A Class of Weighted Estimating Equations for Semiparametric Transformation Models with Missing Covariates

Yang Ning[1], Grace Yun Yi[2*] and Nancy Reid[3]

[1] *Department of Statistical Science, Cornell University, Ithaca, USA*

[2] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada*

[3] *Department of Statistical Sciences, University of Toronto, Toronto, Canada*

[*] *corresponding author: yyi@uwaterloo.ca*

April 12, 2017

Abstract: In survival analysis, covariate measurements often contain missing observations; ignoring this feature can lead to invalid inference. We propose a class of weighted estimating equations for right censored data with missing covariates under semiparametric transformation models. Time-specific and subject-specific weights are accommodated in the formulation of the weighted estimating equations. We establish unified results for estimating missingness probabilities that cover both parametric and nonparametric modeling schemes. To improve estimation efficiency, the weighted estimating equations are augmented by a new set of unbiased estimating equations. The resultant estimator has the so-called "double robustness" property and is optimal within a class of consistent estimators.

Keywords and phrases: Augmented weighted estimating equations; Missing at random; Right censoring; Survival data; Transformation models.

1. Introduction

Semiparametric transformation models provide a general but flexible framework for modeling survival data (Dabrowska and Doksum, 1988). In particular, such models include proportional hazards models (Cox, 1972) and proportional odds models (Bennett, 1983) as special cases. Due to their flexibility, semiparametric transformation models have attracted increasing attention (Cheng et al., 1995; Fine et al., 1998; Chen et al., 2002; Zeng and Lin, 2006, 2007; Chen, 2009; Chen et al., 2012; Kong et al., 2004). Although various methods have been developed for survival data analysis under such models, research gaps still remain. Typically,

existing methods are mostly developed for settings with complete observations for covariates. However, in practice, covariate measurements are often incomplete.

In the proportional hazards model, several methods to accommodate missing covariates have been proposed, including imputation (Paik and Tsai, 1997), regression calibration (Wang et al., 2001), nonparametric maximum likelihood estimation (Chen, 2002; Herring and Ibrahim, 2001; Chen and Little, 1999) and weighted estimating equations (Qi et al., 2005; Luo et al., 2009; Xu et al., 2009; Wang and Chen, 2001).

Despite the popularity of the transformation model in survival analysis, only limited discussion on the missing covariate problem is available. Chen and Little (2001) proposed a pseudo-likelihood approach to estimate the regression coefficients under certain parametric assumptions on the distribution of missing covariates. However, the estimates can be biased if the distribution of missing covariates is misspecified. In addition, their assumption of independence between the censoring indicator and covariates is restrictive. Recently, Huang and Wang (2010) proposed an inverse probability weighted estimator based on the estimating equation method developed by Chen et al. (2002). The weighted estimator in Huang and Wang (2010) suffers from efficiency loss, as the estimator in Chen et al. (2002) is inefficient.

To complement existing work on this problem, we propose a class of weighted estimating equations for right-censored data with missing covariates, under the semiparametric transformation model. Although the general theoretical framework for missing data problems has been well developed (e.g., Robins et al. (1994); Tsiatis (2007)), the application of these general results to the semiparametric transformation model is quite challenging and requires further methodological and theoretical development.

First, Robins' general framework starts from a class of full data influence functions. Because our model has a nonparametric component, it is unclear how to construct a flexible class of full data influence functions. We propose a class of estimating equations, similar in spirit to the profile score functions (Zeng and Lin, 2006), that incorporates the information contained in a set of weighted residuals. This makes our approach more efficient than those in Huang and Wang (2010) and Chen et al. (2002). In addition, we propose a Breslow-type estimator for the nonparametric component, which is different from the nonparametric maximum likelihood estimator in Zeng and Lin (2007) and Chen (2009). This hybrid approach simplifies the adjustment for the missingness effects and therefore makes our estimators easier

to implement than the EM algorithm of Zeng and Lin (2006, 2007) with simpler expressions for the variance estimates. The proposed estimating equation approach provides a better trade-off between estimation efficiency and convenience in implementation.

Second, we further extend the scope of the classical inverse probability weighted estimation methods by accommodating a class of time-specific and subject-specific weights into the weighted estimating equations. We establish unified results on the consistency and asymptotic normality of estimators, which are derived under various modeling schemes for the missingness probabilities. To cover a broad range of missing data scenarios, we explore both parametric and nonparametric modeling schemes for the missing covariate process. Estimation efficiency is investigated together with a geometric explanation which sheds light on the underlying differences among the proposed estimators.

Finally, to improve estimation efficiency, we propose fully augmented weighted estimating equations, which achieve the "double robustness" property (Robins et al., 1994) and optimality within a class of unbiased estimating equations. Although a general projection approach for doubly robust estimators has been well studied by Robins et al. (1994), the construction of such an estimator in the context of transformation models is new. Specifically, we develop estimating equations which iteratively estimate the regression parameters and nuisance functions. Thus, unlike the construction under the standard regression model (Robins et al., 1994) and the proportional hazards model (Qi et al., 2005; Luo et al., 2009; Xu et al., 2009; Wang and Chen, 2001), we need to introduce extra augmented estimating equations for the nonparametric nuisance functions to achieve the double robustness property. Due to the iterative augmentation between parametric and nonparametric components, our theoretical justification of double robustness and optimality is technically more complex than the existing work on the standard regression model and the proportional hazards model.

In Section 2 we present the notation and inference framework. In Section 3, we propose a class of simple weighted estimating equations to account for missingness in covariates. In Section 4, we establish unified results for estimators which are derived from various modeling schemes for missingness probabilities. We further elaborate on both parametric and nonparametric modeling schemes to feature the missing data process. The fully augmented weighted estimating equations are described in Section 5. To illustrate the utility of the proposed methods, in Section 6 we analyze a subcohort of data arising from the Action in

Diabetes and Vascular disease: preterAx and diamicroN-modified release Controlled Evaluation (ADVANCE) clinical trial. In Section 7, we conduct extensive empirical studies to assess the finite sample performance of our estimators. General discussion is included in the last section.

2. Notation and Framework

For subject $i = 1, ..., n$, let T_i and C_i be the failure time and censoring time, respectively and $\delta_i = I(T_i \leq C_i)$ be the censoring indicator. Define $X_i = \min(T_i, C_i)$. Let Z_i denote a p-dimensional vector of covariates for subject i. For simplicity, we assume that Z_i are time independent and C_i is independent of T_i given Z_i. Given the failure time T_i and the covariates Z_i, consider the transformation model

$$\log H(T_i) = -\beta^T Z_i + \epsilon_i, \tag{2.1}$$

where $H(\cdot)$ is an unknown increasing function with $H(0) = 0$, β is a p-dimensional parameter and ϵ_i is a random variable with a known distribution. The model reduces to the proportional hazards model (Cox, 1972) or the proportional odds model (Bennett, 1983; Pettitt, 1984), by assuming that ϵ_i follows the extreme-value distribution or the standard logistic distribution, respectively.

Using counting process notation, we write $N_i(t) = \delta_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$ to reflect information until time t for the counting process and the at risk process for subject i. Let $\lambda(\cdot)$ and $\Lambda(\cdot)$ denote the known hazard and cumulative hazard functions of $\exp(\epsilon_i)$. For a generic function $f(t)$, let $f(t_0-)$ denote the left limit of $f(t)$ with $t \to t_0$. For the model (2.1), the martingale process associated with $N_i(t)$ is given by

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\beta^T Z_i) \lambda_i(s-; \beta, H) dH(s),$$

where $\lambda_i(t; \beta, H) = \lambda\{\zeta_i(t; \beta, H)\}$ with $\zeta_i(t; \beta, H) = \int_0^t Y_i(s) \exp(\beta^T Z_i) dH(s)$. Let

$$k_i(t; \beta, H) = \int_{t+}^{\tau} \phi_i(s-; \beta, H) dM_i(s) \quad \text{and} \quad w_i(t; \beta, H) = 1 - k_i(t; \beta, H)/\lambda_i(t-; \beta, H),$$

where $\phi_i(t; \beta, H) = \dot{\lambda}\{\zeta_i(t; \beta, H)\}/\lambda\{\zeta_i(t; \beta, H)\}$, $\dot{\lambda}(t) = d\lambda(t)/dt$ and τ is the end of study time. It is noted that for the proportional hazards model, we have $\lambda_i(t; \beta, H) = 1$ and hence $w_i(t; \beta, H) = 1$.

The unknown function $H(\cdot)$ is often considered to be an infinite dimensional nuisance parameter. Denote by $dH(t)$ the jump size of H at time t. By nonparametric maximum likelihood theory, $H(t)$ can be estimated by $\widehat{H}(t) = \int_0^t d\widehat{H}(u)$, where $d\widehat{H}(u)$ is the calibrated Breslow-type estimator (Chen, 2009) given by,

$$d\widehat{H}(u) = \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n w_i(u;\beta,\widehat{H})Y_i(u)\exp(\beta^T Z_i)\lambda_i(u-;\beta,\widehat{H})}, \tag{2.2}$$

which is asymptotically equivalent to the nonparametric maximum likelihood estimator (Zeng and Lin, 2007).

The estimation of the parameter of interest β can be obtained by maximizing the profile likelihood for β, which is constructed by replacing $H(t)$ in the likelihood with $\widehat{H}(t)$. Equivalently, β can be estimated by solving the profile score function $U(\beta, \hat{H}) = 0$ (Chen, 2009, eq 8), where

$$U(\beta,\hat{H}) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left[Z_i - \frac{\sum_{j=1}^n Y_j(t)Z_j\exp(\beta^T Z_j)\{\lambda_j(t-;\beta,\widehat{H}) - k_j(t-;\beta,\widehat{H})\}}{\sum_{j=1}^n Y_j(t)\exp(\beta^T Z_j)\{\lambda_j(t-;\beta,\widehat{H}) - k_j(t-;\beta,\widehat{H})\}}\right] dN_i(t). \tag{2.3}$$

In practice, estimation of β and $H(\cdot)$ can be obtained by solving (2.3) and (2.2) iteratively, where the set of jump sizes of $H(\cdot)$, $(dH(X_1), ..., dH(X_n))$, is treated as an n-dimensional parameter. Zeng and Lin (2006) showed that the maximum likelihood estimator for β is semiparametrically efficient. However, the asymptotic variance for the estimator of β has no analytical form, and it can only be estimated by inverting the observed information matrix, which has dimension $(n+p)\times(n+p)$ if the X_i's are all different (Zeng and Lin, 2006). The computation is prohibitive for data with large sample sizes. In contrast, Chen et al. (2002) developed an estimating equation approach by setting the weighted residuals $k_i(t;\beta,H) = 0$ in (2.3) and (2.2). Compared to maximum likelihood estimation, the estimating equation approach is easier to implement and has a simple variance formula for estimators. However, this convenience is achieved at the price of losing efficiency, as information about β contained in the weighted residuals $k_i(t;\beta,H)$ is ignored.

3. Weighted Estimating Equations with Known Missingness Probabilities

3.1. Formulation of Weighted Estimating Equations

We consider settings when some covariates are missing. Let Z_i^m denote the covariates whose

values may be missing, and Z_i^c denote the covariates that are always observable, so $Z_i = (Z_i^m, Z_i^c)$. Let V_i be the missingness indicator taking the value 1 if Z_i^m is observed and 0 otherwise. Assume that data are missing at random with $P(V_i = 1 \mid X_i, Z_i^c, Z_i^m, \delta_i) = P(V_i = 1 \mid W_i) = \pi(W_i)$, where $W_i = (X_i, Z_i^c, \delta_i)$. Assume that $(X_i, Z_i^c, Z_i^m, \delta_i, V_i)$, $i = 1, ..., n$ are independent and identically distributed.

For $i = 1, ..., n$, let $D(t) = D(t, \beta)$ be a nonnegative deterministic function and $B_i(t) = B_i(t, \beta)$ be a nonnegative predictable random process with respect to the data filtration $\mathcal{F}_{t-}$, where $\mathcal{F}_t$ is the σ-field generated by $\{(N_i(u), Y_i(u)) : 0 \leq u \leq t, i = 1, ..., n\}$. To balance computational simplicity and estimation efficiency, we propose a hybrid approach to estimate β with missingness effects properly accounted for. Specifically, we first estimate the nuisance function $H(t)$ by a weighted Breslow estimator, and then estimate β based on the weighted profile score functions. The weighted Breslow estimator $\widehat{H}_W(t)$ is given by

$$d\widehat{H}_W(t) = \frac{\sum_{i=1}^n \Delta_i(t) dN_i(t)}{\sum_{i=1}^n \Delta_i(t) Y_i(t) \exp(\beta^T Z_i) \lambda_i(t-; \beta, \widehat{H}_W)}, \tag{3.1}$$

where $\Delta_i(t) = V_i B_i(t)/\pi(W_i)$. The estimator (3.1) is different from (2.2) in the following two aspects. First, the missingness probability $\pi(W_i)$ is introduced as a weight to account for the missing data effects. Second, we specify $w_i(u; \beta, \widehat{H}) = 1$ in (2.2). At the price of losing some information on estimating the nuisance function, our approach becomes easier to implement with simpler expressions for the asymptotic variance; see Theorem 3.1. To estimate β, we consider the following weighted profile estimating functions,

$$U_W(\beta, \widehat{H}_W, \pi; B, D) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \Delta_i(t) D(t) \Bigg[Z_i - \frac{\sum_{j=1}^n \Delta_j(t) Y_j(t) Z_j \exp(\beta^T Z_j) \{\lambda_j(t-; \beta, \widehat{H}_W) - k_j(t-; \beta, \widehat{H}_W)\}}{\sum_{j=1}^n \Delta_j(t) Y_j(t) \exp(\beta^T Z_j) \{\lambda_j(t-; \beta, \widehat{H}_W) - k_j(t-; \beta, \widehat{H}_W)\}} \Bigg] dN_i(t). \tag{3.2}$$

Our proposed weighted estimating equations accommodate existing methods with different choices of subject-specific weight $B_i(\cdot)$ and time-specific weight $D(\cdot)$. For instance, under the proportional hazards model, we have $\lambda_i(t; \beta, H) = 1$ and $k_i(t; \beta, H) = 0$. Therefore, setting $B_i(t) = 1$ and $D(t) = 1$ yields the inverse probability weighted estimator developed by Qi et al. (2005); letting $B_i(t) = \pi(t, Z_i^c, 1)$ and $D(t) = 1$ leads to the pseudo-partial likelihood estimator proposed by Luo et al. (2009); and the risk set weighted estimator described by Xu et al. (2009) is obtained by the choice of $B_i(t) = 1$ and $D(t) = \pi^*(t)$, where

$\pi^*(t)$ is the risk set selection probability given by the limit of expression (1) in Xu et al. (2009). For the transformation model, the inverse probability weighted estimator developed by Huang and Wang (2010) corresponds to that obtained from (3.2) with $B_i(t) = D(t) = 1$ and $k_i(t; \beta, H) = 0$. Our estimation method not only extends that of Huang and Wang (2010) and others, but more importantly, our method incorporates additional information about β by using the weighted residuals $k_i(t; \beta, H)$ in (3.2); the efficiency improvement is numerically demonstrated in Section 6.

