But since the observed data efficient score differs from the observed data efficient influence function only by a proportionality constant matrix, and an element of $\Lambda^{F\perp}$ differs from the corresponding full data influence function only by a proportionality constant matrix, we only need to identify $B_{\mathit{eff}}^F \in \Lambda^{F\perp}$, such that

$$S_{\mathit{eff}}(O) = \frac{RB_{\mathit{eff}}^F(Z)}{\pi(Q)} - \Pi\left[\left.\frac{RB_{\mathit{eff}}^F(Z)}{\pi(Q)}\right|\Lambda_2\right]. \tag{3.40}$$

Let $\mathbf{m}(\cdot)$ denote the linear operator mapping $\mathcal{H}^F$ to $\mathcal{H}^F$ as

$$\mathbf{m}\{h^F(Z)\} = E[E\{h^F(Z)|O\}|Z],$$

then $(\mathbf{i}-\mathbf{m})$ is a contractor, where $\mathbf{i}$ is the identity operator. Therefore, $\mathbf{m}^{-1}$ exists and is unique. Furthermore, by Proposition 8.1 of Robins, et al. (1994), (3.40) is equivalent to

$$\Pi[\mathbf{m}^{-1}\{B_{\mathit{eff}}^F(Z)\}|\Lambda^{F\perp}] = S_{\mathit{eff}}^F(Z). \tag{3.41}$$

In addition, the solution in $B_{\mathit{eff}}^F(Z) \in \Lambda^{F\perp}$ exists and is unique.

It is straightforward to show that

$$\mathbf{m}^{-1}\{B_{\mathit{eff}}^F(Z)\} = B_{\mathit{eff}}^F(Z) + \frac{\{1-\pi(Q)\}}{\pi(Q)}[B_{\mathit{eff}}^F(Z) - E\{B_{\mathit{eff}}^F(Z)|Q\}].$$

On the other hand, by (3.5),

$$B^F_{\mathit{eff}}(Z) = \int \{a(t, X) - \mu_a(t)\} dM(t), \tag{3.42}$$

for some $q \times 1$ function $a(t, X)$. Therefore,

$$\begin{aligned}\mathbf{m}^{-1}\{B^F_{\mathit{eff}}(Z)\} &= \int \{a(t, X) - \mu_a(t)\} dM(t) \\ &+ \frac{\{1 - \pi(Q)\}}{\pi(Q)} \{I(\Delta = 2) - \varrho(Q)\} \int \{a(t, X) - \mu_a(t)\} dN^*(t)\end{aligned} \tag{3.43}$$

By (3.41) and the projection theorem,

$$0 = E\{[\mathbf{m}^{-1}\{B^F_{\mathit{eff}}(Z)\} - S^F_{\mathit{eff}}(Z)]^T h^F(Z)\}, \ \forall h^F \in \Lambda^{F\perp}.$$

Therefore, by (3.5), (3.8), (3.43), along with martingale properties and the LIE, we have that

$$\begin{aligned} 0 &= E\Bigg\{\Bigg[\int \{a(t, X) - \mu_a(t)\} dM(t) \\ &\quad + \frac{\{1 - \pi(Q)\}}{\pi(Q)} \{I(\Delta = 2) - \varrho(Q)\} \int \{a(t, X) - \mu_a(t)\} dN^*(t) \\ &\quad - \int \{X - \mu_X(t)\} dM(t)\Bigg]^T \\ &\quad \times \int \{b(t, X) - \mu_b(t)\} dM(t)\Bigg\} \\ &= E\Bigg[\int (\{a(t, X) - \mu_a(t)\} - \{X - \mu_X(t)\})^T \{b(t, X) - \mu_b(t)\} \\ &\quad \times \lambda(t) e^{\beta_0^T X} I(T \geq t) dt\Bigg] \\ &\quad + E\Bigg[\frac{\{1 - \pi(Q)\}}{\pi(Q)} \varrho(Q)\{1 - \varrho(Q)\} \\ &\quad \times \{a(T, X) - \mu_a(T)\}^T \{b(T, X) - \mu_b(T)\}\Bigg], \ \forall b(t, X). \end{aligned} \tag{3.44}$$

By martingale properties and the LIE, the first term on the RHS of (3.44) is equal to

$$\begin{aligned} & E\left[\int (\{a(t, X) - \mu_a(t)\} - \{X - \mu_X(t)\})^T \{b(t, X) - \mu_b(t)\} dN(t)\right] \\ = \ & E[\varrho(Q)(\{a(T, X) - \mu_a(T)\} - \{X - \mu_X(T)\})^T \{b(T, X) - \mu_b(T)\}] \\ = \ & E[f^*(T, X)(\{a(T, X) - \mu_a(T)\} - \{X - \mu_X(T)\})^T \{b(T, X) - \mu_b(T)\}], \end{aligned} \tag{3.45}$$

where

$$f^*(T, X) = E\{\varrho(Q) | T, X\}.$$

Similarly, by the LIE, the second term on the RHS of (3.44) is equal to

$$E[g^*(T,X)\{a(T,X)-\mu_a(T)\}^T\{b(T,X)-\mu_b(T)\}], \tag{3.46}$$

where

$$g^*(T,X) = E\left[\frac{\{1-\pi(Q)\}}{\pi(Q)}\varrho(Q)\{1-\varrho(Q)\}\middle| T,X\right].$$

Substituting (3.45) and (3.46) into (3.44), we have that

$$\begin{aligned} 0 &= E[(\{f^*(T,X)+g^*(T,X)\}\{a(T,X)-\mu_a(T)\} \\ &\quad -f^*(T,X)\{X-\mu_X(T)\})^T\{b(T,X)-\mu_b(T)\}],\ \forall b(t,X). \end{aligned} \tag{3.47}$$

Assume that

$$P(\Delta=2|T,X,\Delta>0)>0, \quad P(\Delta>0|T,X)>0.$$

Then by the LIE,

$$\begin{aligned} f^*(T,X) &= P(\Delta=2|T,X) \\ &= P(\Delta=2|T,X,\Delta>0)\ P(\Delta>0|T,X) \\ &> 0. \end{aligned}$$

Consequently, $f^*(T,X)+g^*(T,X)>0$ because $g^*(T,X)\geq 0$.

Let

$$a(t,X)-\mu_a(t) = h(t,X)\{X-\alpha_h(t)\}, \tag{3.48}$$

where

$$\begin{aligned} h(t,X) &= \frac{f^*(t,X)}{f^*(t,X)+g^*(t,X)}, \\ \alpha_h(t) &= \frac{E\{h(t,X)Xe^{\beta_0^TX}I(T\geq t)\}}{E\{h(t,X)e^{\beta_0^TX}I(T\geq t)\}}. \end{aligned}$$

Then, by arguments similar to those used in (3.45) except that we are now working in the reverse direction, the RHS of (3.47) is equal to

$$\begin{aligned} & E[(f^*(T,X)\{X-\alpha_h(T)\}-f^*(T,X)\{X-\mu_X(T)\})^T\{b(T,X)-\mu_b(T)\}] \\ =\ & E[f^*(T,X)\{\mu_X(T)-\alpha_h(T)\}^T\{b(T,X)-\mu_b(T)\}] \\ =\ & E\left[\int\{\mu_X(t)-\alpha_h(t)\}^T\{b(t,X)-\mu_b(t)\}dN(t)\right] \\ =\ & \int\{\mu_X(t)-\alpha_h(t)\}^TE[\{b(t,X)-\mu_b(t)\}e^{\beta_0^TX}I(T\geq t)]\lambda(t)dt \\ =\ & 0,\ \forall b(t,X). \end{aligned}$$

Therefore, by the uniqueness, we have identified the observed data efficient score defined by (3.40), (3.42), and (3.48).

In addition, by results from the previous section, the observed data efficient score can be written as

$$
\begin{aligned}
S_{\mathit{eff}}(O) &= \int h(t,X)\{X-\alpha_h(t)\}dM(t) \\
&\quad +\Omega(R,Z)\int h(t,X)\{X-\alpha_h(t)\}dN^*(t) \\
&= \Phi(R,Z)\int h(t,X)\{X-\alpha_h(t)\}dN^*(t) \\
&\quad -\int h(t,X)\{X-\alpha_h(t)\}\lambda(t)e^{\beta_0^T X}I(T\geq t)dt,
\end{aligned}
$$

where $\Omega(R,Z)=\Omega(R,Z;\psi_0,\gamma_0)$.