3.2. Asymptotic Results

To highlight the key idea, in this section we temporarily treat the missingness probabilities $\pi(W_i)$ as known and focus on the estimation of β. This scenario also features the situation where missingness is created by design. Let $\hat{\beta}_W$ be the estimator of β obtained by solving $U_W(\beta, \widehat{H}_W, \pi; B, D) = 0$. Under suitable regularity conditions, we can establish asymptotic results for $\hat{\beta}_W$. These properties basically follow from the result that the estimating functions $U_W(\beta, \widehat{H}, \pi; B, D)$ are approximately unbiased. The martingale arguments of Andersen and Gill (1982) cannot be applied to show approximate unbiasedness of $U_W(\beta, \widehat{H}, \pi; B, D)$, because the missingness probabilities involve the survival outcome (X_i, δ_i), which makes the estimating functions (3.2) non-adaptive to the given data filtration. To get around this problem, we utilize empirical process theory (Kosorok, 2008).

Let β_0 and $H_0(t)$ be the true values of β and $H(t)$, respectively. For $k = 0, 1, 2$, define

$$s^{(k)}(t; \beta) = E\{B_i(t)Y_i(t)\lambda_i(t; \beta, H)Z_i^{\otimes k} \exp(\beta^T Z_i)\},$$

$$v^{(k)}(t; \beta) = E\{B_i(t)Y_i(t)\dot{\lambda}_i(t; \beta, H)Z_i^{\otimes k} \exp(2\beta^T Z_i)\},$$

and

$$e(t, \beta) = \exp\left\{\int_0^t \frac{v^{(0)}(u, \beta)}{s^{(0)}(u, \beta)} dH_0(u)\right\},$$

where $a^{\otimes 0} = 1, a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^T$. We write $s^{(k)}(t) = s^{(k)}(t; \beta_0)$, $v^{(k)}(t) = v^{(k)}(t; \beta_0)$ and $e(t) = e(t, \beta_0)$. For $k = 0, 1$, we also define $\psi^{(k)}(t; \beta_0)$ and $Q(t)$ whose forms are quite lengthy and are shown in the Supplementary Materials.

Theorem 3.1. *Under the regularity conditions (A1) – (A5) and (B1) in the Supplementary Materials, $\hat{\beta}_W$ is consistent for the parameter β, and as $n \to \infty$,*

$$n^{1/2}(\hat{\beta}_W - \beta_0) \xrightarrow{d} N(0, I_\beta^{-1} \Sigma_W I_\beta^{-1}),$$

where $\Sigma_W = E(M_i^{*\otimes 2}/\pi(W_i))$ *with*

$$M_i^* = \int_0^\tau B_i(t)\left[D(t)\left\{Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)}\right\} + \frac{Q(t)}{s^{(0)}(t)}\right]dM_i(t), \tag{3.3}$$

and

$$I_\beta = \int_0^\tau D(t)\gamma(t;\beta_0,H_0)s^{(0)}(t)dH_0(t)$$

with

$$\gamma(t;\beta_0,H_0) = \frac{s^{(2)}(t)+\psi^{(1)}(t;\beta_0)}{s^{(0)}(t)} - \frac{s^{(1)}(t)\{s^{(1)}(t)+\psi^{(0)}(t;\beta_0)\}^T}{\{s^{(0)}(t)\}^{\otimes 2}}.$$

This theorem generalizes the results by Huang and Wang (2010), who assume that $B_i(t) = D(t) = 1$ and $k_i(t;\beta,H) = 0$. In particular, we note that setting $k_i(t;\beta,H) = 0$ in (3.2) and $B_i(t) = D(t) = 1$ yields that Σ_W reduces to Σ_1 and I_β reduces to A in equation (2.7) of Huang and Wang (2010). As shown by the numerical studies in Section 6, our estimation method is more efficient than that of Huang and Wang (2010) for various types of hazard functions. A detailed discussion of the choice of $B_i(t)$ and $D(t)$ is provided in the Supplementary Materials.

In contrast to the maximum likelihood estimator with complete covariate measurements, one advantage of our estimator lies in the simplicity of the variance estimator. To see this, let

$$\widehat{M_i^*} = \int_0^\tau B_i(t)\left[D(t)\left\{Z_i - \frac{S_W^{(1)}(t;\hat\beta_W,\widehat H_W,\pi)}{S_W^{(0)}(t;\hat\beta_W,\widehat H_W,\pi)}\right\} + \frac{\widehat Q(t)}{S_W^{(0)}(t;\hat\beta_W,\widehat H_W,\pi)}\right]d\widehat{M_i}(t), \tag{3.4}$$

where

$$S_W^{(k)}(t;\beta,H,\pi) = \frac{1}{n}\sum_{i=1}^n \Delta_i(t)Y_i(t)\lambda_i(t;\beta,H)Z_i^{\otimes k}\exp(\beta^T Z_i),$$

$$\widehat{M_i}(t) = N_i(t) - \int_0^t Y_i(s)\exp(\hat\beta_W^T Z_i)\lambda_i(s-;\hat\beta_W,\widehat H_W)d\widehat H_W(s),$$

and $\widehat Q(t)$ is obtained from $Q(t)$ by replacing β_0 with $\hat\beta_W$, $H_0(t)$ with $\widehat H_W(t)$, and the expectation with the sample average, respectively. Then Σ_W can be consistently estimated by $\widehat\Sigma_W = n^{-1}\sum_{i=1}^n \widehat{M_i}^{*\otimes 2}/\pi(W_i)$. Similarly, I_β can be consistently estimated by

$$\widehat I_\beta = \int_0^\tau D(t)\gamma(t;\hat\beta_W,\widehat H_W)S_W^{(0)}(t;\hat\beta_W,\widehat H_W)d\widehat H_W(t),$$

where $\gamma(t;\hat\beta_W,\widehat H_W)$ is obtained from $\gamma(t;\beta_0,H_0)$ by replacing β_0 with $\hat\beta_W$, $H_0(t)$ with $\widehat H_W(t)$, and the expectation with the sample average, respectively. Hence, the asymptotic variance of $\hat\beta_W$ can be consistently estimated by $n^{-1}\widehat I_\beta^{-1}\widehat\Sigma_W\widehat I_\beta^{-1}$.

4. Weighted Estimating Equations with Estimated Missingness Probability

The assumption of known missingness probabilities $\pi(W_i)$ made in Section 3.2 can be restrictive for practical applications; one often needs to estimate the missingness probabilities $\pi(W_i)$. Such estimation, however, introduces additional complications. It not only requires extra care in the estimation of $\pi(W_i)$, but also makes the asymptotic results in Theorem 3.1 invalid. In this section, we address these issues. We first present unified results for estimators of β that are derived under general modeling schemes for missing data processes, and then direct our attention to two useful modeling scenarios for missing covariates.

4.1. General Results

Suppose that an estimator $\tilde{\pi}(\cdot)$ for $\pi(\cdot)$ is available. Replacing $\pi(W_i)$ in (3.1) and (3.2) with $\tilde{\pi}(W_i)$, we obtain the estimator of β, denoted by $\tilde{\beta}$, by solving the estimating equations

$$U_W(\beta, \widetilde{H}, \tilde{\pi}; B, D) = 0, \tag{4.1}$$

where

$$d\widetilde{H}(t) = \frac{\sum_{i=1}^n \widetilde{\Delta}_i(t) dN_i(t)}{\sum_{i=1}^n \widetilde{\Delta}_i(t) Y_i(t) \exp(\beta^T Z_i) \lambda_i(t-; \beta, \widetilde{H})},$$

and $\widetilde{\Delta}_i(t) = V_i B_i(t)/\tilde{\pi}(W_i)$. The following theorem establishes the asymptotic properties for $\tilde{\beta}$, under conditions on the estimator of $\pi(W_i)$.

Theorem 4.1. *Suppose that the estimator $\tilde{\pi}(\cdot)$ satisfies the following conditions:*

(D1) $\sup_w |\tilde{\pi}(w) - \pi(w)| = O_p(n^{-c})$ *for some* $c \in (1/4, 1/2]$,

(D2) there exists a p-dimensional function $m(W_i; \beta_0, H_0, \pi)$ *of* W_i, β_0, H_0 *and* π, *such that*

(i)

$$\frac{1}{n^{1/2}} \sum_{i=1}^n \frac{V_i}{\tilde{\pi}(W_i)} M_i^* = \frac{1}{n^{1/2}} \sum_{i=1}^n \frac{V_i}{\pi(W_i)} M_i^* + \frac{1}{n^{1/2}} \sum_{i=1}^n \left(1 - \frac{V_i}{\pi(W_i)}\right) m(W_i; \beta_0, H_0, \pi) + o_p(1), \tag{4.2}$$

(ii) $\widetilde{\Sigma} = \Sigma_W + I_1 - 2I_2$ *exists and is positive definite, where*

$$I_1 = E\left\{ \frac{1 - \pi(W_i)}{\pi(W_i)} m(W_i; \beta_0, H_0, \pi)^{\otimes 2} \right\}, I_2 = E\left\{ \frac{1 - \pi(W_i)}{\pi(W_i)} M_i^{*T} m(W_i; \beta_0, H_0, \pi) \right\},$$

and M_i^* *is defined in (3.3).*

Assume that the regularity conditions (A1) – (A5) and (B1) in the Supplementary Materials hold. Then (1) there exists a solution $\tilde{\beta}$ of (4.1), such that $\tilde{\beta}$ is a consistent estimator of β; and (2)

$$n^{1/2}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, I_\beta^{-1}\widetilde{\Sigma}I_\beta^{-1}), \quad as \;\; n \to \infty,$$

where I_β is defined as in Theorem 3.1.

This theorem has important implications. It offers an analogue to Theorem 3.1 but covers a more realistic scenario with unknown missingness probabilities $\pi(W_i)$. The asymptotic results in Theorem 4.1 apply to a broad scope of missing data models. The associated conditions (D1) and (D2) for an estimator $\tilde{\pi}(\cdot)$ are generally standard and can be satisfied for practical use. Condition (D1) says that the estimator $\tilde{\pi}(\cdot)$ needs to converge uniformly to $\pi(\cdot)$; the convergence rate can be slower than $n^{1/2}$. If a parametric model is assumed for π, then the convergence rate of $\tilde{\pi}$ is usually $n^{1/2}$, i.e., $c = 1/2$ in (D1). Such a case is discussed in Section 4.2 in detail. Moreover, Condition (D1) also allows $\tilde{\pi}$ to be a nonparametric estimator, which usually converges to π at a rate slower than $n^{1/2}$. However, to control the uncertainty of $\tilde{\pi}$, we cannot allow $\tilde{\pi}$ to converge slower than $n^{1/4}$. In Section 4.3, we describe a nonparametric method based on the kernel estimator that satisfies this condition. Condition (D2) offers a decomposition of the key structure, $n^{1/2}\sum_{i=1}^n V_i/\tilde{\pi}(W_i)M_i^*$, which is used for establishing asymptotic results for $\tilde{\beta}$, and it also ensures that the asymptotic covariance matrix exists and is positive definite. This decomposition allows us to study the difference between the estimator $\tilde{\pi}$ and π. If π is modeled parametrically with a finite dimensional parameter, then an explicit form for $m(W_i; \beta_0, H_0, \pi)$ can be derived using the standard Taylor series expansion. However, extra care is required to establish Condition (D2) if $\tilde{\pi}$ is a nonparametric estimator. Under suitable regularity conditions, we can establish Condition (D2) by following the proof of Theorem 21.1 in Kosorok (2008). The validity of Condition (D2) is established in Sections 4.2 and 4.3 for parametric estimators and kernel based nonparametric estimators, respectively.

In Theorem 3.1 $\hat{\beta}_W$ depends only on the complete observations. The second term on the right hand side of the equation (4.2) incorporates the information from incomplete observations. This suggests the possibility of improving the efficiency of $\hat{\beta}_W$ by estimating $\pi(W_i)$. By the form of $\widetilde{\Sigma}$ in (D2), if $I_1 - 2I_2$ is negative definite, then $\Sigma_W - \widetilde{\Sigma}$ is positive definite, suggesting that $\tilde{\beta}$ can be more efficient than $\hat{\beta}_W$. A sufficient condition for this scenario is

included in the following corollary.

Corollary 4.1. *If*

$$E\left[\frac{1-\pi(W_i)}{\pi(W_i)}m(W_i;\beta_0,H_0,\pi)\{M_i^*-m(W_i;\beta_0,H_0,\pi)\}^T\right]=0, \tag{4.3}$$

then

$$\widetilde{\Sigma}=\Sigma_W-E\left\{\frac{1-\pi(W_i)}{\pi(W_i)}m(W_i;\beta_0,H_0,\pi)^{\otimes 2}\right\}.$$

Therefore, $\tilde{\beta}$ is more efficient than $\hat{\beta}_W$.

This result is immediate from Theorem 4.1. Intuitively, this result can be viewed from a geometric point of view. Let

$$h_{1i}(\beta_0,H_0,\pi)=\{V_i/\pi(W_i)\}M_i^*. \tag{4.4}$$

For a given function $\tilde{m}(W_i;\beta_0,H_0,\pi)$ satisfying (D2) in Theorem 4.1, define

$$h_{2i}(\beta_0,H_0,\pi)=\{1-V_i/\pi(W_i)\}\tilde{m}(W_i;\beta_0,H_0,\pi). \tag{4.5}$$

Corollary 4.1 says that if we choose a function $\tilde{m}(W_i;\beta_0,H_0,\pi)$ such that the resulting $h_{2i}(\beta_0,H_0,\pi)$ is the projection of $h_{1i}(\beta_0,H_0,\pi)$, then the estimator $\tilde{\beta}$ obtained from $h_{2i}(\beta_0,H_0,\pi)$, is more efficient than $\hat{\beta}_W$ which is obtained from $h_{1i}(\beta_0,H_0,\pi)$. A random variable U_1 is defined to be the projection of random variable U_2 if $E\{(U_2-U_1)U_1\}=0$, i.e., U_2-U_1 and U_1 are orthogonal.