In fact, the observed data efficient score can also be obtained directly by projecting the observed data score for β onto the space orthogonal to the observed data nuisance tangent space as we will now demonstrate.

By (3.6), the observed data score for β is given by

$$
\begin{aligned}
S_\beta(O) &= E\{S_\beta^F(Z)|O\} \\
&= E\left[\int XdM(t)\middle| O\right] \\
&= R\int XdM(t)+(1-R)\int X\{\varrho(Q)dN^*(t)-\lambda(t)e^{\beta_0^T X}I(T\geq t)dt\} \\
&= \int XdM(t)-(1-R)\{I(\Delta=2)-\varrho(Q)\}\int XdN^*(t).
\end{aligned}
$$

Note that $\Lambda=\Lambda_\psi\oplus\Lambda_\eta$, where $\Lambda_\eta=E\{\Lambda^F|O\}$, $\Lambda_\eta^\perp=\frac{R}{\pi(Q)}\Lambda^{F\perp}+\Lambda_2=\Lambda_0^\perp$. In addition, $S_\beta\perp\Lambda_\psi$, hence $\Pi[S_\beta|\Lambda_\psi]=0$. Consequently,

$$
S_{\mathit{eff}}=\Pi[S_\beta|\Lambda^\perp]=\Pi[S_\beta|\Lambda_\eta^\perp].
$$

Therefore, by the projection theorem, we have

$$
\begin{aligned}
0 &= E\Bigg\{\Bigg(\left[\int XdM(t)-(1-R)\{I(\Delta=2)-\varrho(Q)\}\int XdN^*(t)\right] \\
&\quad -\Bigg[\int\{a(t,X)-\mu_a(t)\}dM(t) \\
&\quad +\Omega(R,Z)\int\{a(t,X)-\mu_a(t)\}dN^*(t)\Bigg]\Bigg)^T \\
&\quad \times\Bigg[\frac{R}{\pi(Q)}\int\{b(t,X)-\mu_b(t)\}dM(t) \\
&\quad -\frac{\{R-\pi(Q)\}}{\pi(Q)}L(Q)\Bigg]\Bigg\},\ \forall b(t,X),\forall L(Q).
\end{aligned}
$$

Equivalently,

$$
\begin{aligned}
0 \quad = \quad & E\Bigg\{\Bigg[\int\{a(t,X)-\mu_a(t)-X\}dM(t) \\
& +\Omega(R,Z)\int\{a(t,X)-\mu_a(t)\}dN^*(t) \\
& +(1-R)\{I(\Delta=2)-\varrho(Q)\}\int XdN^*(t)\Bigg]^T \\
& \times\left[\frac{R}{\pi(Q)}\int\{b(t,X)-\mu_b(t)\}dM(t)\right. \\
& \left.-\frac{\{R-\pi(Q)\}}{\pi(Q)}L(Q)\right]\Bigg\},\ \forall b(t,X),\forall L(Q).
\end{aligned}
$$

By martingale properties and the LIE, this reduces to

$$
\begin{aligned}
0 \quad = \quad & E\left[\int\{a(t,X)-\mu_a(t)-X\}^T\{b(t,X)-\mu_b(t)\}\right. \\
& \left.\times\lambda(t)e^{\beta_0^TX}I(T\geq t)dt\right] \\
& +E\left[\frac{\{1-\pi(Q)\}}{\pi(Q)}\varrho(Q)\{1-\varrho(Q)\}\right. \\
& \left.\times\{a(T,X)-\mu_a(T)\}^T\{b(T,X)-\mu_b(T)\}\right],\ \forall b(t,X). \qquad (3.49)
\end{aligned}
$$

Arguments similar to those applied to (3.44) can be applied to (3.49) to show that the observed data efficient score is given by (3.40), (3.42), and (3.48).

Denote

$$
\begin{aligned}
f(T,X) &= E\{\varrho(W)|T,X,\Delta>0\}, \\
g(T,X) &= E\left[\frac{\{1-\pi(W)\}}{\pi(W)}\varrho(W)\{1-\varrho(W)\}\middle|T,X,\Delta>0\right],
\end{aligned}
$$

By the LIE, we have that

$$
\begin{aligned}
f^*(T,X) &= f(T,X)P(\Delta>0|T,X), \\
g^*(T,X) &= g(T,X)P(\Delta>0|T,X).
\end{aligned}
$$

Therefore,

$$
h(t,X)=\frac{f(t,X)}{f(t,X)+g(t,X)}.
$$

To construct estimating equations for β based on the observed data efficient score, we need to estimate $f(T_i,X_i)$ and $g(T_i,X_i)$ for each individual, which may involve modeling the conditional expectations, e.g., we may consider a logit model for f and a loglinear model for g, respectively. Let θ denote the finite-dimensional parameters introduced in this modeling process, and let $\hat{\theta}_n$ denote an estimator for θ satisfying

$$
\hat{\theta}_n \xrightarrow{p} \theta^*, \quad n^{1/2}(\hat{\theta}_n-\theta^*)=O_p(1). \qquad (3.50)
$$

Then we can estimate $f(T_i, X_i)$ and $g(T_i, X_i)$ using $f(T_i, X_i, \hat{\theta}_n)$ and $g(T_i, X_i, \hat{\theta}_n)$, respectively.

Let

$$\begin{aligned} h(t, X_i, \theta) &= \frac{f(t, X_i, \theta)}{f(t, X_i, \theta) + g(t, X_i, \theta)}, \\ \bar{\alpha}_h(t, \beta, \theta) &= \frac{\sum_{j=1}^n h(t, X_j, \theta) X_j e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^n h(t, X_j, \theta) e^{\beta^T X_j} I(T_j \geq t)}. \end{aligned}$$

Then it is straightforward to show that

$$\sum_{i=1}^n \int h(t, X_i, \theta)\{X_i - \bar{\alpha}_h(t, \beta, \theta)\}\lambda(t) e^{\beta^T X_i} I(T_i \geq t) dt = 0, \ \forall \beta, \forall \theta. \tag{3.51}$$

Consequently, S_{eff} suggests the following estimating equation for β,

$$0 = \sum_{i=1}^n \Phi(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int h(t, X_i, \hat{\theta}_n)\{X_i - \bar{\alpha}_h(t, \beta, \hat{\theta}_n)\} dN_i^*(t). \tag{3.52}$$

By (3.27) and (3.51), this is identical to

$$\begin{aligned} 0 = n^{-1} \sum_{i=1}^n \Bigg[& \int h(t, X_i, \hat{\theta}_n)\{X_i - \bar{\alpha}_h(t, \beta, \hat{\theta}_n)\} dM_i(t, \beta) \\ & + \Omega(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int h(t, X_i, \hat{\theta}_n)\{X_i - \bar{\alpha}_h(t, \beta, \hat{\theta}_n)\} dN_i^*(t) \Bigg]. \end{aligned} \tag{3.53}$$

By the WLLN, $\bar{\alpha}_h(t, \beta_0, \hat{\theta}_n) \xrightarrow{p} \alpha_h(t, \theta^*)$, where

$$\alpha_h(t, \theta^*) = \frac{E\{h(t, X, \theta^*) X e^{\beta_0^T X} I(T \geq t)\}}{E\{h(t, X, \theta^*) e^{\beta_0^T X} I(T \geq t)\}}.$$

Assume that $\hat{\psi}_n \xrightarrow{p} \psi^*$, $\hat{\gamma}_n \xrightarrow{p} \gamma^*$, and that either of the two parametric models is correctly specified, then, when evaluated at β_0, a typical summand of (3.53) is asymptotically equivalent to $\varphi_i^* \equiv T_{1i} + T_{2i}$, where

$$\begin{aligned} T_1 &= \int h(t, X, \theta^*)\{X - \alpha_h(t, \theta^*)\} dM(t), \\ T_2 &= \Omega(R, Z; \psi^*, \gamma^*) \int h(t, X, \theta^*)\{X - \alpha_h(t, \theta^*)\} dN^*(t). \end{aligned}$$

Note that had the parametric models for the conditional expectations involved in modeling $f(t, X)$ and $g(t, X)$ been correctly specified, we would have $\theta^* = \theta_0$. In addition, if both the π model and the ϱ model were also correctly specified, then we would have $\varphi^* = S_{eff}$ as expected.

By martingale properties and the DR argument, $E(\varphi^*) = 0$. Therefore, (3.52) is an asymptotically unbiased estimating equation for β. Consequently, under certain regularity conditions, the resulting estimator is consistent.

By definition, we have that

$$\begin{aligned} \frac{\partial \bar{\alpha}_h(t,\beta,\theta)}{\partial \beta^T} &= \frac{\sum_{j=1}^n h(t,X_j,\theta)\{X_j - \bar{\alpha}_h(t,\beta,\theta)\}^{\otimes 2} e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^n h(t,X_j,\theta) e^{\beta^T X_j} I(T_j \geq t)} \\ &\equiv K_n(h,\alpha_h;t,\beta,\theta). \end{aligned}$$

Expanding (3.52) about β_0, we have that

$$\begin{aligned} & n^{1/2}(\hat{\beta}_n - \beta_0) \\ = \; & \left\{ n^{-1} \sum_{i=1}^n \Phi(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int h(t, X_i, \hat{\theta}_n) K_n(h, \alpha_h; t, \beta_n^*, \hat{\theta}_n) dN_i^*(t) \right\}^{-1} \\ & \times n^{-1/2} \sum_{i=1}^n \Phi(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int h(t, X_i, \hat{\theta}_n) \{X_i - \bar{\alpha}_h(t, \beta_0, \hat{\theta}_n)\} dN_i^*(t), \qquad (3.54) \end{aligned}$$

where β_n^* lies between $\hat{\beta}_n$ and β_0.

Since $\hat{\beta}_n \xrightarrow{p} \beta_0$, $\beta_n^* \xrightarrow{p} \beta_0$. By the WLLN, $\bar{\alpha}_h(t, \beta_n^*, \hat{\theta}_n) \xrightarrow{p} \alpha_h(t, \theta^*)$, so that $K_n(h, \alpha_h; t, \beta_n^*, \hat{\theta}_n) \xrightarrow{p} \tau(h, \alpha_h; t, \theta^*)$, where

$$\tau(h, \alpha_h; t, \theta^*) = \frac{E[h(t, X, \theta^*)\{X - \alpha_h(t, \theta^*)\}^{\otimes 2} e^{\beta_0^T X} I(T \geq t)]}{E\{h(t, X, \theta^*) e^{\beta_0^T X} I(T \geq t)\}}.$$

Similar to the previous section, by the DR argument, the leading matrix inside the bracket on the RHS of (3.54) converges in probability to

$$V(h, \alpha_h; \theta^*) = \int E[h(t, X, \theta^*)\{X - \alpha_h(t, \theta^*)\}^{\otimes 2} e^{\beta_0^T X} I(T \geq t)] \lambda(t) dt.$$

This suggests that $V(h, \alpha_h; \theta^*)$ can be estimated by

$$\hat{V}(h, \alpha_h; \theta^*) = n^{-1} \sum_{i=1}^n \Phi(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int h(t, X_i, \hat{\theta}_n) K_n(h, \alpha_h; t, \hat{\beta}_n, \hat{\theta}_n) dN_i^*(t).$$

An alternative estimator of $V(h, \alpha_h; \theta^*)$ is provided by

$$\tilde{V}(h, \alpha_h; \theta^*) = n^{-1} \sum_{i=1}^n \Phi(R_i, Z_i; \hat{\psi}_n, \hat{\gamma}_n) \int h(t, X_i, \hat{\theta}_n) \{X_i - \bar{\alpha}_h(t, \hat{\beta}_n, \hat{\theta}_n)\}^{\otimes 2} dN_i^*(t).$$

It is straightforward to show that

$$\begin{aligned} \frac{\partial \bar{\alpha}_h(t,\beta,\theta)}{\partial \theta^T} &= \frac{\sum_{j=1}^n \{X_j - \bar{\alpha}_h(t,\beta,\theta)\} h_\theta^T(t,X_j,\theta) e^{\beta^T X_j} I(T_j \geq t)}{\sum_{j=1}^n h(t,X_j,\theta) e^{\beta^T X_j} I(T_j \geq t)} \\ &\equiv \zeta_n(t,\beta,\theta). \end{aligned}$$

Therefore,

$$\frac{\partial h(t,X,\theta)\{X-\bar{\alpha}_h(t,\beta,\theta)\}}{\partial\theta^T} \;=\; \{X-\bar{\alpha}_h(t,\beta,\theta)\}h_\theta^T(t,X,\theta) - h(t,X,\theta)\zeta_n(t,\beta,\theta).$$

By (3.27), (3.51), and by expanding about $(\psi^*,\gamma^*,\theta^*)$, the $q\times 1$ vector on the RHS of (3.54) is equal to

$$\begin{aligned}
& n^{-1/2}\sum_{i=1}^n\Big[\int h(t,X_i,\hat{\theta}_n)\{X_i-\bar{\alpha}_h(t,\beta_0,\hat{\theta}_n)\}dM_i(t) \\
&\qquad +\Omega(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int h(t,X_i,\hat{\theta}_n)\{X_i-\bar{\alpha}_h(t,\beta_0,\hat{\theta}_n)\}dN_i^*(t)\Big] \\
= \;& n^{-1/2}\sum_{i=1}^n\Big[\int h(t,X_i,\theta^*)\{X_i-\bar{\alpha}_h(t,\beta_0,\theta^*)\}dM_i(t) \\
&\qquad +\Omega(R_i,Z_i;\psi^*,\gamma^*)\int h(t,X_i,\theta^*)\{X_i-\bar{\alpha}_h(t,\beta_0,\theta^*)\}dN_i^*(t)\Big] \\
& -\Big[n^{-1}\sum_{i=1}^n\int h(t,X_i,\theta_n^*)\{X_i-\bar{\alpha}_h(t,\beta_0,\theta_n^*)\}dN_i^*(t) \\
&\qquad \times\{I(\Delta_i=2)-\varrho(W_i,\gamma_n^*)\}\frac{R_i\pi_\psi^T(Q_i,\psi_n^*)}{\pi^2(Q_i,\psi_n^*)}\Big]n^{1/2}(\hat{\psi}_n-\psi^*) \\
& -\Big[n^{-1}\sum_{i=1}^n\int h(t,X_i,\theta_n^*)\{X_i-\bar{\alpha}_h(t,\beta_0,\theta_n^*)\}dN_i^*(t) \\
&\qquad \times\frac{\{R_i-\pi(Q_i,\psi_n^*)\}}{\pi(Q_i,\psi_n^*)}\varrho_\gamma^T(W_i,\gamma_n^*)\Big]n^{1/2}(\hat{\gamma}_n-\gamma^*) \\
& +\Big[n^{-1}\sum_{i=1}^n\Big\{\int[\{X_i-\bar{\alpha}_h(t,\beta_0,\theta_n^*)\}h_\theta^T(t,X_i,\theta_n^*) \\
&\qquad\qquad - h(t,X_i,\theta_n^*)\zeta_n(t,\beta_0,\theta_n^*)]dM_i(t) \\
&\qquad +\Omega(R_i,Z_i;\psi_n^*,\gamma_n^*)\int[\{X_i-\bar{\alpha}_h(t,\beta_0,\theta_n^*)\}h_\theta^T(t,X_i,\theta_n^*) \\
&\qquad\qquad - h(t,X_i,\theta_n^*)\zeta_n(t,\beta_0,\theta_n^*)]dN_i^*(t)\Big\}\Big]n^{1/2}(\hat{\theta}_n-\theta^*),
\end{aligned} \tag{3.55}$$

where ψ_n^* lies between $\hat{\psi}_n$ and ψ^*, γ_n^* lies between $\hat{\gamma}_n$ and γ^*, and θ_n^* lies between $\hat{\theta}_n$ and θ^*.