4.2. Parametric Modeling of Missingness Probability

To estimate the missingness probability, we might specify a parametric model, $\pi(W_i;\alpha)=\pi_i(\alpha)$, for the missing data process, where α is a finite dimensional parameter vector. Let $\ell(\alpha)=\sum_{i=1}^n V_i\log\pi_i(\alpha)+(1-V_i)\log\{1-\pi_i(\alpha)\}$ denote the resulting log likelihood. Then the score function for α is $U_\alpha(\alpha)=n^{-1}\sum_{i=1}^n U_{\alpha,i}(\alpha)$, where

$$U_{\alpha,i}(\alpha)=\frac{V_i-\pi_i(\alpha)}{\pi_i(\alpha)\{1-\pi_i(\alpha)\}}\dot{\pi}_i(\alpha),\quad\text{and}\quad\dot{\pi}_i(\alpha)=\frac{d\pi_i(\alpha)}{d\alpha}.$$

Let $\hat{\alpha}$ be the root of $U_\alpha(\alpha)=0$, and α_0 be the true value of α. Define $I_\alpha=E[\{U_{\alpha,i}(\alpha_0)\}^{\otimes 2}]$, and $I_{\alpha\beta}=E\{M_i^*\dot{\pi}_i(\alpha_0)/\pi_i(\alpha_0)\}$. An estimator of β, denoted by $\hat{\beta}_{SW}$, can be obtained by solving the estimating equations

$$U_W(\beta,\widehat{H}_{SW},\pi(\hat{\alpha});B,D)=0,$$

where

$$d\widehat{H}_{SW}(t) = \frac{\sum_{i=1}^n \Delta_i(t,\hat{\alpha})dN_i(t)}{\sum_{i=1}^n \Delta_i(t,\hat{\alpha})Y_i(t)\exp(\beta^T Z_i)\lambda_i(t-;\beta,\widehat{H}_{SW})},$$

and $\Delta_i(t,\hat{\alpha}) = V_i B_i(t)/\pi_i(\hat{\alpha})$. Asymptotic properties of $\hat{\beta}_{SW}$ are established in the following theorem.

Theorem 4.2. *Under the regularity conditions (A1) – (A5) and (B1) in the Supplementary Materials, $\hat{\beta}_{SW}$ is consistent for the parameter β, and as $n \to \infty$,*

$$n^{1/2}(\hat{\beta}_{SW} - \beta_0) \xrightarrow{d} N(0, I_\beta^{-1}\Sigma_{SW}I_\beta^{-1}),$$

where $\Sigma_{SW} = \Sigma_W - I_{\alpha\beta}^{-1}I_\alpha I_{\alpha\beta}^{-1}$.

To see the connection between Theorem 4.2 and Theorem 4.1, we note that using the Taylor series expansion, $m(W_i;\beta_0,H_0,\pi)$ in (4.2) is taken as $m(W_i;\beta_0,H_0,\pi) = I_{\alpha\beta}^{-1}I_\alpha\dot{\pi}_i(\alpha_0)/\{1-\pi_i(\alpha_0)\}$. Thus, Theorem 4.2 is a special case of Theorem 4.1. Moreover, one can verify that (4.3) in Corollary 4.1 holds in this case, suggesting that $\hat{\beta}_{SW}$ is more efficient than $\hat{\beta}_W$. The phenomenon was also observed by Xu et al. (2009) for the proportional hazards model. In general missing data problems, an interpretation of estimation efficiency based on semiparametric theory is given by Robins et al. (1994); see also Tsiatis (2007).

Finally, we note that when applying Theorem 4.2 for inference, an asymptotic variance estimate of $\hat{\beta}_{SW}$ is given by

$$\frac{1}{n}\widehat{I}_\beta^{-1}\{\widehat{\Sigma}_W - \widehat{I}_{\alpha\beta}^{-1}\widehat{I}_\alpha\widehat{I}_{\alpha\beta}^{-1}\}\widehat{I}_\beta^{-1},$$

where $\widehat{I}_{\alpha\beta} = n^{-1}\sum_{i=1}^n \widehat{M}_i^*\dot{\pi}_i(\hat{\alpha})/\pi_i(\hat{\alpha})$, and $\widehat{I}_\alpha = n^{-1}\sum_{i=1}^n\{U_{\alpha,i}(\hat{\alpha})\}^{\otimes 2}$.

4.3. Nonparametric Modeling of Missingness Probability

The validity of the weighted estimating equation approach relies on the correct specification of the missingness probability. The estimator $\hat{\beta}_{SW}$ can be inconsistent if the model $\pi(W_i;\alpha)$ in Section 4.2 is misspecified. To avoid possible model misspecification, we propose a nonparametric approach to handle the missingness probability $\pi(W_i)$. We write $W_i = (W_i^{(1)}, W_i^{(2)})$, so that $W_i^{(1)}$ is a vector of continuous variables and $W_i^{(2)}$ is a vector of discrete variables. Then $\pi(W)$ can be estimated by the kernel estimator

$$\hat{\pi}(w) = \hat{\pi}(w^{(1)},w^{(2)}) = \frac{\sum_{i=1}^n V_i I(W_i^{(2)} = w^{(2)})K_h(w^{(1)} - W_i^{(1)})}{\sum_{i=1}^n I(W_i^{(2)} = w^{(2)})K_h(w^{(1)} - W_i^{(1)})}, \tag{4.6}$$

where $K_h(\cdot) = K(\cdot/h)$, $K(\cdot)$ is a kernel function described in the Supplementary Materials and h is a smoothing parameter. Such a nonparametric estimator $\hat{\pi}(w^{(1)}, w^{(2)})$ is often feasible for the case where W_i is low dimensional.

Replacing $\pi(W_i)$ in (3.1) and (3.2) with $\hat{\pi}(W_i)$, we obtain the estimator of β, denoted by $\hat{\beta}_{NW}$, by solving the estimating equations $U_W(\beta, \widehat{H}_{NW}, \hat{\pi}; B, D) = 0$, where

$$d\widehat{H}_{NW}(t) = \frac{\sum_{i=1}^n \widehat{\Delta}_i(t) dN_i(t)}{\sum_{i=1}^n \widehat{\Delta}_i(t) Y_i(t) \exp(\beta^T Z_i) \lambda_i(t-; \beta, \widehat{H}_{NW})},$$

and $\widehat{\Delta}_i(t) = V_i B_i(t)/\hat{\pi}(W_i)$. Asymptotic properties of $\hat{\beta}_{NW}$ are established in the following theorem.

Theorem 4.3. *Under the regularity conditions (A1) – (A5), (B1) – (B3) and (C1)–(C3) in the Supplementary Materials, $\hat{\beta}_{NW}$ is consistent for the parameter β, and as $n \to \infty$,*

$$n^{1/2}(\hat{\beta}_{NW} - \beta_0) \xrightarrow{d} N\left[0, I_\beta^{-1}\left\{\Sigma_W - E\left(\frac{1-\pi(W_i)}{\pi(W_i)} M_i^{*o\otimes 2}\right)\right\} I_\beta^{-1}\right],$$

where $M_i^{*o} = E(M_i^* \mid W_i)$.

The asymptotic covariance of $n^{1/2}(\hat{\beta}_{NW} - \beta_0)$ in Theorem 4.3 can be consistently estimated by

$$\widehat{I}_\beta^{-1}\left\{\widehat{\Sigma}_W - \frac{1}{n}\sum_{i=1}^n \frac{1-\hat{\pi}(W_i)}{\hat{\pi}(W_i)} \widehat{M}_i^{*o\otimes 2}\right\} \widehat{I}_\beta^{-1},$$

where

$$\widehat{M}_i^{*o} = \frac{\sum_{j=1}^n \widehat{M}_j^* V_j I(W_j^{(2)} = W_i^{(2)}) K_h(W_i^{(1)} - W_j^{(1)})}{\sum_{j=1}^n V_j I(W_j^{(2)} = W_i^{(2)}) K_h(W_i^{(1)} - W_j^{(1)})}.$$

We comment that Theorem 4.3 is essentially a special case of Theorem 4.1. Under conditions (B1) – (B3) and (C1)–(C3) in the Supplementary Materials, we have that $\sup_w |\hat{\pi}(w) - \pi(w)| = O_p(h^r + (nh^d)^{-1/2})$, where r and d are defined in condition (B2). To meet the condition (C2), we choose a smoothing parameter $h = O(n^{-1/a})$ for some integer a with $2d < a < 2r$. Consequently, the resulting convergence rate for $\hat{\pi}(w)$ is slower than $n^{1/2}$, which makes condition (D1) true. Equation (4.2) is satisfied with $m(W_i; \beta_0, H_0, \pi) = E(M_i^* \mid W_i)$, as established in the Supplementary Materials. In addition, one can verify that (4.3) in Corollary 4.1 holds in this case. Thus, similar to $\hat{\beta}_{SW}$, $\hat{\beta}_{NW}$ is more efficient than $\hat{\beta}_W$. Moreover, for $\hat{\beta}_{NW}$ we have the following optimality result.

Corollary 4.2. *Let $\tilde{\beta}$ be an estimator obtained from any choice of $\tilde{\pi}(W_i)$ in Theorem 4.1. Then, under the regularity conditions of Theorem 4.3, we have*

$$Avar(\hat{\beta}_{NW}) \leq Avar(\tilde{\beta}),$$

where $Avar(\tilde{\beta})$ represents the asymptotic variance of an estimator $\tilde{\beta}$, and the inequality $\leq$ is the Loewner order.

Corollary 4.2 implies that for any specified $B_i(t)$ and $D(t)$, $\hat{\beta}_{NW}$ is most efficient within the class of estimators which solve equation (4.1), where $\tilde{\pi}(W_i)$ in (4.1) is any estimator of $\pi(W_i)$ satisfying the conditions in Theorem 4.1. This optimality result does not imply that $\hat{\beta}_{NW}$ is most efficient among all regular estimators. However, with the flexibility of $B_i(t)$ and $D(t)$ as well as the arbitrariness of an estimator of $\pi(\cdot)$, the optimality result of Corollary 4.2 enables us to have a better understanding of the estimator $\hat{\beta}_{NW}$. In the Supplementary Materials, we give a toy example to show several possibilities for constructing an estimator of $\pi(\cdot)$.

The optimality result of Corollary 4.2 can be visualized from a geometric point of view. Consider estimating functions, $h_{1i}(\beta_0, H_0, \pi)$ and $h_{2i}(\beta_0, H_0, \pi)$ as defined by (4.4) and (4.5), respectively. Define $h_{3i}(\beta_0, H_0, \pi) = (1 - V_i/\pi(W_i))E(M_i^* \mid W_i)$. Now we examine the relationship among those estimating functions.

Let $\mathcal{S}_u(V_i, W_i)$ denote the class of all p-dimensional functions of β and data (V_i, W_i), which are unbiased and have finite covariances. The class $\mathcal{S}_u(V_i, W_i)$ can be viewed as a Hilbert space (Small and McLeish, 2011). Consider the following subspaces of $\mathcal{S}_u(V_i, W_i)$:

$$\mathcal{S}_{u2}(V_i, W_i) = \{a(1 - V_i/\pi(W_i))\tilde{m}(W_i; \beta_0, H_0, \pi) : a \in \mathbb{R}\},$$

and

$$\begin{aligned}\mathcal{S}_{u3}(V_i, W_i) \;=\; & \{(1 - V_i/\pi(W_i))g(W_i; \beta_0, H_0, \pi) : g(W_i; \beta_0, H_0, \pi) \in \mathcal{S}_u(V_i, W_i), \\ & \text{and } g(W_i; \beta_0, H_0, \pi) \text{ is free of } V_i\}.\end{aligned}$$

As shown in Corollary 4.1, $h_{2i}(\beta_0, H_0, \pi)$ and $h_{1i}(\beta_0, H_0, \pi) - h_{2i}(\beta_0, H_0, \pi)$ are orthogonal, indicating that $h_{2i}(\beta_0, H_0, \pi)$ can be regarded as the projection of $h_{1i}(\beta_0, H_0, \pi)$ onto the space $\mathcal{S}_{u2}(V_i, W_i)$. By the proof of Corollary 4.2, we establish that $h_{3i}(\beta_0, H_0, \pi)$ can be treated as the projection of $h_{1i}(\beta_0, H_0, \pi)$ onto the space $\mathcal{S}_{u3}(V_i, W_i)$. It is noted that

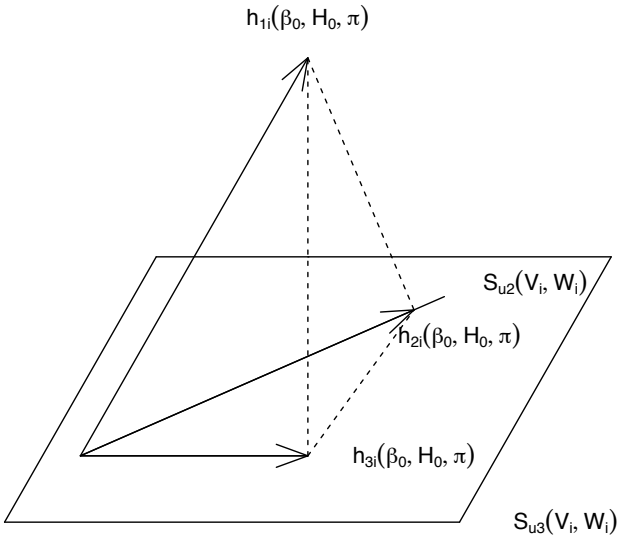

FIG 1. *Illustration of geometric relationship among* $h_{1i}(\beta_0, H_0, \pi)$, $h_{2i}(\beta_0, H_0, \pi)$, $h_{3i}(\beta_0, H_0, \pi)$, $\mathcal{S}_{u2}(V_i, W_i)$ *and* $\mathcal{S}_{u3}(V_i, W_i)$.

$\mathcal{S}_{u2}(V_i, W_i)$ is a subspace of $\mathcal{S}_{u3}(V_i, W_i)$. Therefore, $h_{3i}(\beta_0, H_0, \pi)$ has shorter "length" than $h_{2i}(\beta_0, H_0, \pi)$, as displayed in Figure 1. This suggests that $h_{3i}(\beta_0, H_0, \pi)$ has the smallest variance (in the Loewner order) within the class of unbiased estimating functions $h_{2i}(\beta_0, H_0, \pi)$ with different choices of $\tilde{m}(W_i; \beta_0, H_0, \pi)$. Following standard Taylor series expansion, the optimality property of $h_{3i}(\beta_0, H_0, \pi)$ can be transferred to the corresponding point estimator $\hat{\beta}_{NW}$.