Since $\hat{\psi}_n\xrightarrow{p}\psi^*$, $\hat{\gamma}_n\xrightarrow{p}\gamma^*$, $\hat{\theta}_n\xrightarrow{p}\theta^*$, we have $\psi_n^*\xrightarrow{p}\psi^*$, $\gamma_n^*\xrightarrow{p}\gamma^*$, $\theta_n^*\xrightarrow{p}\theta^*$. By the WLLN, $\bar{\alpha}_h(t,\beta_0,\theta_n^*)\xrightarrow{p}\alpha_h(t,\theta^*)$, so that $\zeta_n(t,\beta_0,\theta_n^*)\xrightarrow{p}\zeta(t,\theta^*)$, where

$$\zeta(t,\theta^*)=\frac{E[\{X-\alpha_h(t,\theta^*)\}h_\theta^T(t,X,\theta^*)e^{\beta_0^TX}I(T\ge t)]}{E[h(t,X,\theta^*)e^{\beta_0^TX}I(T\ge t)]}.$$

Therefore, the leading matrix inside the bracket of the fourth term of (3.55) converges in probability to

$$E\left[\int [\{X-\alpha_h(t,\theta^*)\}h_\theta^T(t,X,\theta^*) - h(t,X,\theta^*)\zeta(t,\theta^*)]dM(t)\right.$$
$$\left. +\Omega(R,Z;\psi^*,\gamma^*)\int [\{X-\alpha_h(t,\theta^*)\}h_\theta^T(t,X,\theta^*) - h(t,X,\theta^*)\zeta(t,\theta^*)]dN^*(t)\right].$$

By martingale properties and the DR argument, this is equal to zero. Consequently, by (3.50), the fourth term on the RHS of (3.55), which measures the effect of the estimation of θ on the estimation of β, is negligible.

Let

$$a(t,X) - \mu_a(t) = h(t,X,\theta^*)\{X-\alpha_h(t,\theta^*)\}.$$

Then arguments similar to those used in the previous section can be applied to reach almost identical conclusions. For instance, the influence function for $\hat{\beta}_n$ is given by $\varphi = V^{-1}(h,\alpha_h;\theta^*)\{\varphi^* - P_\psi^* I_\psi^{-1} S_\psi - P_\gamma I_\gamma^{-1} S_\gamma\}$. By the CLT, $n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0,\Sigma)$, where

$$\begin{aligned}\Sigma \;=\; & V^{-1}(h,\alpha_h;\theta^*)[V(h^2,\alpha_h;\theta^*) + E(T_2T_2^T) + 2E(T_1T_2^T) - P_\psi I_\psi^{-1} P_\psi^T + P_\gamma I_\gamma^{-1} P_\gamma^T \\ & -E(\varphi^* S_\gamma^T) I_\gamma^{-1} P_\gamma^T - P_\gamma I_\gamma^{-1}\{E(\varphi^* S_\gamma^T)\}^T]V^{-T}(h,\alpha_h;\theta^*).\end{aligned}$$

Similar estimators can be used by plugging in $\hat{\theta}_n$ for θ^* with the only exception being $\hat{V}(h,\alpha_h;\theta^*)$ and $\hat{V}(h^2,\alpha_h;\theta^*)$, where

$$\hat{V}(h^2,\alpha_h;\theta^*) = n^{-1}\sum_{i=1}^n \Phi(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int h^2(t,X_i,\hat{\theta}_n)K_n(h^2,\alpha_h;t,\hat{\beta}_n,\hat{\theta}_n)dN_i^*(t).$$

3.8 Simulation Study

We considered three IPW semiparametric estimators. The first one is the simple IPWCC estimator with estimating equation given by

$$0 \;=\; \sum_{i=1}^n \frac{R_i}{\pi(Q_i,\hat{\psi}_n)}\int \{X_i - \bar{X}(t,\beta,\hat{\psi}_n)\}dN_i(t).$$

The second one is the IPWDR estimator which is guaranteed to have improvement on robustness and efficiency over the IPWCC estimator. The estimating equation is given by

$$0 \;=\; \sum_{i=1}^n \Phi(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int \{X_i - \bar{X}(t,\beta)\}dN_i^*(t).$$

The third one is the IPWLE estimator with estimating equation given by

$$0 \;=\; \sum_{i=1}^n \Phi(R_i,Z_i;\hat{\psi}_n,\hat{\gamma}_n)\int h(t,X_i,\hat{\theta}_n)\{X_i - \bar{\alpha}_h(t,\beta,\hat{\theta}_n)\}dN_i^*(t).$$

We will fit a logit π model and a logit ϱ model in the simulation study no matter what the true models are. In addition, to obtain the IPWLE estimator, we will fit a logit model to f and a loglinear model to g among those individuals who have failed. To avoid nonconvergence, we will run a linear regression on the transformed scale of f and g, and then transform the predictions back to the original scale.

For comparison purpose, we also considered the complete case estimator and the (single) imputation estimator. In order to investigate the performance in terms of robustness of these estimators, we considered two situations.

In the first situation, we generated a random treatment assignment X_i from a Bernoulli distribution with success probability 0.5. Given $X_i = x$, we generated two failure times $\{T_{2i}, T_{1i}\}$ and a censoring time C_i, independently, as we will now describe. The failure time due to the cause of interest, i.e., T_{2i}, was generated from an exponential distribution with hazard $\lambda(t|x) = \phi e^{\beta x}$, where $\phi = 1$ and $\beta = 0.4$. The failure time due to competing causes, i.e., T_{1i}, was generated from a Weibull distribution with hazard $\lambda_1(t|x) = t^{\gamma_1 - 1} e^{\gamma_2 x}$, where $\gamma_1 = 1.4$ and $\gamma_2 = 0.2$. The censoring time C_i was generated from an exponential distribution with hazard $\lambda_0(t|x) = \lambda_C = 0.3$. Compute $T_i = \min\{T_{2i}, T_{1i}, C_i\}$, and let $\Delta_i = 2$ if $\{T_{2i} \leq T_{1i}, T_{2i} \leq C_i\}$, $\Delta_i = 1$ if $\{T_{1i} < T_{2i}, T_{1i} \leq C_i\}$, and $\Delta_i = 0$ if $\{C_i < T_{2i}, C_i < T_{1i}\}$. Given $\{T_i = t, X_i = x, \Delta_i = \delta\}$, we generated an auxiliary covariate A_i from an exponential distribution with hazard $\lambda(a|t, x, \delta) = \xi_0 + \xi_1 \delta$, where $\xi_0 = 1$ and $\xi_1 = 1$. Consequently,

$$\text{logit } \varrho(t, x, a) = -\log[(\xi_0 + \xi_1)/\{\phi(\xi_0 + 2\xi_1)\}] - (\gamma_1 - 1)\log t - (\gamma_2 - \beta)x - \xi_1 a.$$

Note that, without prior knowledge, we will fit a logit model linear in t instead of $\log t$, i.e., logit $\varrho(t, x, a) = \gamma_0^* + \gamma_1^* t + \gamma_2^* x + \gamma_3^* a$. To comply with the MAR assumption, we generated the missingness indicator R_i from a Bernoulli distribution with success probability depending only on (T_i, X_i, A_i). In particular, we let

$$\text{logit } \pi(t, x, a) = \psi_0 + \psi_1 t + \psi_2 x + \psi_3 a,$$

where $\psi_0 = 1$, $\psi_1 = 1$, $\psi_2 = -2$, $\psi_3 = 1$. Therefore, this is the same π model as we will fit.

In the second situation, we changed the specification for the conditional distribution of T_{1i} given X_i, which now follows a Gompertz distribution with hazard $\lambda_1(t|x) = e^{\gamma_1 t + \gamma_2 x}$, where $\gamma_1 = 0.5$ and $\gamma_2 = -0.5$, so that the true ϱ model is now given by

$$\text{logit } \varrho(t, x, a) = -\log[(\xi_0 + \xi_1)/\{\phi(\xi_0 + 2\xi_1)\}] - \gamma_1 t - (\gamma_2 - \beta)x - \xi_1 a.$$

Note that this is the same ϱ model as we will fit. In addition, we changed the specification for the π model, which is now given by

$$\text{logit } \pi(t, x, a) = \psi_0 + t^{\psi_1} + \psi_2 x + \psi_3 a,$$

where $\psi_0 = 1$, $\psi_1 = 0.5$, $\psi_2 = -2$, $\psi_3 = 1$. Note that, without prior knowledge, we will fit a logit π model.

For each situation, 500 simulation data sets were generated for sample sizes $n = 200, 500$. Because it is commonly believed that most IPW estimators and especially their variance estimators are not stable when some of the probabilities of having a complete case are close to zero, we also generated 200 bootstrap data sets for each simulation data set with the first bootstrap data set being the simulation data set itself. For the two situations considered above, there are about 52% to 55% failures from the cause of interest, 31% to 35% failures from the competing causes, and 13% to 14% censored observations. The proportion of missing cause of failure ranges from 28% to 30%. The results are summarized in Tables 3.1 and 3.2, where a "good" fit means that all variables in W are used in the logistic regression model and a "poor" bit means that only the intercept term is used in the logistic regression model and both refer to the fit of the misspecified model.