5. Fully Augmented Weighted Estimating Equations

5.1. Formulation and Theory

The methods developed in Section 3 are easy to implement, but they may not necessarily be very efficient. On the other hand, while the estimator $\hat{\beta}_{NW}$ of Section 4 enjoys a certain optimality property, it may be numerically unstable when the sample size is small or W_i is high dimensional. To address these issues, we further develop augmented weighted estimating equations, which can typically be applied for parametric modeling of $\pi(W_i)$ for the case with high dimensional W_i. Following Robins et al. (1994), the key idea is to project the estimating equations constructed in Section 3 onto the orthogonal complement of the tangent space for the nuisance missing data process, and then construct new estimating equations by removing

redundant information. However, the construction of the estimator under the transformation model requires further augmentations of estimating equations which will be described as follows. For $k = 0, 1, 2$, let

$$\begin{aligned} S_A^{(k)}(t;\beta,H,\pi) &= \frac{1}{n}\sum_{j=1}^{n}\Bigg[\Delta_j(t)Y_j(t)Z_j^{\otimes k}\exp(\beta^T Z_j)\lambda_j(t-;\beta,H) \\ &\quad + \left(1-\frac{V_j}{\pi(W_j)}\right)Y_j(t)E\{B_j(t)Z_j^{\otimes k}\exp(\beta^T Z_j)\lambda_j(t-;\beta,H)\mid W_j\}\Bigg], \end{aligned}$$

$$\begin{aligned} R_A^{(k)}(t;\beta,H,\pi) &= \frac{1}{n}\sum_{j=1}^{n}\Bigg[\Delta_j(t)Y_j(t)Z_j^{\otimes k}\exp(\beta^T Z_j)k_j(t-;\beta,H) \\ &\quad + \left(1-\frac{V_j}{\pi(W_j)}\right)Y_j(t)E\{B_j(t)Z_j^{\otimes k}\exp(\beta^T Z_j)k_j(t-;\beta,H)\mid W_j\}\Bigg], \end{aligned}$$

and let

$$\begin{aligned} A_i^F(\beta,H,\pi;B,D) &= \left(1-\frac{V_i}{\pi(W_i)}\right)\int_0^\tau\Bigg[E\{B_i(t)D(t)Z_i dN_i(t)\mid W_i\} \\ &\quad - \frac{S_A^{(1)}(t;\beta,H,\pi)-R_A^{(1)}(t;\beta,H,\pi)}{S_A^{(0)}(t;\beta,H,\pi)-R_A^{(0)}(t;\beta,H,\pi)}E\{B_i(t)D(t)dN_i(t)\mid W_i\}\Bigg]. \end{aligned}$$

Then augmented weighted estimating equations for β are defined to be

$$\begin{aligned} U_{FA}(\beta,H,\pi;B,D) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Delta_i(t)D(t)\left\{Z_i-\frac{S_A^{(1)}(t;\beta,H,\pi)-R_A^{(1)}(t;\beta,H,\pi)}{S_A^{(0)}(t;\beta,H,\pi)-R_A^{(0)}(t;\beta,H,\pi)}\right\}dN_i(t) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}A_i^F(\beta,H,\pi;B,D). \end{aligned}$$

Due to the presence of $H(\cdot)$ in the above estimating equation, we further introduce a new set of augmented estimators $\int_0^t d\widehat{H}_A(u)$ for the nuisance function $H(t)$, where

$$d\widehat{H}_A(t) = \frac{1}{n}\frac{\sum_{i=1}^{n}[\Delta_i(t)dN_i(t)-\{V_i/\pi(W_i)-1\}E\{B_i(t)dN_i(t)\mid W_i\}]}{S_A^{(0)}(t;\beta,\widehat{H}_A,\pi)}.$$

Let $f(Z_i^m \mid W_i)$ denote the conditional distribution of Z_i^m given $W_i = (X_i, Z_i^c, \delta_i)$. If $\pi(W_i)$ and $f(Z_i^m \mid W_i)$ are known, then solving $U_{FA}(\beta, \widehat{H}_A, \pi; B, D) = 0$ gives us the augmented weighted estimator $\hat{\beta}_{FA}$. Asymptotic properties of estimator $\hat{\beta}_{FA}$ are established in the following theorem.

Theorem 5.1. *Under the regularity conditions (A1) – (A6) and (B1) in the Supplementary Materials, $\hat{\beta}_{FA}$ is consistent for the parameter β, provided either $\pi(W_i)$ or $f(Z_i^m \mid W_i)$ is correctly specified. When both $\pi(W_i)$ and $f(Z_i^m \mid W_i)$ are correctly specified, we have that as $n \to \infty$,*

$$n^{1/2}(\hat{\beta}_{FA} - \beta_0) \xrightarrow{d} N\left[0, I_\beta^{-1}\left\{\Sigma_W - E\left(\frac{1-\pi(W_i)}{\pi(W_i)} M_i^{*o\otimes 2}\right)\right\} I_\beta^{-1}\right],$$

and $Avar(\hat{\beta}_{FA}) \leq Avar(\tilde{\beta})$, where $\tilde{\beta}$ is given in Theorem 4.1.

This theorem shows that $\hat{\beta}_{FA}$ is more efficient than the estimator $\hat{\beta}_W$ when $\pi(W_i)$ and $f(Z_i^m \mid W_i)$ are correctly specified; it also says that the augmented estimator $\hat{\beta}_{FA}$ has the so-called "double robustness" property (Robins et al., 1994). That is, $\hat{\beta}_{FA}$ is still consistent even when one of the models for $\pi(W_i)$ and $f(Z_i^m \mid W_i)$ is misspecified. Following Corollary 4.2, $\hat{\beta}_{FA}$ has the same optimality as $\hat{\beta}_{NW}$, because $\hat{\beta}_{FA} = \hat{\beta}_{NW} + o_p(n^{-1/2})$ as shown in the online Supplementary Materials. Despite the asymptotic equivalence between $\hat{\beta}_{FA}$ and $\hat{\beta}_{NW}$, $\hat{\beta}_{FA}$ is more suitable for the situation with high dimensional W_i. On the other hand, $\hat{\beta}_{NW}$ can be feasible for settings with fairly small dimension of the covariates.

When $\pi(W_i)$ and $f(Z_i^m \mid W_i)$ are unknown, one may estimate them using parametric models, say $\pi(W_i;\alpha)$ and $f(Z_i^m \mid W_i;\phi)$, where α and ϕ are associated finite dimensional parameters. We adopt the same parametric model for $\pi(W_i;\alpha) = \pi_i(\alpha)$ as in Section 4, and let $\hat{\alpha}$ be the root of $U_\alpha(\alpha) = 0$, where

$$U_\alpha(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{V_i - \pi_i(\alpha)}{\pi_i(\alpha)\{1-\pi_i(\alpha)\}} \dot{\pi}_i(\alpha).$$

Following Xu et al. (2009), ϕ is estimated by $\hat{\phi}(\alpha)$, the root of $U_\phi(\alpha,\phi) = 0$, where

$$U_\phi(\alpha,\phi) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{V_i}{\pi_i(\alpha)}\frac{\partial}{\partial\phi} f(Z_i^m \mid W_i;\phi) - \frac{V_i - \pi_i(\alpha)}{\pi_i(\alpha)} E\left\{\frac{\partial}{\partial\phi} f(Z_i^m \mid W_i;\phi) \mid W_i;\phi\right\}\right].$$

Denote $\hat{\phi} = \hat{\phi}(\hat{\alpha})$. By Taylor series expansion, it can be shown that $\hat{\phi}$ is asymptotically equivalent to the estimator $\hat{\phi}(\alpha)$ for which α is assumed known. Let $\hat{\beta}_{FA}(\alpha,\phi)$ denote the root of $U_{FA}(\beta, \widehat{H}_A, \pi(\alpha), \phi) = 0$. The following theorem shows that estimation of α and ϕ has no effect on the asymptotic distribution of $\hat{\beta}_{FA}(\alpha,\phi)$.

Theorem 5.2. *Suppose the regularity conditions (A1) – (A6) and (B1) in the Supplementary Materials hold. Assume that $\pi(W_i;\alpha)$ and $f(Z_i^m \mid W_i;\phi)$ are correctly specified. Then*

$\hat{\beta}_{FA}(\alpha, \hat{\phi})$, $\hat{\beta}_{FA}(\hat{\alpha}, \phi)$ and $\hat{\beta}_{FA}(\hat{\alpha}, \hat{\phi})$ are all asymptotically equivalent to $\hat{\beta}_{FA}(\alpha, \phi)$. That is

$$\hat{\beta}_{FA}(\hat{\alpha}, \hat{\phi}) = \hat{\beta}_{FA}(\alpha, \hat{\phi}) + o_p(n^{-1/2}) = \hat{\beta}_{FA}(\hat{\alpha}, \phi) + o_p(n^{-1/2}) = \hat{\beta}_{FA}(\alpha, \phi) + o_p(n^{-1/2}).$$

Augmented weighted estimating equations proposed here are developed for semiparametric transformation models, therefore, our methods extend those used by Xu et al. (2009), Wang and Chen (2001), Qi et al. (2005), and Luo et al. (2009), which are addressed to the proportional hazards model only. In contrast to the proportional hazards model, $U_{FA}(\beta, H, \pi; B, D)$ requires estimating the nuisance function $H(t)$ in the transformation model. Hence, to attain the double robustness property, the weighted Breslow estimator $d\widehat{H}_W(t)$ needs to be augmented as well, leading to the augmented Breslow estimator $d\widehat{H}_A(t)$. Thus, one has to account for the extra uncertainty of the augmented Breslow estimator, which makes our theoretical analysis more challenging than the existing work on the proportional hazards model.

5.2. Computational Algorithm

To implement the proposed methods, we develop an iterative reweighting algorithm. To be specific, we describe the algorithm for calculating the estimator $\hat{\beta}_{SW}$; similar algorithms can be applied to calculate the estimators $\hat{\beta}_W$, $\hat{\beta}_{FA}$ and $\hat{\beta}_{FA}(\hat{\alpha}, \hat{\phi})$.

Step 1: Calculate $\hat{\alpha} = \text{argmax}_\alpha \ell(\alpha)$, and obtain the fitted missingness probabilities, $\pi_i(\hat{\alpha})$, for $i = 1, ..., n$.

Step 2: Set $d\widehat{H}^{(0)}_{SW}(t_*) = 1/n$, where t_* is an observed failure time, $\hat{\beta}^{(0)}_{SW} = 0$ and $k = 0$.

Step 3: Given $d\widehat{H}^{(k)}_{SW}$ and $\hat{\beta}^{(k)}_{SW}$, we consider the estimator $d\widehat{H}_{SW}$ given by

$$d\widehat{H}^{(k+1)}_{SW}(t) = \frac{\sum_{i=1}^n \Delta_i(t, \hat{\alpha}) dN_i(t)}{\sum_{i=1}^n \Delta_i(t, \hat{\alpha}) Y_i(t) \exp(\hat{\beta}^{(k)T}_{SW} Z_i) \lambda_i(t-; \hat{\beta}^{(k)}_{SW}, \widehat{H}^{(k)}_{SW})}.$$

Step 4: Given $d\widehat{H}^{(k+1)}_{SW}$ and $\hat{\beta}^{(k)}_{SW}$, the estimating equations $U_W(\beta, \widehat{H}^{(k+1)}_{SW}, \pi(\hat{\alpha}); B, D) = 0$ are solved to obtain $\hat{\beta}^{(k+1)}_{SW}$ with fixed $k_j(t-; \hat{\beta}^{(k)}_{SW}, \widehat{H}^{(k)}_{SW})$ in equation (3.2).

Step 5: Repeat steps 3 and 4 until convergence, and let $(\hat{\beta}_{SW}, \widehat{H}_{SW})$ denote the resulting limits.

Our algorithm is similar to the iterative algorithm considered by Chen (2009) for calculating the maximum likelihood estimator under the transformation model without missing

covariates. Thus, we expect the algorithm to be computationally more efficient than the EM algorithm of Zeng and Lin (2006), as argued by Chen (2009). To estimate the asymptotic variance of estimators, the methods by Chen (2009) and Zeng and Lin (2006, 2007) may require inversion of $(n+p) \times (n+p)$ information matrices, which can be unstable for large n or p. In contrast, our numerical experience suggests that our algorithm is computationally efficient and the estimated variance via the asymptotic variance formula is fairly stable.

6. Application to the ADVANCE Data

The Action in Diabetes and Vascular disease: preterAx and diamicroN-modified release Controlled Evaluation (ADVANCE) trial is one of the largest clinical trials to investigate diabetes-related diseases. One objective of this study is to compare intensive and standard glycemic control, applied to over eleven thousand subjects in twenty countries. Survival time, type of death, and various covariate information, such as glycated hemoglobin (HbA1c) and urinary albumin creatinine ratio (ACR), are recorded. While HbA1c is observed for all individuals, ACR contains missing values. Discarding subjects with missing ACR, Zoungas et al. (2012) fitted the proportional hazards model and showed that HbA1c and ACR are important risk factors for various outcomes, including all-cause mortality. Here, we analyze a subcohort which includes 146 observations for female smokers. Among those individuals, 23 (16%) have missing ACR values. In this subcohort, the mean follow-up time is 1697 days and the all-cause mortality rate is 13%.

To study the effect of HbA1c and ACR on all-cause mortality, we fit a sequence of transformation models

$$\log H(T_i) = -\beta_1 \mathrm{HbA1c}_i - \beta_2 \mathrm{ACR}_i + \epsilon_i, \tag{6.1}$$

where the hazard function of ϵ_i at time t is specified as $\exp(t)/\{1 + r\exp(t)\}$ with $r = 1, 1.5$ and 2.

In contrast to Huang and Wang (2010), who set $B_i(t) = D(t) = 1$ and $k_i(t; \beta, H) = 0$ in equations (3.1) and (3.2), we conduct various analyses with different weighting schemes and choices of $k_i(t; \beta, H)$. In particular, we consider the following four sets of weighting schemes:

(W1) $B_i(t) = D(t) = 1$;
(W2) $B_i(t) = 1$, $D(t) = \{\sum_{i=1}^n V_i Y_i(t)\}/\{\sum_{i=1}^n Y_i(t)\}$;
(W3) $B_i(t) = \pi(t, Z_i^c, 1)$, $D(t) = 1$;

(W4) $B_i(t) = \pi(t, Z_i^c, 1)$, $D(t) = \{\sum_{i=1}^n V_i Y_i(t)\}/\{\sum_{i=1}^n Y_i(t)\}$.