As can be seen from the two tables, the complete case estimator is biased in all cases, which is expected because cause of failure is not missing completely at random. The imputation estimator behaves fairly well, even in the case when the ϱ model is misspecified. However, further research is needed to draw any confirmative conclusions about the robustness of the imputation estimator against misspecification of the ϱ model. When the π model is correctly specified, the IPWCC estimator is unbiased but less efficient than the IPWDR estimator in all cases. When the π model is misspecified, the IPWCC estimator is biased, although the bias may not be pronounced if the modeling is carefully carried out so that the fitted π model is close to the truth. Also notice that the variance estimator for IPWCC overestimates the true sampling variation in all cases. However, this should not cause any particular concern because the IPWDR estimator is almost always a better choice compared to the IPWCC estimator. It is interesting to notice that the IPWDR estimator is very close to the IPWLE estimator in terms of robustness and efficiency. This suggests that it is not worth the effort to go through the complicated modeling processes for $f(T, X)$ and $g(T, X)$ in an attempt to gain extra efficiency if any. The IPWDR estimator is recommended for its simplicity.

3.9 Breast Cancer Example

The data from a clinical trial in elderly women with stage II breast cancer were analyzed to identify covariates that were significantly associated with death due to breast cancer. There were 169 eligible patients enrolled in this study, among which 90 patients had censored death times. Of the 79 patients who died, 18 of them had incomplete cause-of-death information. For the remaining patients with known cause of death, 44 died from breast cancer and the other 17 died of other causes. Cummings et al. (1986) reported two covariates, presence of 1 to 3 positive axillary lymph nodes and having an ER-positive primary tumor, as being significantly associated with overall survival. Goetghebeur and Ryan (1995) conducted a cause-specific survival analysis based on the standard proportional hazards structure for both failure types. Lu & Tsiatis (2000) used the same data to illustrate multiple imputation methods. We summarize their results in Table 3.3 along with the proposed IPWDR estimator as we will now describe. First, we had to establish a model for $\pi(W)$ and $\varrho(W)$. For the covariates W, we considered ER-status, number of positive axillary lymph nodes, tumor size, treatment assignment (Tamoxifen versus placebo), and time of relapse. As noted by the latter two authors, because among the 6 patients with ER-negative status 5 had died and all died from breast cancer, we can not use a logistic regression model that includes ER-status as a covariate to fit the π model or the ϱ model as the MLE does not exist. On the other hand, since only patients with ER-positive status had unknown cause of death, we can fit a logistic regression model for π using the subset of patients who died and were ER-positive, and a logistic regression model for ϱ using the subset of patients who died with known cause of death and were ER-positive. It turned out that none of the covariates except the intercept term was significant for both logistic regression models. The results are shown in Table 3.3, where the numbers inside the brackets denote the standard errors associated with the parameter estimates.

It can be seen from Table 3.3 that the hazard of death from breast cancer is significantly associated with ER-status, but no firm conclusions can be reached regarding the effect of number of positive axillary lymph nodes.

3.10 Discussion

We have investigated various inverse probability weighted semiparametric estimators which allow the inclusion of additional auxiliary covariates. We recommend to use the IPWDR estimator for its simplicity, flexibility, robustness and high efficiency.

Table 3.1: Monte Carlo comparison of complete cases, imputation, and inverse probability weighted estimators with sample size of 200

	CC	SI	IPWCC	IPWDR	IPWLE
π model correctly specified, ϱ model misspecified, good fit					
Bias	-0.2268	0.0030	-0.0001	-0.0017	-0.0453
SSE	0.2496	0.2239	0.2288	0.2148	0.2214
SEE	0.2574	0.2243	0.2978	0.2161	0.2164
CP	0.874	0.944	0.998	0.944	0.942
SEEBT	–	–	0.2346	0.2201	0.2282
CPBT	–	–	0.960	0.946	0.954
π model correctly specified, ϱ model misspecified, poor fit					
Bias	$*$	-0.0532	$*$	-0.0007	-0.0846
SSE	$*$	0.1993	$*$	0.2158	0.2232
SEE	$*$	0.1975	$*$	0.2168	0.2178
CP	$*$	0.940	$*$	0.940	0.922
SEEBT	–	–	$*$	0.2203	0.2286
CPBT	–	–	$*$	0.946	0.946
ϱ model correctly specified, π model misspecified, good fit					
Bias	-0.2019	-0.0011	-0.0047	-0.0017	-0.0421
SSE	0.2529	0.2293	0.2374	0.2266	0.2362
SEE	0.2550	0.2284	0.3167	0.2210	0.2206
CP	0.8700	0.964	0.996	0.950	0.930
SEEBT	–	–	0.2385	0.2246	0.2346
CPBT	–	–	0.958	0.952	0.956
ϱ model correctly specified, π model misspecified, poor fit					
Bias	$*$	$*$	-0.1799	-0.0019	-0.0083
SSE	$*$	$*$	0.2490	0.2236	0.2232
SEE	$*$	$*$	0.3652	0.2199	0.2198
CP	$*$	$*$	0.988	0.948	0.950
SEEBT	–	–	0.2533	0.2226	0.2220
CPBT	–	–	0.890	0.954	0.958

Table 3.2: Monte Carlo comparison of complete cases, imputation, and inverse probability weighted estimators with sample size of 500

	CC	SI	IPWCC	IPWDR	IPWLE
π model correctly specified, ϱ model misspecified, good fit					
Bias	-0.2257	-0.0002	-0.0039	-0.0006	-0.0417
SSE	0.1518	0.1357	0.1406	0.1299	0.1346
SEE	0.1604	0.1403	0.1869	0.1350	0.1352
CP	0.710	0.960	0.994	0.966	0.958
SEEBT	–	–	0.1457	0.1357	0.1404
CPBT	–	–	0.952	0.968	0.964
π model correctly specified, ϱ model misspecified, poor fit					
Bias	$*$	-0.0571	$*$	0.0005	-0.0796
SSE	$*$	0.1201	$*$	0.1305	0.1343
SEE	$*$	0.1236	$*$	0.1355	0.1361
CP	$*$	0.946	$*$	0.960	0.918
SEEBT	–	–	$*$	0.1361	0.1412
CPBT	–	–	$*$	0.964	0.934
ϱ model correctly specified, π model misspecified, good fit					
Bias	-0.1947	0.0021	-0.0041	0.0003	-0.0383
SSE	0.1493	0.1380	0.1400	0.1336	0.1390
SEE	0.1584	0.1426	0.1960	0.1377	0.1375
CP	0.768	0.970	0.996	0.962	0.946
SEEBT	–	–	0.1463	0.1385	0.1444
CPBT	–	–	0.962	0.964	0.960
ϱ model correctly specified, π model misspecified, poor fit					
Bias	$*$	$*$	-0.1735	0.0004	-0.0035
SSE	$*$	$*$	0.1457	0.1323	0.1320
SEE	$*$	$*$	0.2267	0.1371	0.1370
CP	$*$	$*$	0.968	0.968	0.968
SEEBT	–	–	0.1554	0.1377	0.1373
CPBT	–	–	0.828	0.966	0.964

Table 3.3: Comparison of complete cases, Goetghebeur and Ryan, imputation, and doubly robust estimator using the breast cancer data

	CC	GR	MI[a]	IPWDR
4+ nodes	0.71[0.3065]	0.57[0.2803]	0.60[0.2618]	0.53[0.2808]
ER-neg.	1.70[0.4861]	1.59[0.4822]	1.61[0.4794]	1.71[0.4809]

[a] $m = 10$

Chapter 4

Conclusions

In this chapter, we first summarize the results of comparison among different approaches in terms of flexibility, robustness and efficiency. Based on these comparisons, we then make some practical recommendations. Finally we will point out some directions in this area that might be of interest for future research.

4.1 Comparison

First let us find out which of the three approaches allow us to exploit important information contained in auxiliary covariates. Intuitively, the complete case approach does not make use of auxiliary covariates. Neither of the two partial likelihood approaches allows auxiliary covariates because each covariate is associated with either of the two cause-specific hazards. However, the imputation approach and the three inverse probability weighting (IPW) approaches treat one of the two causes as the cause of interest, and then posit and fit a parametric model for the conditional probability of failing from the cause of interest given that a failure has occurred. In addition, the three IPW approaches explicitly model the missing data mechanism. The strategy of fitting the two parametric models necessitates the discussion of auxiliary covariates. The results are summarized in Table 4.1.

Now let us take a look at the corresponding missing data mechanism that is needed to warrant the validity of each approach. For the complete case approach, we need to assume that cause of failure is missing completely at random. For the Goetghebeur and Ryan partial likelihood approach, we need to assume that the missingness probability may depend on time, but not on covariates. For the imputation approach, the efficient partial likelihood approach, and the three IPW approaches, we only need to assume that the missingness probability may depend on both time and covariates, but not on cause of failure that might be missing. When the missingness probability may depend on the missing data, none of the approaches are expected to work. The results are summarized in Table 4.2.

In order to assess the robustness of the approaches, we consider two cases where the missingness probability depends on both time and covariates and where both the complete case estimator and the Goetghebeur and Ryan estimator are biased. In the fist case, the parametric model for the missing data mechanism is correctly specified, but the parametric model for the distribution of the complete data is misspecified, i.e., the π model is correctly specified, but the ϱ model is misspecified. For this case, the efficient partial likelihood estimator does not perform well as confirmed by simulation. The imputation estimator is not expected to perform well in theory, but it is surprisingly good in simulation. All IPW estimators perform well as confirmed by simulation. In the second case, the π model is misspecified, but the ϱ model is correctly specified. For this case, the inverse probability weighted complete case (IPWCC) estimator does not perform well because it relies critically on the modeling of the π model, but both the imputation estimator and the efficient partial likelihood estimator do not model the missing data mechanism at all, hence are valid, and the IPW doubly robust (IPWDR) estimator and the IPW locally efficient (IPWLE) estimator are doubly robust, hence are also valid. The results are summarized in Table 4.3.

To compare efficiency, we must assume that all approaches are valid. For example, to compare the efficiency of the complete case estimator and the imputation estimator, we need the missing-completely-at-random (MCAR) assumption and the specification of the ϱ model to be correct. Because both the imputation estimator and the IPWCC estimator exploit additional information from the competing cause, they are more efficient than the complete case estimator. The Goetghebeur and Ryan partial likelihood estimator is more efficient and the efficient partial likelihood estimator is most efficient. The IPWDR estimator gains significant efficiency over the IPWCC estimator and the IPWLE estimator should be more efficient than the IPWDR estimator had the h model been correctly specified, but the efficiency gain is minimal because the true h model is not known in practice. The results are summarized in Table 4.4.

Based on the above discussion, the IPWDR estimator allows for inclusion of auxiliary covariates, is valid under the general missing at random assumption, is doubly robust against misspecification of either the π model or the ϱ model, has satisfactory efficiency performance, and has the appeal of easy implementation, therefore we recommend it to be used in practice.

4.2 Future Research

It is not clear why the imputation estimator performs so well in simulation studies when the ϱ model is misspecified. In addition, we have not discussed the situation when the missingness probability may depend on the unobserved data, and we have not investigated the sensitivity of these estimators for very small sample sizes.

Table 4.1: Inclusion of Auxiliary Covariates

CC	MI	GR	EPL	IPW
$\times$	$\surd$	$\times$	$\times$	$\surd$

Table 4.2: Missing Data Mechanism

	CC	MI	GR	EPL	IPW
MCAR	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
RMAR	$\times$	$\surd$	$\surd$	$\surd$	$\surd$
MAR	$\times$	$\surd$	$\times$	$\surd$	$\surd$
NINR	$\times$	$\times$	$\times$	$\times$	$\times$

Table 4.3: Robustness

(π, ϱ)	CC	MI	GR	EPL	IPWCC	IPWDR	IPWLE
(1,0)[a]	$\times$	$\times[\surd]$[b]	$\times$	$\times$	$\surd$	$\surd$	$\surd$
(0,1)	$\times$	$\surd$	$\times$	$\surd$	$\times$	$\surd$	$\surd$

[a]1 = correctly specified, 0 = misspecified
[b]statements inside brackets are simulation properties

Table 4.4: Efficiency

CC	MI	GR	EPL	IPWCC	IPWDR	IPWLE
poor	good	excellent	best	good	excellent	best[excellent]

Bibliography

Andersen, J., Goetghebeur, E., and Ryan, L. (1996). Missing cause of death information in the analysis of survival data. *Statistics in Medicine* **15**, pp. 1191-2201.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1997). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

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

Bickel, P. J., Klaasen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: John Hopkins University Press.

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

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

Cox, D. R., and Oakes, D. (1984). *Analysis of Survival Data.* New York: Chapman and Hall.

Cummings, F. J., Gray, R., Davis, T. E., Tormey, D. C., Harris, J. E., Falkson, G. G., and Arseneau, J. (1986). Tamoxifen versus placebo: double blind adjuvant trial in elderly women with stage II breast cancer. *National Cancer Institute Monographs* **1**, pp. 119-123.

Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika* **79**, pp. 855-857.

Fleming, T. R., and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Gill, R. D., and Johansen, S. (1990). A survey of product-integration with a view towards application in survival analysis. *Annals of Statistics* **18**, pp. 1501-1555.

Goetghebeur, E., and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82**, pp. 821-833.

Holt, J. D. (1978). Competing risk analyses with special reference to matched pair experiments. *Biometrika* **65**, pp. 159-165.

Kalbfleisch, J. D., and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Lu, K., and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* **57**, pp. 1191-1197.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, pp. 99-135.

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**, pp. 846-866.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. *Statistical Models in Epidemiology: The Environment and Clinical Trials.* Halloran, E., and Berry, D., editors. IMA Volume 116, New York: Springer-Verlag, pp. 1-92.

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

Rubin, D. B. (1996). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, pp. 473-489.

Satten, G. A., Datta, S., and Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* **93**, pp. 318-327.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Annals of Statistics* **9**, pp. 93-108.

van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Wang, N. and Robins, J. M. (1998). Large sample inference in parametric multiple imputation. *Biometrika* **85**, pp. 935-948.

Appendices

A Asymptotic Properties of Imputation Estimators

Consider the contribution from the jth imputed data set. Define the counting process

$$N_{ij}(R_i, \gamma, t) = I\{T_i \le t, D_{ij}(R_i, \gamma) = 1\}.$$

Then the log partial likelihood function can be written as

$$\log L_j(\hat{\gamma}, \beta) = \sum_{i=1}^{n} \int \left[\beta^T X_i - \log \left\{ \sum_{k=1}^{n} e^{\beta^T X_k} I(T_k \ge t) \right\} \right] dN_{ij}(R_i, \hat{\gamma}, t).$$

Let $\hat{\beta}_j$ be the MPLE, then it maximizes the above log partial likelihood, or, equivalently, solves the corresponding score equation $S_j(\hat{\gamma}, \beta) = 0$, where

$$S_j(\hat{\gamma}, \beta) = \sum_{i=1}^{n} \int \{X_i - \bar{X}(t, \beta)\} dN_{ij}(R_i, \hat{\gamma}, t),$$

$$\bar{X}(t, \beta) = \frac{\sum_{k=1}^{n} X_k e^{\beta^T X_k} I(T_k \ge t)}{\sum_{k=1}^{n} e^{\beta^T X_k} I(T_k \ge t)}.$$

The concave function arguments in Andersen and Gill (1982) can be used to establish the consistency of $\hat{\beta}_j$.