The weighting scheme (W1) corresponds to the inverse probability weighted estimating equations, and the weighting scheme (W2) corresponds to the risk set selection probability weighted estimating equations, because $\{\sum_{i=1}^n V_i Y_i(t)\}/\{\sum_{i=1}^n Y_i(t)\}$ is a consistent estimator of the selection probability given a risk set at time t (Xu et al., 2009). The weighting scheme (W3) yields the pseudo-partial estimating equations in Luo et al. (2009), while the weighting scheme (W4) can be regarded as a combination of (W2) and (W3).

In terms of the treatment of $k_i(t; \beta, H)$, we first consider a simplest method with $k_i(t; \beta, H) = 0$ in equation (3.2) (Huang and Wang, 2010). We calculate two estimators: one based on the complete-case only (Complete-A), and the other a weighted estimator with estimated missingness probabilities under a parametric model (WE-$\hat{\alpha}$-A). Secondly, with the martingale type residual $k_i(t; \beta, H) = \int_{t+}^{\tau} \phi_i(s-; \beta, H) dM_i(s)$ incorporated in (3.2), we compute estimators respectively based on the complete-case only (Complete-B), the weighted estimator with estimated missingness probabilities under a parametric model (WE-$\hat{\alpha}$-B), the weighted estimator with estimated missingness probabilities under a nonparametric model (WE-$\hat{\pi}$-B) and the fully augmented weighted estimator with estimated missingness probabilities under a parametric model (FAW-$\hat{\alpha}$-B). In particular, the logistic regression model $\text{logit}(\pi_i) = \alpha^T W_i$ is used for WE-$\hat{\alpha}$-A and WE-$\hat{\alpha}$-B, and the kernel function $K(u) = 3(1-u^2)/4$, $|u| \leq 1$ with the bandwidth $h_n = n^{-1/3}$ is employed for WE-$\hat{\pi}$-B.

The results are presented in Table 1. All the methods indicate that HbA1c and ACR are positively associated with mortality. The differences between the complete data analyses (Complete-A and Complete-B) and other weighted estimators (WE-$\hat{\alpha}$-A, WE-$\hat{\alpha}$-B, WE-$\hat{\pi}$-B and FAW-$\hat{\alpha}$-B) suggest that the weighted estimators reduce the bias incurred by using only complete cases. All the weighted estimation methods for HbA1c have similar point estimates and standard errors. In contrast, for ACR, the estimated standard error of WE-$\hat{\alpha}$-B which has $k_i(t; \beta, H) \neq 0$ and estimated missingness probability is as much as 34% smaller than that of WE-$\hat{\alpha}$-A which sets $k_i(t; \beta, H) = 0$, suggesting that our proposed method is empirically more efficient than the counterpart of Huang and Wang (2010). The agreement between WE-$\hat{\alpha}$-A and WE-$\hat{\alpha}$-B provides some support for the validity of the model assumption. In addition, the transformation model (6.1) with $r = 1$ seems to fit the data better than the models (6.1) with $r = 1.5$ and $r = 2$. When $r = 1$, the estimators with weight (W2) are most

efficient among the estimators based on using weights (W4), (W3) and (W1). In particular, standard error of the estimators WE-$\hat{\alpha}$-B, WE-$\hat{\pi}$-B and FAW-$\hat{\alpha}$-B for HbA1c with weight (W2) can be smaller than those with other weights by 27%. This suggests that choosing a proper weight other than the inverse missingness probability may help improve estimation efficiency. In order to investigate this, we turn to simulations.

7. Empirical Studies

7.1. Performance of the Proposed Methods

We conducted simulation studies to assess the finite sample performance of the proposed methods. The number of simulation replications is 1000. The failure time T_i is generated according to the transformation model (2.1) with $p = 2$. We set $\beta_0 = (\beta_{01}, \beta_{02}) = (-0.5, 0.5)$, $H_0(t) = t^2$ and the hazard function of ϵ_i at time t to be $\exp(t)/\{1 + r\exp(t)\}$ with $r = 1, 1.5$ and 2. The censoring time C_i is generated from the exponential distribution with density $\lambda_1 \exp(-\lambda_1 t)$, where λ_1 is chosen to yield 15%, 18% and 22% censoring proportions corresponding to $r = 1, 1.5$ and 2 respectively. We consider the case with two independent covariates, one observed covariate Z_i^c and one missing covariate Z_i^m, where Z_i^c is generated from the Bernoulli distribution with success probability 0.5 and Z_i^m is generated from the standard normal distribution. The missing data indicator V_i is generated from a Bernoulli distribution with probability π_i. We consider three scenarios for the missingness probability π_i. Under the first simulation scenario with $r = 1$, the missingness probability is associated with the censoring indicator, i.e., $\pi_i = 0.9\delta_i + 0.4(1 - \delta_i)$, yielding about 17% missingness proportion. Under the second simulation scenario with $r = 1.5$, the missingness probability is generated to be $\pi_i = \exp(0.8X_i)/(1 + \exp(0.8X_i))$, yielding about 28% missingness proportion, where $X_i = \min(T_i, C_i)$. Under the third simulation scenario with $r = 2$, the missingness probability is associated with the observed data (X_i, δ_i, Z_i^c), i.e., $\pi_i = \exp(3 - 0.5X_i - 0.5\delta_i - 0.5Z_i^c)/\{1 + \exp(3 - 0.5X_i - 0.5\delta_i - 0.5Z_i^c)\}$ yielding about 23% missingness proportion.

Various weighted estimators with different weighting schemes are examined. As in Section 6, we consider four weighting schemes (W1), (W2), (W3) and (W4) and two methods indexed by A and B for the treatment of $k_i(t; \beta, H)$. The weighting schemes (W1), (W2), (W3) and (W4) are described in Section 6. In terms of the treatment of $k_i(t; \beta, H)$, we first

TABLE 1

Analyses of the ADVANCE subcohort data under different transformation models ($r = 1, 1.5, 2$) and various weighting schemes

Weight	Method	model with $r = 1$				model with $r = 1.5$				model with $r = 2$			
		HbA1c		ACR		HbA1c		ACR		HbA1c		ACR	
		Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
(W1)	Complete-A	0.71	0.24	0.42	0.31	0.69	0.19	0.44	0.34	0.64	0.25	0.47	0.38
	WE-$\hat{\alpha}$-A	0.77	0.27	0.41	0.31	0.73	0.23	0.40	0.35	0.68	0.26	0.43	0.40
	Complete-B	0.73	0.27	0.34	0.22	0.80	0.20	0.32	0.21	0.89	0.24	0.32	0.25
	WE-$\hat{\alpha}$-B	0.78	0.29	0.39	0.24	0.76	0.24	0.35	0.23	0.80	0.26	0.33	0.30
	WE-$\hat{\pi}$-B	0.80	0.28	0.35	0.23	0.79	0.22	0.31	0.22	0.83	0.29	0.30	0.30
	FAW-$\hat{\alpha}$-B	0.77	0.32	0.37	0.27	0.76	0.25	0.35	0.23	0.81	0.28	0.34	0.32
(W2)	Complete-A	0.71	0.24	0.42	0.32	0.67	0.21	0.46	0.35	0.59	0.28	0.47	0.38
	WE-$\hat{\alpha}$-A	0.77	0.28	0.41	0.32	0.72	0.26	0.44	0.35	0.64	0.26	0.43	0.40
	Complete-B	0.75	0.23	0.35	0.24	0.78	0.24	0.33	0.20	0.84	0.30	0.33	0.32
	WE-$\hat{\alpha}$-B	0.79	0.24	0.37	0.24	0.74	0.26	0.36	0.23	0.78	0.28	0.35	0.23
	WE-$\hat{\pi}$-B	0.80	0.22	0.35	0.22	0.79	0.28	0.32	0.25	0.81	0.30	0.31	0.26
	FAW-$\hat{\alpha}$-B	0.77	0.24	0.38	0.24	0.75	0.27	0.36	0.23	0.78	0.31	0.34	0.24
(W3)	Complete-A	0.72	0.25	0.40	0.31	0.71	0.22	0.45	0.36	0.68	0.27	0.50	0.41
	WE-$\hat{\alpha}$-A	0.79	0.28	0.39	0.33	0.77	0.26	0.42	0.38	0.69	0.29	0.46	0.45
	Complete-B	0.73	0.29	0.32	0.24	0.83	0.23	0.32	0.23	0.86	0.25	0.34	0.27
	WE-$\hat{\alpha}$-B	0.77	0.27	0.39	0.22	0.77	0.25	0.37	0.25	0.85	0.26	0.32	0.32
	WE-$\hat{\pi}$-B	0.80	0.28	0.33	0.23	0.81	0.25	0.34	0.26	0.85	0.27	0.33	0.33
	FAW-$\hat{\alpha}$-B	0.77	0.29	0.35	0.24	0.79	0.24	0.36	0.24	0.87	0.27	0.35	0.33
(W4)	Complete-A	0.78	0.25	0.46	0.35	0.72	0.23	0.47	0.37	0.66	0.30	0.49	0.39
	WE-$\hat{\alpha}$-A	0.84	0.31	0.43	0.38	0.75	0.29	0.43	0.39	0.68	0.28	0.47	0.43
	Complete-B	0.81	0.26	0.36	0.28	0.81	0.27	0.33	0.24	0.85	0.29	0.33	0.35
	WE-$\hat{\alpha}$-B	0.85	0.25	0.41	0.26	0.79	0.27	0.36	0.25	0.80	0.29	0.38	0.25
	WE-$\hat{\pi}$-B	0.83	0.24	0.39	0.25	0.82	0.28	0.32	0.25	0.83	0.32	0.34	0.26
	FAW-$\hat{\alpha}$-B	0.79	0.26	0.41	0.25	0.82	0.27	0.36	0.25	0.82	0.34	0.38	0.24

consider Huang and Wang's type estimators (2010) with $k_i(t;\beta,H) = 0$ in equation (3.2). This leads to four estimators respectively based on the full cohort (Full-A), complete-case only (Complete-A), the weighted estimator with true missingness probabilities (WE-A) and the weighted estimator with estimated missingness probabilities under a parametric model (WE-$\hat{\alpha}$-A). Secondly, with the $k_i(t;\beta,H)$ terms incorporated in (3.2), we compute estimators respectively based on the full cohort (Full-B), complete-case only (Complete-B), the weighted estimator with true missingness probabilities (WE-B), the weighted estimator with estimated missingness probabilities under a parametric model (WE-$\hat{\alpha}$-B), the weighted estimator with estimated missingness probabilities under a nonparametric model (WE-$\hat{\pi}$-B), the fully augmented weighted estimator with true missingness probabilities (FAW-B) and the fully augmented weighted estimator with estimated missingness probabilities under a parametric model (FAW-$\hat{\alpha}$-B). As in Section 6, the same logistic model and the kernel function are employed for WE-$\hat{\alpha}$-B and WE-$\hat{\pi}$-B.

The full cohort estimates (Full-A and Full-B), and the estimates with true missingness probabilities (WE-A and WE-B) would not be available in practice, but serve as a benchmark for comparing the other methods. Tables 2 – 4 display the bias, empirical standard error, model-based standard error and 95% coverage rate for various estimators with weights (W1) and (W2). The simulation results with weights (W3) and (W4) are provided in the Supplementary Materials. The complete case analyses (Complete-A and Complete-B) yield biased estimates and inaccurate coverage probabilities, which agrees with the theory. The performance of the proposed method depends on the specification of $k_i(t;\beta,H)$. The method with a non-zero $k_i(t;\beta,H)$ may outperform the method with a zero $k_i(t;\beta,H)$, which is evident from empirical standard errors reported for the methods of Full-A versus Full-B, WE-A versus WE-B, and WE-$\hat{\alpha}$-A versus WE-$\hat{\alpha}$-B. This demonstrates that our proposed estimators are more efficient than the counterparts of Huang and Wang (2010). Among all estimators based on method B, we find that the model-based standard error of WE-$\hat{\alpha}$-B and WE-$\hat{\pi}$-B is smaller than that of WE-B, confirming our theoretical finding in Theorems 4.2 and 4.3. The augmented estimators FAW-B and FAW-$\hat{\alpha}$-B are at least as efficient as the simple weighted estimators, and have more accurate coverage probabilities than WE-$\hat{\alpha}$-B and WE-$\hat{\pi}$-B. This agrees with our theoretical findings that the augmented estimators are optimal.

Our simulation studies also reveal that the weighting scheme may have a non-ignorable effect on estimation of the regression parameters that are of primary interest. For instance, the variance of the estimator for β_2 obtained from the FAW-$\hat{\alpha}$-B method is different under the four weighting schemes (W1)–(W4). In particular, we find that the weighting schemes (W1) and (W3) produce the most efficient estimators in the first and second scenarios, respectively, and (W1) and (W3) have similar performance in the third scenario. A similar phenomenon holds for other types of estimators.

To further evaluate how the sample size may affect the finite sample performance of the proposed methods, we conduct simulation studies with two sample sizes: $n = 100$ and $n = 200$. We report the results for $n = 100$ here but defer those for $n = 200$ to the Supplementary Materials. The results for $n = 200$ confirm that the performance of the proposed methods improves significantly as the sample size increases. Model-based standard errors (MSE) are fairly close to the empirical standard errors (ESE) for the proposed estimators. The 95% coverage rates for our estimators are all varying between 93% and 95% when the missingness probability is known or estimated parametrically. As expected, the estimator WE-$\hat{\pi}$-B sometimes may be unstable due to nonparametric estimation of the missingness probability. But its 95% coverage rates still vary within a reasonable range. These results suggest that our variance estimators are fairly accurate, even for a moderate sample size.

In summary, the augmented estimator $\hat{\beta}_{FA}(\hat{\alpha}, \hat{\phi})$ (FAW-$\hat{\alpha}$-B) described in Section 5 tends to have the best finite sample performance in terms of the coverage probability. In addition, the performance of the proposed methods depends on the selection of the weighting schemes. Our numerical experience shows that the choice of the best weighting scheme, among (W1), (W2), (W3) and (W4), depends on the data generating procedure. In practice, when the selection probability is close to 0 for some subjects, the estimator with inverse probability weighting scheme $B_i(t) = D(t) = 1$ may be numerically unstable, leading to inflated standard errors. If we find that many estimated selection probabilities are close to 0, then it is advisable to use an alternative weighting scheme that can better compensate for the small selection probability. One such weighting scheme is (W3) with $B_i(t) = \pi(t, Z_i^c, 1)$ and $D(t) = 1$.

TABLE 2

Simulation results for the first scenario with $n = 100$: Bias, empirical standard error (ESE), model-based standard error (MSE) and 95% coverage rate (CR, in percent) of the proposed estimators using weights (W1) and (W2)

		β_1				β_2			
Weight	Method	Bias	ESE	MSE	CR	Bias	ESE	MSE	CR
(W1)	Full-A	-0.033	0.35	0.32	93	0.018	0.22	0.20	94
	Complete-A	-0.057	0.41	0.37	90	0.027	0.25	0.23	89
	WE-A	-0.049	0.44	0.39	92	0.021	0.29	0.25	93
	WE-$\hat{\alpha}$-A	-0.041	0.43	0.38	91	0.023	0.27	0.23	91
	Full-B	-0.024	0.33	0.30	94	0.012	0.20	0.18	94
	Complete-B	-0.061	0.39	0.34	87	0.028	0.23	0.19	90
	WE-B	-0.037	0.40	0.37	93	0.015	0.24	0.22	93
	WE-$\hat{\alpha}$-B	-0.030	0.39	0.35	91	0.014	0.23	0.20	90
	WE-$\hat{\pi}$-B	-0.031	0.39	0.33	89	0.014	0.23	0.20	90
	FAW-B	-0.029	0.39	0.37	92	0.015	0.23	0.21	93
	FAW-$\hat{\alpha}$-B	-0.030	0.40	0.37	94	0.017	0.24	0.21	92
(W2)	Full-A	-0.077	0.38	0.35	93	0.064	0.24	0.21	92
	Complete-A	-0.108	0.44	0.39	86	0.075	0.27	0.24	88
	WE-A	-0.092	0.48	0.45	92	0.065	0.32	0.29	93
	WE-$\hat{\alpha}$-A	-0.084	0.47	0.43	89	0.068	0.30	0.26	90
	Full-B	-0.054	0.37	0.33	93	0.050	0.22	0.19	94
	Complete-B	-0.094	0.41	0.36	84	0.063	0.25	0.19	88
	WE-B	-0.071	0.43	0.41	93	0.050	0.27	0.25	92
	WE-$\hat{\alpha}$-B	-0.064	0.43	0.38	90	0.051	0.26	0.22	90
	WE-$\hat{\pi}$-B	-0.068	0.43	0.37	89	0.056	0.26	0.21	89
	FAW-B	-0.070	0.42	0.39	92	0.053	0.26	0.24	93
	FAW-$\hat{\alpha}$-B	-0.072	0.43	0.39	93	0.051	0.27	0.25	95

TABLE 3

Simulation results for the second scenario with $n = 100$: Bias, empirical standard error (ESE), model-based standard error (MSE) and 95% coverage rate (CR, in percent) of proposed estimators using weights (W1) and (W2)

		β_1				β_2			
Weight	Method	Bias	ESE	MSE	CR	Bias	ESE	MSE	CR
(W1)	Full-A	-0.058	0.49	0.43	93	-0.004	0.24	0.20	92
	Complete-A	-0.074	0.59	0.50	87	-0.026	0.31	0.26	90
	WE-A	-0.052	0.57	0.52	92	0.018	0.30	0.24	91
	WE-$\hat{\alpha}$-A	-0.049	0.55	0.49	92	0.016	0.28	0.23	92
	Full-B	-0.062	0.41	0.37	94	0.012	0.21	0.18	93
	Complete-B	-0.081	0.50	0.45	89	0.032	0.25	0.22	88
	WE-B	-0.043	0.52	0.49	94	0.011	0.27	0.25	93
	WE-$\hat{\alpha}$-B	-0.032	0.48	0.42	91	0.022	0.26	0.22	92
	WE-$\hat{\pi}$-B	-0.040	0.49	0.41	90	0.016	0.27	0.22	91
	FAW-B	-0.036	0.50	0.43	92	0.018	0.27	0.22	93
	FAW-$\hat{\alpha}$-B	-0.034	0.48	0.43	93	0.016	0.26	0.22	92
(W2)	Full-A	-0.078	0.59	0.55	92	0.022	0.30	0.26	93
	Complete-A	-0.093	0.69	0.58	84	0.043	0.37	0.32	89
	WE-A	-0.064	0.68	0.63	92	0.031	0.35	0.29	90
	WE-$\hat{\alpha}$-A	-0.058	0.64	0.60	92	0.030	0.35	0.30	91
	Full-B	-0.052	0.45	0.42	94	0.011	0.26	0.24	94
	Complete-B	-0.128	0.54	0.48	81	0.033	0.30	0.27	90
	WE-B	-0.057	0.58	0.54	92	0.025	0.32	0.27	91
	WE-$\hat{\alpha}$-B	-0.050	0.55	0.50	91	0.029	0.31	0.26	91
	WE-$\hat{\pi}$-B	-0.063	0.55	0.49	90	0.027	0.32	0.25	91
	FAW-B	-0.053	0.54	0.50	92	0.024	0.30	0.26	92
	FAW-$\hat{\alpha}$-B	-0.061	0.53	0.50	93	0.023	0.30	0.26	92

TABLE 4

Simulation results for the third scenario with $n = 100$: Bias, empirical standard error (ESE), model-based standard error (MSE) and 95% coverage rate (CR, in percent) of the proposed estimators using weights (W1) and (W2)

		β_1				β_2			
Weight	Method	Bias	ESE	MSE	CR	Bias	ESE	MSE	CR
(W1)	Full-A	0.017	0.55	0.49	94	0.045	0.29	0.25	93
	Complete-A	0.080	0.62	0.52	86	0.066	0.35	0.27	88
	WE-A	0.027	0.61	0.62	91	0.068	0.40	0.36	93
	WE-$\hat{\alpha}$-A	0.030	0.64	0.60	92	0.042	0.38	0.34	92
	Full-B	0.020	0.41	0.37	93	0.050	0.23	0.21	93
	Complete-B	0.079	0.50	0.44	88	0.048	0.31	0.25	90
	WE-B	0.032	0.57	0.54	92	0.024	0.34	0.32	93
	WE-$\hat{\alpha}$-B	0.040	0.54	0.51	92	0.027	0.33	0.30	93
	WE-$\hat{\pi}$-B	0.028	0.53	0.48	91	0.030	0.33	0.28	91
	FAW-B	0.031	0.54	0.51	92	0.026	0.32	0.30	93
	FAW-$\hat{\alpha}$-B	0.029	0.52	0.51	92	0.026	0.32	0.30	93
(W2)	Full-A	0.031	0.58	0.53	93	0.040	0.31	0.27	93
	Complete-A	0.099	0.67	0.58	86	0.064	0.41	0.37	85
	WE-A	0.046	0.73	0.67	91	0.044	0.43	0.39	92
	WE-$\hat{\alpha}$-A	0.052	0.68	0.65	93	0.047	0.39	0.36	92
	Full-B	0.047	0.42	0.40	94	0.062	0.26	0.23	93
	Complete-B	-0.113	0.50	0.37	82	0.089	0.37	0.28	85
	WE-B	0.053	0.57	0.52	92	0.058	0.40	0.35	92
	WE-$\hat{\alpha}$-B	0.060	0.53	0.50	92	0.062	0.39	0.35	92
	WE-$\hat{\pi}$-B	0.071	0.53	0.48	89	0.066	0.37	0.33	90
	FAW-B	0.067	0.55	0.51	93	0.061	0.37	0.34	92
	FAW-$\hat{\alpha}$-B	0.068	0.54	0.52	93	0.060	0.37	0.34	93

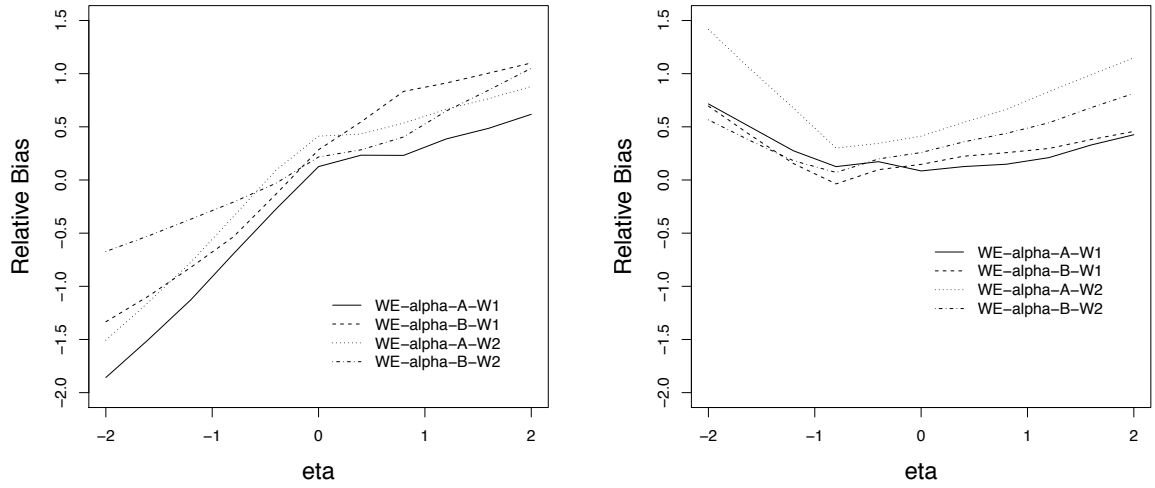

FIG 2. *Effects of misspecifying $\pi_i(\alpha)$ on the proposed estimators for β_1 and β_2 under weights (W1) and (W2). The left panel is the averaged relative bias for estimation of β_1 and the right panel is for β_2.*

7.2. Sensitivity Analysis under Misspecification of $\pi_i(\alpha)$

To evaluate the finite sample performance of the estimator $\hat{\beta}_{SW}$ (WE-$\hat{\alpha}$-A and WE-$\hat{\alpha}$-B) when the missingness probabilities are incorrectly estimated, we consider the following simulation scenario. The failure time T_i is generated according to the transformation model (2.1) with $p = 2$, $\beta_0 = (\beta_{01}, \beta_{02}) = (-0.5, 0.5)$, $H_0(t) = t^2$ and the hazard function of ϵ at time t being $\exp(t)/\{1 + \exp(t)\}$. The censoring time C_i is generated from the exponential distribution with density $0.1\exp(-0.1t)$. We use the same procedure to generate covariates. The missingness probability is given by $\pi_i = \exp(3 - 0.5X_i - 0.5\delta_i - 0.5Z_i^c - \eta\delta_i Z_i^c)/\{1 + \exp(3 - 0.5X_i - 0.5\delta_i - 0.5Z_i^c - \eta\delta_i Z_i^c)\}$, where η ranges from -2 to 2. The sample size for each simulated data is 100, and the number of simulation replications is 100.

When fitting the data, we use a misspecified logistic model $\text{logit}(\pi_i) = \alpha^T W_i$, which ignores the interaction between δ_i and Z_i^c. Unless $\eta = 0$, the estimator $\hat{\beta}_{SW}$ would be inconsistent. Figure 2 shows the relative bias of estimators using weight (W1) (WE-$\hat{\alpha}$-A-W1, WE-$\hat{\alpha}$-B-W1) and weight (W2) (WE-$\hat{\alpha}$-A-W2, WE-$\hat{\alpha}$-B-W2) averaged over 100 replications. The relative bias is defined as the ratio of the bias of the estimates to the true parameter value. All estimators have little bias when η is close to 0. As expected, the magnitude of the bias of estimators increases as η further departs from 0. The estimators of β_2 seem less sensitive with

respect to the specification of $\pi_i(\alpha)$ than those of β_1. In particular, the estimator WE-$\hat{\alpha}$-A-W2 for β_2 shows largest bias among the estimators we consider. The comparison of the relative bias of various estimators with weights (W3) and (W4) is included in the Supplementary Materials. In general, we find that the estimators are quite robust to the misspecification.

8. Discussion

Missing covariates arise commonly in survival data analysis, and it is crucial to adjust for the missingness effects in order to conduct valid inference. In this paper, we propose a class of weighted estimating equations to handle survival data with missing covariates under semiparametric transformation models. We consider weighted estimating equations, where both parametric and nonparametric methods are used for modeling the missing covariate process. We explore the doubly robust estimator derived by augmenting the weighted estimating equations. The theoretical results for the proposed methods are rigorously established. The numerical studies demonstrate satisfactory performance of our methods under various settings.

Our methods have several advantages over the existing methods. For instance, in the absence of missing covariates, our methods are easier to implement than the maximum likelihood approach proposed by Zeng and Lin (2006). On the other hand, compared to the estimating equation approach by Chen et al. (2002), our approaches utilize the information contained in the weighted residuals, and are potentially more efficient. In addition, our methods accommodate time-specific and subject-specific weights. In contrast to Huang and Wang (2010), our methods retain the information about β in the weighted residuals $k_i(t;\beta,H)$, leading to more efficient estimators. Recently, Zeng and Lin (2014) proposed a kernel based maximum likelihood estimation of semiparametric transformation models in the case-cohort study. However, such kernel based methods may not be applied when the observed covariate is high dimensional.

One of our main contributions is to incorporate the time dependent weights $B_i(t)$ and $D(t)$ into the estimating equations. Theoretically speaking, optimal $B_i(t)$ and $D(t)$ can be derived by minimizing the asymptotic variance of the resulting estimator. However, this is quite difficult to implement due to the complexity of the asymptotic variance expression. Some discussion of optimal weights is provided in the Supplementary Materials.

Throughout the paper, we adopt the counting process notation in Zeng and Lin (2006). As commented by Zeng and Lin (2006), our proposed methods and theory can be extended to accommodate time-varying covariates and recurrent events. In our development here, we assume that Z_i^m are either all observed or all missing, so that the missingness indicator V_i is a scalar. When an arbitrary nonmonotonic missing pattern for Z_i^m exists, a missingness indicator vector, say $V_i = (V_{i1}, ..., V_{ik})$ is needed. Modeling of the missing covariate process will be more notationally involved, while the same technique discussed in the paper can be employed.

In this paper, we focus on the inference on the regression parameter β and treat the function $H(\cdot)$ of little interest; this treatment is consistent with many existing methods on this topic. To reflect different roles of β and $H(\cdot)$, we employ two separate stages to handle $H(\cdot)$ and β differently. In particular, given the missingness probabilities, we first estimate the function $H(t)$ by $\hat{H}_W(t) = \int_0^t d\hat{H}_W(u)$, where $d\hat{H}_W(u)$ is defined by (3.1), and then estimate the parameter β using the weighted profile estimating equations $U_W(\beta, \hat{H}_W, \pi; B, D) = 0$, given by (3.2).