Let

$$dM_{ij}(R_i, \hat{\gamma}, t, \beta) = dN_{ij}(R_i, \hat{\gamma}, t) - \lambda(t) e^{\beta^T X_i} I(T_i \ge t) dt.$$

Then

$$S_j(\hat{\gamma}, \beta) = \sum_{i=1}^{n} \int \{X_i - \bar{X}(t, \beta)\} dM_{ij}(R_i, \hat{\gamma}, t, \beta).$$

By the WLLN, $\bar{X}(t, \beta_0) \xrightarrow{p} \mu_X(t)$. Similar to Tsiatis (1981), it can be shown that

$$n^{-1/2} \sum_{i=1}^{n} \int \{\bar{X}(t, \beta_0) - \mu_X(t)\} dM_{ij}(R_i, \hat{\gamma}, t, \beta_0) \xrightarrow{p} 0;$$

hence,

$$n^{-1/2} S_j(\hat{\gamma}, \beta_0) = n^{-1/2} \sum_{i=1}^{n} \Psi_{ij}(\hat{\gamma}, \beta_0) + o_p(1),$$

where $\Psi_{ij}(\gamma, \beta_0) = \int \{X_i - \mu_X(t)\} dM_{ij}(R_i, \gamma, t, \beta_0)$. Let $\mu_{\Psi,\beta_0}(\gamma) = E\{\Psi_{ij}(\gamma, \beta_0)\}$, define

$$H_j(\gamma, \beta_0) = \sum_{i=1}^{n} \{\Psi_{ij}(\gamma, \beta_0) - \mu_\Psi(\gamma, \beta_0)\}.$$

Then it can be shown using the theory of empirical processes (van der Vaart, 2000) that

$$n^{-1/2} \{H_j(\hat{\gamma}, \beta_0) - H_j(\gamma_0, \beta_0)\} \xrightarrow{p} 0;$$

hence,

$$n^{-1/2}S_j(\hat{\gamma}, \beta_0) = n^{-1/2}\sum_{i=1}^{n}\Psi_{ij}(\gamma_0, \beta_0) + \left\{\frac{\partial\mu_\Psi(\gamma_0, \beta_0)}{\partial\gamma^T}\right\} n^{1/2}(\hat{\gamma} - \gamma_0) + o_p(1). \tag{A.1}$$

We have shown that $\{D_{ij}(R_i, \gamma_0), W_i\}$ has the same joint distribution as (D_i, W_i); hence, $\Psi_{ij}(\gamma_0, \beta_0)$ has the same distribution as $\int\{X_i - \mu_X(t)\}dM_i(t)$, where $dM_i(t) = dN_i(t) - \lambda(t)e^{\beta_0^T X_i}I(T_i \geq t)dt$, $N_i(t) = I(T_i \leq t, D_i = 1)$. Note that $dM_i(t)$ is the increment of a martingale process; hence, $\mu_\Psi(\gamma_0, \beta_0) = 0$. In general, it can be shown that

$$\mu_\Psi(\gamma, \beta_0) = E\left[\{X - \mu_X(T)\}P(R = 0|W)\{\varrho(W, \gamma) - \varrho(W, \gamma_0)\}\right];$$

hence,

$$\frac{\partial\mu_\Psi(\gamma_0, \beta_0)}{\partial\gamma^T} = E[\{X - \mu_X(T)\}P(R = 0|W)\varrho_\gamma^T(W)]. \tag{A.2}$$

Because $\hat{\gamma}$ is the maximum likelihood estimator of γ based on complete cases,

$$n^{1/2}(\hat{\gamma} - \gamma_0) = n^{-1/2}\sum_{i=1}^{n}\phi(O_i, \gamma_0) + o_p(1), \tag{A.3}$$

where

$$\phi(O, \gamma_0) = I_\gamma^{-1}\varrho_\gamma(W)\left[\frac{RI(\Delta > 0)\{D - \varrho(W)\}}{\varrho(W)\{1 - \varrho(W)\}}\right].$$

Substituting (A.2) and (A.3) into (A.1), we see that $n^{-1/2}S_j(\hat{\gamma}, \beta_0)$ is asymptotically equivalent to

$$n^{-1/2}\sum_{i=1}^{n}\{\Psi_{ij}(\gamma_0, \beta_0) + E[\{X - \mu_X(T)\}P(R = 0|W)\varrho_\gamma^T(W)]\phi(O_i, \gamma_0)\}.$$

Because this is a normalized sum of i.i.d. mean zero random variables, asymptotic normality follows from the usual central limit theorem and the asymptotic variance is given by the variance of a single summand, or V_{SI}. Mean value expansion of $S_j(\hat{\gamma}, \hat{\beta}_j) = 0$ can then be used to prove Proposition 1.

The consistency of the multiple imputation estimator, $\hat{\beta}$, follows from the consistency of single imputation estimators. In addition, $n^{1/2}(\hat{\beta} - \beta_0) = V_S^{-1}n^{-1/2}S(\hat{\gamma}, \beta_0) + o_p(1)$, where $n^{-1/2}S(\hat{\gamma}, \beta_0)$ is asymptotically equivalent to

$$n^{-1/2}\sum_{i=1}^{n}\{\bar{\Psi}_{i\cdot}(\gamma_0, \beta_0) + E[\{X - \mu_X(T)\}P(R = 0|W)\varrho_\gamma^T(W)]\phi(O_i, \gamma_0)\}.$$

Because this is a normalized sum of i.i.d. mean zero random variables, asymptotic normality follows from the usual central limit theorem and the asymptotic variance is given by the variance of a single summand, or V_{MI}, which leads to Proposition 2.

B Semiparametric Efficiency of Partial Likelihood Estimator Appendix

Proof of Semiparametric Efficiency. The model is characterized by the $q \times 1$ parameter of interest β and the infinite dimensional nuisance parameters $\{\lambda(t), \lambda_0(t|x), p_X(x), \pi(t,x)\}$. Similar to Newey (1990), Bickel et al. (1993), and Robins et al. (1994), we consider the Hilbert space $\mathcal{H}$ of all q-dimensional mean zero and square-integrable measurable functions of the observed data $O = \{R, T, I(\Delta = 0), RI(\Delta = 1), RI(\Delta = 2), X\}$. The nuisance tangent space Λ is the linear subspace of $\mathcal{H}$ spanned by the scores for the nuisance parameters of all parametric submodels and their mean-square closure. It follows from the semiparametric theory that the solution to the estimating equation based on the efficient score is most efficient among all semiparametric estimators, where the efficient score is defined as the residual of the score vector for β after being projected onto the nuisance tangent space, $S_{\mathit{eff}} = S_\beta - \Pi(S_\beta|\Lambda)$. To establish the semiparametric efficiency, we only need to identify the score vector S_β and the nuisance tangent space Λ, carry out the projection, and verify the asymptotic equivalency of the estimating equation based on the efficient score and the estimating equation we have used to obtain the EPL estimator.

It is straightforward to show that the log likelihood for a single observation is given by

$$
\begin{aligned}
\ell(\beta, O) \quad = \quad & I(R=1)\log \pi(T,X) \\
& + I(R=0)\log\{1-\pi(T,X)\} \\
& + I(\Delta=0)\log \lambda_0(T|X) - \Lambda_0(T|X) \\
& + I(\Delta>0)\log \lambda(T) - \int \lambda(t) r_{\cdot}(t,X,\beta)Y(t)dt \\
& + I(R=1,\Delta=1)\log r_1(T,X,\beta) \\
& + I(R=1,\Delta=2)\log r_2(T,X,\beta) \\
& + I(R=0)\log r_{\cdot}(T,X,\beta) \\
& + \log p_X(X),
\end{aligned}
$$

where $\{\Lambda_\delta(t|x), \delta = 2,1,0\}$ are the cumulative cause-specific hazards.

Since the nuisance parameters are functionally independent and separate from each other in the log likelihood, the nuisance tangent space can be written as a direct sum of four orthogonal spaces,

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s} \oplus \Lambda_{3s} \oplus \Lambda_{4s},$$

where Λ_{1s} is associated with $\lambda(t)$, Λ_{2s} is associated with $\lambda_0(t|x)$, Λ_{3s} is associated with $p_X(x)$, and Λ_{4s} is associated with $\pi(t,x)$.

Let $dM_\delta(t) = dN_\delta(t) - \lambda^*_\delta(t|X)dt$ be the martingale increments for the corresponding counting processes, then standard techniques of semiparametric theory can be used to show that a typical element of Λ_{1s} is given by

$$\int a(t)dM_.(t),$$

where $M_. = M_1 + M_2 + M_u$, and $a(\cdot)$ is some arbitrary $q \times 1$ function of t.

To simplify notation, write $r_\delta(t,x) = r_\delta(t,x,\beta_0)$, then the score vector for β evaluated at the truth is given by

$$\begin{aligned} S_\beta(O) &= \int \frac{r_1'(t,X)}{r_1(t,X)} dM_1(t) \\ &+ \int \frac{r_2'(t,X)}{r_2(t,X)} dM_2(t) \\ &+ \int \frac{r_.'(t,X)}{r_.(t,X)} dM_u(t). \end{aligned}$$

Note that this is orthogonal to $\Lambda_{2s}, \Lambda_{3s}$ and Λ_{4s}. Therefore, using the projection theorem, the efficient score, derived as the residual after projecting S_β onto Λ, or in this case, Λ_{1s}, is given by

$$\begin{aligned} S_{\mathit{eff}}(O) &= \int \frac{r_1'(t,X)}{r_1(t,X)} dM_1(t) \\ &+ \int \frac{r_2'(t,X)}{r_2(t,X)} dM_2(t) \\ &+ \int \frac{r_.'(t,X)}{r_.(t,X)} dM_u(t) \\ &- \int a^*(t) dM_.(t), \end{aligned}$$

where

$$a^*(t) = \frac{E\{r_.'(t,X)Y(t)\}}{E\{r_.(t,X)Y(t)\}}.$$

The corresponding estimating equation is asymptotically equivalent to $\mathbf{U}(\beta) = 0$, so that the EPL estimator is semiparametric efficient.