As a reviewer of an earlier draft noted, if prediction of survival time in model (2.1) is of interest, then the function $H(\cdot)$ is not a nuisance. We can modify our development to simultaneously estimate $H(\cdot)$ and β. Using the notation of Section 3.1, for any $t \in [0, \tau]$, we solve both

$$\sum_{i=1}^{n} \Delta_i(t)\{dN_i(t) - Y_i(t)\exp(\beta^T Z_i)\lambda_i(t-; \beta, H)dH(t)\} = 0, \tag{8.1}$$

and $U_W(\beta, H, \pi; B, D) = 0$ in (3.2) to estimate β and $dH(t)$ simultaneously. Then prediction of the survival function $S(t, z) = P(T \leq t | Z = z)$ with a given covariate $Z = z$ can be derived by plugging in the estimators of β and $dH(t)$. Associated asymptotic properties may be established along the lines of Eriksson et al. (2015); Martinussen and Scheike (2007) by exploiting the conditional multiplier resampling method for the construction of confidence bands for $S(t, z)$. A rigorous development is, however, beyond the scope of this article though.

Finally, we note that (8.1) defines a recursive estimating equation for $dH(t)$ which depends on the unknown parameter β while estimating function $U_W(\beta, H, \pi; B, D)$ for β involves $H(t)$. Simultaneously solving (8.1) and (3.2) with a single step is impossible, and one has to use iterations to obtain approximate solutions. Discussion on this algorithm can be found in Gorfine et al. (2006); Martinussen et al. (2011), and Eriksson et al. (2015), among others.

Acknowledgements

This research was supported by Natural Sciences and Engineering Research Council of Canada.

Supplementary Materials

The Supplementary Materials contain discussion on the optimal weights in Theorem 3.1, regularity conditions, proofs of Theorems 3.1, 4.1, 4.2, 4.3, 5.1 and 5.2 and Corollary 4.2, and additional simulation results.

References

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

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

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

Chen, H. Y. and Little, R. J. (2001). A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. *Lifetime Data Analysis* **7** 207–224.

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

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

Chen, K., Sun, L. and Tong, X. (2012). Analysis of cohort survival data with transformation model. *Statistica Sinica* **22** 489–508.

Chen, Y. H. (2009). Weighted Breslow-type and maximum likelihood estimation in semiparametric transformation models. *Biometrika* **96** 591–600.

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

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

Dabrowska, D. M. and Doksum, K. A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics* **15** 1–23.

Eriksson, F., Li, J., Scheike, T. and Zhang, M.-J. (2015). The proportional odds cumulative incidence model for competing risks. *Biometrics* **71** 687–695.

Fine, J. P., Ying, Z. and Wei, L. G. (1998). On the linear transformation model for censored data. *Biometrika* **85** 980–986.

Gorfine, M., Zucker, D. M. and Hsu, L. (2006). Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika* 735–741.

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

Huang, B. and Wang, Q. (2010). Semiparametric analysis based on weighted estimating equations for transformation models with missing covariates. *Journal of Multivariate Analysis* **101** 2078–2090.

Kong, L., Cai, J. and Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika* **91** 305–319.

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

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

Martinussen, T. and Scheike, T. H. (2007). *Dynamic Regression Models for Survival Data*. Springer, New York.

Martinussen, T., Scheike, T. H. and Zucker, D. M. (2011). The Aalen additive gamma frailty hazards model. *Biometrika* **98** 831–843.

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

Pettitt, A. N. (1984). Proportional odds models for survival data and estimates using

ranks. *Applied Statistics* **33** 169–175.

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

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

Small, C. G. and McLeish, D. L. (2011). *Hilbert Space Methods in Probability and Statistical Inference*. John Wiley & Sons, New York.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer, New York.

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

Wang, C. Y., Xie, S. X. and Prentice, R. L. (2001). Recalibration based on an approximate relative risk estimator in Cox regression with missing covariates. *Statistica Sinica* **11** 1081–1104.

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

Zeng, D. and Lin, D. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association* **109** 371–383.

Zeng, D. and Lin, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93** 627–640.

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

Zoungas, S., Chalmers, J., Ninomiya, T., Li, Q., Cooper, M. E., Colagiuri, S., Fulcher, G., de Galan, B. E., Harrap, S., Hamet, P., Heller, S., MacMahon, S., Marre, M., Poulter, N., Travert, F., Patel, A., Neal, B. and Woodward, M. (2012). Association of HbA(1c) levels with vascular complications and death in patients with type 2 diabetes: evidence of glycaemic thresholds. *Diabetologia* **55** 636–43.

Supplementary Materials of "A Class of Weighted Estimating Equations for Semiparametric Transformation Models With Missing Covariates"

Yang Ning[1], **Grace Yun Yi**[2] **and Nancy Reid**[3]

[1]*Department of Statistical Science, Cornell University, Ithaca, USA*

[2]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada*

[3]*Department of Statistical Science, University of Toronto, Toronto, Canada*

April 12, 2017

Abstract: This supplementary materials contain discussion on the optimal weights in Theorem 3.1, proofs of Theorems 4.1, 4.2, 4.3, 5.1 and 5.2 and Corollary 4.2, and additional simulation results.

1. Expressions of $\boldsymbol{Q(t)}$ and $\boldsymbol{\psi^{(k)}(t;\beta_0)}$

For $k = 0, 1$, let

$$\eta_1^{(k)}(s,t) = E\bigg\{B_i(t)Z_i^{\otimes k}\exp(3\beta_0^T Z_i)Y_i(s)\frac{\dot\lambda_i^2(s;\beta_0,H_0)}{\lambda_i(s;\beta_0,H_0)}\bigg\},$$

$$\eta_2^{(k)}(s,t) = E\{B_i(t)Z_i^{\otimes k}\exp(2\beta_0^T Z_i)Y_i(s)\dot\lambda_i(s;\beta_0,H_0)\},$$

$$\xi_1^{(k)}(s,t) = \int_s^\tau \frac{\eta_1^{(k)}(u,t)}{e(u)}dH_0(u), \quad \xi_2^{(k)}(s,t) = \int_s^\tau \frac{\eta_2^{(k)}(u,t)v^{(0)}(u)}{e(u)s^{(0)}(u)}dH_0(u),$$

$$q_1(t) = \int_t^\tau D(u)\frac{s^{(1)}(u)v^{(0)}(u) - v^{(1)}(u)s^{(0)}(u)}{s^{(0)}(u)e(u)}dH_0(u),$$

$$q_2(t) = \int_t^\tau \frac{D(u)s^{(1)}(u)}{s^{(0)}(u)}\{\xi_1^{(0)}(u,u) - \xi_2^{(0)}(u,u)\}dH_0(u),$$

$$q_3(t) = \int_0^t \frac{D(u)s^{(1)}(u)}{s^{(0)}(u)}\eta_2^{(0)}(t,u)dH_0(u),$$

$$q_4(t) = \int_0^t \frac{D(u)s^{(1)}(u)}{s^{(0)}(u)}\{\xi_1^{(0)}(t,u) - \xi_2^{(0)}(t,u)\}dH_0(u),$$

$$q_5(t) = \int_t^\tau D(u)\{\xi_1^{(1)}(u,u) - \xi_2^{(1)}(u,u)\}dH_0(u),$$

$$q_6(t) = \int_0^t D(u)\eta_2^{(1)}(t,u)dH_0(u),$$

$$q_7(t) = \int_0^t D(u)\{\xi_1^{(1)}(t,u) - \xi_2^{(1)}(t,u)\}dH_0(u),$$

$$Q(t) = e(t)\{q_1(t) + q_2(t) + q_4(t) - q_5(t) - q_7(t)\} + q_3(t) - q_6(t),$$

$$K(t) = -\frac{1}{e(t)}\int_0^t \frac{e(u)\{s^{(1)}(u) + v^{(1)}(u)H_0(u)\}}{s^{(0)}(u)}dH_0(u),$$

and

$$\begin{aligned}\psi^{(k)}(t;\beta_0) &= v^{(k+1)}(t)H_0(t) + v^{(k)}(t)K(t) \\ &\quad - \int_t^\tau \frac{\eta_2^{(k)}(u,t)\{s^{(1)}(u) + v^{(1)}(u)H_0(u) + v^{(0)}(u)K(u)\}}{s^{(0)}(u)}dH_0(u) \\ &\quad + \int_t^\tau \left\{\eta_2^{(k+1)}(u,t) + \eta_1^{(k+1)}(u,t)H_0(u) + \eta_2^{(k)}(u,t)K(u)\right\}dH_0(u).\end{aligned}$$

2. Discussion on the Optimal Weights in Theorem 3.1

In this section, we discuss the optimal choice of time-varying weights in Theorem 3.1. In particular, we consider a special transformation model: the proportional hazards model. Under this model, the hazard function of failure time T_i given the covariates Z_i satisfies

$$\lambda(t) = \lambda_0(t)\exp(\beta^T Z_i),$$

where $\lambda_0(t) = dH(t)/dt$. Recall that $X_i = \min(T_i, C_i)$ is the observed survival time, $\delta_i = I(T_i \le C_i)$ is the censoring indicator, $D(t) = D(t,\beta)$ is a nonnegative deterministic function and $B_i(t) = B_i(t,\beta)$ is a nonnegative predictable random process with respect to the data filtration $\mathcal{F}_{t-}$, where $\mathcal{F}_t$ is the σ-field generated by $\{(N_i(u), Y_i(u)) : 0 \le u \le t, i = 1, ..., n\}$, where $N_i(t) = \delta_i I(X_i \le t)$ and $Y_i(t) = I(X_i \ge t)$. Note that under the proportional hazards model, we can show that

$$\lambda_i(t-;\beta,H) = 1, \quad \text{and} \quad k_i(t-;\beta,H) = 0,$$

for any $i = 1, ..., n$. Hence, the weighted Breslow estimator $\widehat{H}_W(t)$ can be simplifies to

$$d\widehat{H}_W(t) = \frac{\sum_{i=1}^n \Delta_i(t)dN_i(t)}{\sum_{i=1}^n \Delta_i(t)Y_i(t)\exp(\beta^T Z_i)},$$

where $\Delta_i(t) = V_i B_i(t)/\pi(W_i)$ and $\pi(W_i) = P(V_i = 1 \mid W_i)$. As a result, the weighted estimating equations for β are given by,

$$U_W(\beta, \widehat{H}_W, \pi; B, D) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Delta_i(t)D(t)\left[Z_i - \frac{\sum_{j=1}^n \Delta_j(t)Y_j(t)Z_j \exp(\beta^T Z_j)}{\sum_{j=1}^n \Delta_j(t)Y_j(t)\exp(\beta^T Z_j)}\right] dN_i(t).$$

We find that the weighted estimating equations for β do not depend on $\widehat{H}_W$ under the proportional hazards model, which can be used to simplify the asymptotic variance formula of $\hat{\beta}_W$ in Theorem 3.1. Furthermore, denote

$$dM_i(t) = dN_i(t) - Y_i(t)\exp(\beta^T Z_i)dH(t), \tag{2.1}$$

$$s^{(k)}(t;\beta) = E\{B_i(t)Y_i(t)Z_i^{\otimes k}\exp(\beta^T Z_i)\}, \quad k = 0, 1, 2$$

and

$$\gamma(t;\beta_0, H_0) = \frac{s^{(2)}(t)}{s^{(0)}(t)} - \frac{\{s^{(1)}(t)\}^{\otimes 2}}{\{s^{(0)}(t)\}^2},$$

where $s^{(k)}(t) = s^{(k)}(t, \beta_0)$ for $k = 0, 1, 2$. Theorem 3.1 leads to the result

$$\sqrt{n}(\hat{\beta}_W - \beta_0) \xrightarrow{d} N(0, I_\beta^{-1}\Sigma_W I_\beta^{-1}),$$

where $\Sigma_W = E(M_i^{*\otimes 2}/\pi(W_i))$ with

$$M_i^* = \int_0^\tau B_i(t)D(t)\left\{Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)}\right\}dM_i(t), \tag{2.2}$$

and

$$I_\beta = \int_0^\tau D(t)\gamma(t;\beta_0, H_0)s^{(0)}(t)dH_0(t). \tag{2.3}$$

When β is a scalar, one can choose $D(t)$ and $B_i(t)$ such that Σ_W/I_β^2 is minimized. To illustrate the usage of the time-varying weights, we assume that $B_i(t)$ is identical for all subjects. In this case, $B_i(t)D(t)$ can be treated as a new time-varying weight function. Without loss of generality, we assume that $B_i(t) = 1$. We now minimize the asymptotic variance Σ_W/I_β^2 with respect to $D(t)$.

Define the following functionals

$$F_1(D) = \int_0^\tau D(t)A_1(t, Z)dM(t), \quad \text{where} \quad A_1(t, Z) = Z - \frac{s^{(1)}(t)}{s^{(0)}(t)}.$$

and

$$F_2(D) = \int_0^\tau D(t)A_2(t)dH_0(t), \quad \text{where} \quad A_2(t) = \gamma(t;\beta_0,H_0)s^{(0)}(t).$$

Let $D_\alpha(t) = D(t)+\alpha u(t)$ denote a curve passing through $D(t)$ indexed by a scalar parameter α with direction determined by a function $u(t)$. Then, we take derivative of Σ_W/I_β^2 with respect to α and evaluate it at $\alpha = 0$,

$$\begin{aligned}\frac{d}{d\alpha}\left(\Sigma_W/I_\beta^2\right) &= \frac{d}{d\alpha}\left[\frac{E\{F_1^2(D_\alpha)/\pi(W)\}}{F_2^2(D_\alpha)}\right] \\ &= \frac{2F_3(D,u)F_2(D) - 2F_3(D,D)F_2(u)}{F_2^3(D)},\end{aligned}$$

where

$$F_3(D,u) = E\left\{\frac{1}{\pi(W)}F_1(D)F_1(u)\right\}.$$

Hence, a necessary condition for the optimality of $D(t)$ is that $D(t)$ is the solution of the following integral equation

$$F_3(D,u)F_2(D) = F_3(D,D)F_2(u), \tag{2.4}$$

for any $u(t)$.