C Notion of Auxiliary Covariates

First let us explain why auxiliary covariates should be introduced. For simplicity, let us only consider $\{\Delta = 2, 1\}$. In some circumstances, there exist covariates which are not of inherent interest for modeling the cause-specific hazard of interest, but which may be related to the missingness mechanism. For example, we may be able to identify some post-treatment variable which is related to the reason why the cause of death information was not collected, but we would not include it in the proportional hazards model because it may affect the causal interpretation associated with the parameters for treatment effects. Suppose that cause of failure is missing at random when the auxiliary covariates are included. For example, $P(R=1|T=t, X=x, A=a, \Delta=\delta)=e^{-a}$. In addition, assume that the auxiliary covariates are involved in the relationship between the two causes of death so that the conditional independence, $\{A \perp\!\!\!\perp \Delta|T, X\}$, does not hold. For example, A follows an exponential distribution with hazard $\lambda(a|t,x,\delta)=\delta$. Then by the LIE, we have

$$\begin{aligned} P(R=1|t,x,\delta) &= \int_0^\infty P(R=1|t,x,a,\delta) f(a|t,x,\delta) da \\ &= \int_0^\infty e^{-a}\delta e^{-\delta a} da \\ &= \delta/(1+\delta). \end{aligned}$$

Therefore, the MAR assumption will be violated if we exclude the auxiliary covariates from the data. On the other hand, if the auxiliary covariates are not related to the two causes of death so that $\{A \perp\!\!\!\perp \Delta|T, X\}$, then by the LIE, we have

$$\begin{aligned} P(R=1|t,x,\delta) &= \int P(R=1|t,x,a,\delta) f(a|t,x,\delta) da \\ &= \int P(R=1|t,x,a) f(a|t,x) da \\ &= P(R=1|t,x). \end{aligned}$$

Therefore, the MAR assumption will still hold even if the auxiliary covariates are excluded from the data.

Now let us see how the ϱ model is related to the conditional distribution of the auxiliary covariates. Using Bayes' rule, we have

$$\begin{aligned} \varrho(w) &= P(\Delta=2|t,x,a) \\ &= \frac{f(t,\Delta=2|x) f(a|t,x,\delta=2)}{f(t,\Delta=2|x) f(a|t,x,\delta=2) + f(t,\Delta=1|x) f(a|t,x,\delta=1)} \\ &= 1/\left\{1 + \frac{f(t,\Delta=1|x)}{f(t,\Delta=2|x)} \frac{f(a|t,x,\delta=1)}{f(a|t,x,\delta=2)}\right\} \\ &= 1/\left\{1 + \frac{\lambda_1(t|x)}{\lambda(t|x)} \frac{f(a|t,x,\delta=1)}{f(a|t,x,\delta=2)}\right\}, \end{aligned} \tag{C.1}$$

where the last equation follows from

$$f(t,\Delta=\delta|x) = \lambda_\delta(t|x) e^{-\{\Lambda(t|x)+\Lambda_1(t|x)\}}, \ \delta=1,2. \tag{C.2}$$

In particular, if $f(a|t,x,\delta=1)=f(a|t,x,\delta=2)$, then

$$\varrho(w) = 1/\left\{1 + \frac{\lambda_1(t|x)}{\lambda(t|x)}\right\} = \varrho(t,x).$$

The converse is also true. In fact, they are two equivalent ways to indicate the conditional independence, $\{A \perp\!\!\!\perp \Delta | T, X\}$.

Finally, let us investigate the effects of auxiliary covariates A on the relationship of the cause-specific hazard of interest with respect to the covariates of interest X. By definition,

$$\begin{aligned}
\lambda(t|x,a) &= \lim_{h\to 0} h^{-1} P(t \le T < t+h, \Delta = 2 | T \ge t, x, a) \\
&= \lim_{h\to 0} h^{-1} \frac{P(t \le T < t+h, \Delta = 2, a|x)}{P(T \ge t, a|x)} \\
&= \lim_{h\to 0} h^{-1} \frac{f(a|t \le T < t+h, x, \delta = 2) P(t \le T < t+h, \Delta = 2|x)}{f(a|T \ge t, x) P(T \ge t|x)} \\
&= \frac{f(a|t, x, \delta = 2)}{f(a|T \ge t, x)} \lim_{h\to 0} h^{-1} \frac{P(t \le T < t+h, \Delta = 2|x)}{P(T \ge t|x)} \\
&= \frac{f(a|t, x, \delta = 2)}{f(a|T \ge t, x)} \lambda(t|x). \qquad \text{(C.3)}
\end{aligned}$$

Consequently, if $\{A \perp\!\!\!\perp (T, \Delta)|X\}$, then $\lambda(t|x, a) = \lambda(t|x)$. Otherwise, $\lambda(t|x, a)$ may not even retain the proportional structure of $\lambda(t|x)$, let alone having same parameter values associated with x. For example, if $\lambda(t|x) = e^x$, $\lambda_1(t|x) = e^{2x}$, and $\lambda(a|t, x, \delta) = \delta$, then it is straightforward to show that $\lambda(t|x, a) = e^x(1 + e^x)/\{1 + 0.5e^{a+x}\}$.

D Bias of Complete Case Estimator

For the complete case estimator to be unbiased, it must satisfy

$$\frac{\lambda(t|x = 1, r = 1)}{\lambda(t|x = 0, r = 1)} = \frac{\lambda(t|x = 1)}{\lambda(t|x = 0)}. \qquad \text{(D.1)}$$

Similar to (C.3), we have that

$$\lambda(t|x, r = 1) = \frac{P(R = 1|t, x, \delta = 2)}{P(R = 1|T \ge t, x)} \lambda(t|x). \qquad \text{(D.2)}$$

By the LIE and the MAR assumption,

$$P(R = 1|t, \delta, x) = \int \pi(t, x, a) f(a|t, x, \delta) da, \ \ \delta = 1, 2. \qquad \text{(D.3)}$$

On the other hand,

$$P(R = 1|T \ge t, x) = \frac{P(R = 1, T \ge t|x)}{P(T \ge t|x)}. \qquad \text{(D.4)}$$

Denote $\Lambda_T(t|x) = \Lambda(t|x) + \Lambda_1(t|x)$, then $P(T \ge t|x) = e^{-\Lambda_T(t|x)}$. Moreover,

$$P(R = 1, T \ge t|x) = \int_t^\infty P(R = 1, u|x) du. \qquad \text{(D.5)}$$

Using Bayes' rule, we have

$$P(R=1,t|x) = \sum_{\delta=1}^{2} P(R=1|t,\delta,x) f(t,\Delta=\delta|x). \tag{D.6}$$

Now substituting (C.2) and (D.3) into (D.6), then into (D.5), then into (D.4), then into (D.2), we have that

$$\frac{\lambda(t|x,r=1)}{\lambda(t|x)} = \frac{\int \pi(t,x,a) f(a|t,x,\delta=2)da\ e^{-\Lambda_T(t|x)}}{\int_t^\infty \sum_{\delta=1}^2 \int \pi(u,x,a) f(a|u,x,\delta)da\ \lambda_\delta(u|x) e^{-\Lambda_T(u|x)} du}.$$

Denote this ratio as $\kappa(t,x)$, then compute $\phi(t) = \kappa(t,1)/\kappa(t,0)$. By (D.1), if the complete case estimator is unbiased, then $\phi(t) = 1$. For example, if cause of failure is missing completely at random, then $\pi(t,x,a) = \pi$ is a constant, hence $\kappa(t,x) = 1$, $\phi(t) = 1$. Therefore, we can plot $\phi(t)$ versus t to assess the bias of the complete case estimator based on the level of deviation of $\phi(t)$ from one.