However, the integral equation (2.4) does not have an explicit solution; even the existence and uniqueness of the solution are not automatically guaranteed. To provide further insight into the optimal weights, we consider the proportional hazards model with discrete time. For simplicity, assume that failure and censoring are only observed at time $t = 1$ and $t = 2$. Denote $\Delta H(t) = H(t) - H(t-1)$, where $t = 1, 2$. Recall that $H(0) = 0$ by definition. Then equations (2.2) and (2.3) reduce to

$$I_\beta = D(1)J_1 + D(2)J_2, \quad \text{where} \quad J_t = A_2(t)\Delta H(t)$$

and

$$M_i^* = D(1)A_1(1,Z_i)\Delta M_i(1) + D(2)A_1(2,Z_i)\Delta M_i(2),$$

where,

$$\Delta M_i(t) = \Delta N_i(t) - Y_i(t)\exp(\beta^T Z_i)\Delta H(t),$$

and $\Delta N_i(t) = N_i(t) - N_i(t-1)$. Therefore, Σ_W is given by

$$\begin{aligned}\Sigma_W &= E\left[\frac{1}{\pi(W_i)}\{D(1)A_1(1,Z_i)\Delta M_i(1) + D(2)A_1(2,Z_i)\Delta M_i(2)\}^2\right] \\ &= D(1)^2C_1 + D(2)^2C_2 + 2D(1)D(2)C_{12},\end{aligned}$$

where

$$C_t = E\left[\frac{1}{\pi(W_i)}\{A_1(t, Z_i)\Delta M_i(t)\}^2\right], \quad t = 1, 2,$$

$$C_{12} = E\left[\frac{1}{\pi(W_i)}A_1(1, Z_i)\Delta M_i(1)A_1(2, Z_i)\Delta M_i(2)\right].$$

Our goal is to find $D(1)$ and $D(2)$ such that

$$\frac{\Sigma_W}{I_\beta^2} = \frac{D(1)^2C_1 + D(2)^2C_2 + 2D(1)D(2)C_{12}}{(D(1)J_1 + D(2)J_2)^2}, \tag{2.5}$$

is minimized. Note that if we inflate $D(1)$ and $D(2)$ by the same positive constant, (2.5) remains identical. Without loss of generality, assume that $D(2) = 1$. The optimal weight $D(1)$ is given by $\arg\min_{x\geq 0} f(x)$, where

$$f(x) = \frac{x^2C_1 + 2C_{12}x + C_2}{(J_1x + J_2)^2}.$$

Simple algebra shows that

$$f'(x) = \frac{2(C_1J_2 - C_{12}J_1)x - 2(C_2J_1 - C_{12}J_2)}{(J_1x + J_2)^3}.$$

Hence, we consider the following five situations, according to the sign of $C_{12}J_2 - C_2J_1$ and $C_1J_2 - C_{12}J_1$.

(1) If $C_1J_2 - C_{12}J_1 > 0$ and $C_2J_1 - C_{12}J_2 \geq 0$, then $f'(x) < 0$ for $0 < x < \frac{C_2J_1 - C_{12}J_2}{C_1J_2 - C_{12}J_1}$ and $f'(x) > 0$ for $x > \frac{C_2J_1 - C_{12}J_2}{C_1J_2 - C_{12}J_1}$. Hence, $f(x)$ has a unique minimizer at $x = \frac{C_2J_1 - C_{12}J_2}{C_1J_2 - C_{12}J_1}$.

(2) If $C_1J_2 - C_{12}J_1 < 0$ and $C_2J_1 - C_{12}J_2 \leq 0$, then $f'(x) > 0$ for $0 < x < \frac{C_2J_1 - C_{12}J_2}{C_1J_2 - C_{12}J_1}$ and $f'(x) < 0$ for $x > \frac{C_2J_1 - C_{12}J_2}{C_1J_2 - C_{12}J_1}$. Hence, the minimizer of $f(x)$ is either $x = 0$ or $x = +\infty$.

(3) If $C_1J_2 - C_{12}J_1 \geq 0$ and $C_2J_1 - C_{12}J_2 < 0$, then $f'(x) > 0$ for $x \geq 0$. Hence, the minimizer of $f(x)$ is $x = 0$.

(4) If $C_1J_2 - C_{12}J_1 \leq 0$ and $C_2J_1 - C_{12}J_2 > 0$, then $f'(x) < 0$ for $x \geq 0$. Hence, the minimizer of $f(x)$ is $x = +\infty$.

(5) If $C_1J_2 - C_{12}J_1 = 0$ and $C_2J_1 - C_{12}J_2 = 0$, then $f'(x) = 0$ for $x \geq 0$. Hence, $f(x)$ is a constant function.

For instance, in situation (1), the optimal weight given by $D(1) = \frac{C_2J_1 - C_{12}J_2}{C_1J_2 - C_{12}J_1}$ and $D(2) = 1$ exists and is unique. Note that, if the covariates are missing completely at random (MCAR), the missingness probability $\pi(W_i) = \pi$ is a positive constant. By the property of the martingale, we obtain $C_{12} = 0$, $\pi C_1 = J_1$ and $\pi C_2 = J_2$. From our derivation, the optimal weight

under MCAR is given by $D(1) = D(2) = 1$. However, if the covariates are missing at random (MAR), typically we have $C_2J_1 - C_{12}J_2 \neq C_1J_2 - C_{12}J_1$, and therefore the optimal $D(1)$ is not 1. Recall that in this situation the inverse probability weighting scheme corresponds to $D(1) = 1$. This suggests that the optimal time-varying weight is not the inverse probability weight.

For the proportional hazards model with continuous failure time, in principle, the cumulative hazard function can be approximated by a discrete function taking values at the observed failure time. One may adopt the similar procedure to find the optimal weights $D(t_i)$ at each time point t_i, where $i = 1, ..., n$. However, the optimization with respect to $(D(t_1), ..., D(t_n))$ may be computationally intractable for large n. In the simulation studies, instead of deriving the optimal weights, we evaluate the efficiency of the estimators under many commonly used weight functions proposed in the literature.

3. Toy Example for Corollary 4.2

In this section, we present a toy example to illustrate how the estimation efficiency may be affected by different estimators of $\pi(\cdot)$. Assume that W_i is partitioned as $(W_{1i}, ..., W_{Ji})$, and the true missing data process π_i only depends on the first covariate of W_i, i.e., $\pi_i = \pi(W_{1i})$. In this setting, there exist many possibilities for constructing the nonparametric estimator of π_i. For $j = 1, ..., J$, denote $\overline{W}_{ji} = (W_{1i}, ..., W_{ji})$ and assume that $\overline{W}_{ji} = (\overline{W}_{ji}^{(1)}, \overline{W}_{ji}^{(2)})$, where $\overline{W}_{ji}^{(1)}$ is a vector of continuous variables and $\overline{W}_{ji}^{(2)}$ is a vector of discrete variables. Consider the following kernel estimators based on $\overline{W}_{ji}$,

$$\hat{\pi}^j(\bar{w}_j) = \hat{\pi}^j(\bar{w}_j^{(1)}, \bar{w}_j^{(2)}) = \frac{\sum_{i=1}^n V_i I(\overline{W}_{ji}^{(2)} = \bar{w}_j^{(2)}) K_h(\bar{w}_j^{(1)} - \overline{W}_{ji}^{(1)})}{\sum_{i=1}^n I(\overline{W}_{ji}^{(2)} = \bar{w}_j^{(2)}) K_h(\bar{w}_j^{(1)} - \overline{W}_{ji}^{(1)})}, \tag{3.1}$$

where $\bar{w}_j = (w_1, ...w_j)$. Then π_i can be estimated by a sequence of estimators $\hat{\pi}^j(\overline{W}_{ji})$, for $j = 1, ..., J$.

The resulting estimator of β with the kernel estimator $\hat{\pi}^j(\overline{W}_{ji})$ is denoted by $\hat{\beta}_{NW}^{(j)}$. Note that $\hat{\beta}_{NW}^{(J)}$ corresponds to the estimator $\hat{\beta}_{NW}$ in Theorem 4.3. We can show that for $\hat{\beta}_{NW}^{(j)}$, condition (D2) in Theorem 4.1 is satisfied with $m(W_i; \beta_0, H_0, \pi) = E(M_i^* \mid \overline{W}_{ji})$. Therefore, as a corollary of Theorem 4.1, we have

$$\sqrt{n}(\hat{\beta}_{NW}^{(j)} - \beta_0) \xrightarrow{d} N\left[0, I_\beta^{-1}\left\{\Sigma_W - E\left(\frac{1-\pi_i}{\pi_i} M_i^{*oj\otimes 2}\right)\right\} I_\beta^{-1}\right],$$

where $M_i^{*oj} = E(M_i^* \mid \overline{W}_{ji})$. Similar to Corollary 4.2, we have $Avar(\hat{\beta}_{NW}^{(J)}) \leq Avar(\hat{\beta}_{NW}^{(J-1)}) \leq ... \leq Avar(\hat{\beta}_{NW}^{(1)})$.

This property implies that the estimator of β may gain efficiency by incorporating more variables in the estimation of the missingness probability, even if these covariates are not associated with the missing data process. Such a phenomenon is also observed by Qi et al. (2005) for the proportional hazards model. This property also shows that $\hat{\beta}_{NW}^{(J)}$ is optimal among the class of estimators $\{\hat{\beta}_{NW}^{(j)}, j = 1, ..., J\}$.

4. Regularity Conditions

For the survival process, we require the following regularity conditions.

(A1) For $i = 1, ..., n$, $P(Y_i(\tau) = 1) > 0$, and $H_0(\tau) \leq \infty$.

(A2) For $i = 1, ..., n$, the covariates Z_i are bounded.

(A3) For $i = 1, ..., n$, $B_i(t)$ and $D(t)$ are predictable processes with bounded variation and $B_i(t) > \epsilon_2$, for some $\epsilon_2 > 0$.

(A4) The quantities $s^{(k)}(t;\beta)$, $v^{(k)}(t;\beta)$, $\eta_1^{(k)}(s,t)$, and $\eta_2^{(k)}(s,t)$ exist. The matrices I_β and Σ_W are positive definite.

(A5) The matrices $I_{\alpha\beta}$, I_α and Σ_{SW} are positive definite.

(A6) The matrices $\text{var}[R_i/\pi(W_i)M_i^* - \{V_i - \pi(W_i)\}/\pi(W_i)E(M_i^* \mid W_i)]$ and $I_\phi = E\{\frac{\partial^2}{\partial\phi^2} f(Z_i^m \mid W_i;\phi)\}$ are positive definite.

For the missing data process, we assume the following regularity conditions.

(B1) There exists $\epsilon_1 > 0$ such that the missingness probability satisfies $\inf_w \pi(w) \geq \epsilon_1$.

(B2) There exists an integer $r > d$ so that the missingness probability $\pi(w^{(1)}, w^{(2)})$ has rth order continuous and bounded partial derivatives with respect to $w^{(1)}$ almost surely, where $d = \dim(w^{(1)})$.

(B3) The probability density function $f(w)$ of W has rth order continuous and bounded partial derivatives with respect to $w^{(1)}$ almost surely, and $0 < \inf_w f(w) \leq \sup_w f(w) < \infty$.

For using the kernel smoothing method, we consider the following regularity conditions.

(C1) The kernel function $K(\cdot)$ is a rth order kernel function with bounded support.

(C2) $nh^{2d} \to \infty$ and $nh^{2r} \to 0$, as $n \to \infty$.

(C3) The conditional expectation $E(M_i^* \mid W_i)$ has rth order continuous and bounded partial derivatives with respect to $W_i^{(1)}$.

Conditions (A1) and (A2) are common for survival models (Huang and Wang, 2010; Chen et al., 2002; Andersen and Gill, 1982). Condition (A3) is used to show the consistency of the estimate of $H(t)$. Conditions (A4) – (A6) are assumed to guarantee the positive definiteness of the asymptotic variances in Theorems 3.1 – 5.2. Condition (B1) is used to avoid the situation that the denominator in the weighted estimating equations is 0. Conditions (B1) – (B3) and (C1) – (C3) are typical for the kernel smoothing method in survival models. Similar regularity conditions are adopted by Qi et al. (2005) for the proportional hazards model and by Huang and Wang (2010) for the transformation model.

5. Proofs

Proof of Theorem 3.1

The proof consists of six parts which are sketched as follows. For simplicity, we assume $B_i(t)$ and $D(t)$ are independent of β. The arguments as in Appendix A1 of Xu et al. (2009) can be applied to show that the same asymptotic results still hold if $B_i(t,\beta)$ and $D(t,\beta)$ are consistently estimated by $B_i(t,\hat{\beta}_W)$ and $D(t,\hat{\beta}_W)$, respectively.

Part 1: Consistency of $\widehat{H}_W$

Define a metric for nondecreasing functions H_1 and H_2 on $[0,\tau]$ as

$$d(H_1,H_2) = \sup_{t\in[0,\tau]} |H_1(t) - H_2(t)|.$$

Let $\lambda_i(H) = \lambda_i(s-;\beta,H)$, $k_i(H) = k_i(s-;\beta,H)$ and

$$A(H)(t) \;=\; \frac{1}{n}\sum_{i=1}^{n}\int_0^t \Delta_i(s)\{dN_i(s) - Y_i(s)\exp(\beta^T Z_i)\lambda_i(s-;\beta,H)dH(s)\}.$$

For an arbitrary but fixed $\epsilon > 0$, consider nondecreasing functions $H_1(t)$ and $H_2(t)$ with $H_1(0) = H_2(0) = 0$ such that $d(H_1,H_2) > \epsilon$. Without loss of generality, there exists $\tilde{\tau} > 0$ such that $H_1(t) \geq H_2(t)$ for any $t \in [0,\tilde{\tau}]$, and for some $t \in [0,\tilde{\tau}]$, $H_1(t) > H_2(t)$. Since $H_1(0) = H_2(0)$, there must exist $\tilde{t}$ such that $dH_1(t) \geq dH_2(t)$, for any $t \in [0,\tilde{t}]$, and $\epsilon_* =$

$H_1(\tilde{t}) - H_2(\tilde{t}) > 0$. By the regularity conditions (A1) – (A3), we have

$$\begin{aligned}
&\sup_{t\in[0,\tau]} \mid A(H_1)(t) - A(H_2)(t) \mid \\
= \;& \frac{1}{n} \sup_{t\in[0,\tau]} \left| \sum_{i=1}^{n} \int_0^t \Delta_i(s) Y_i(s) \exp(\beta^T Z_i) \{\lambda_i(s-;\beta,H_1) dH_1(s) - \lambda_i(s-;\beta,H_2) dH_2(s)\} \right| \\
\geq \;& \frac{1}{n} \left| \sum_{i=1}^{n} \frac{V_i \epsilon_2}{\pi(W_i)} \int_0^{\tilde{t}} Y_i(s) \exp(\beta^T Z_i) \{\lambda_i(s-;\beta,H_1) dH_1(s) - \lambda_i(s-;\beta,H_2) dH_2(s)\} \right| \\
\geq \;& \frac{1}{n} \left| \sum_{i=1}^{n} \frac{V_i \epsilon_2}{\pi(W_i)} \int_{H_2(\tilde{t}\wedge X_i)}^{H_1(\tilde{t}\wedge X_i)} \exp(\beta^T Z_i) \lambda \{\exp(\beta^T Z_i) s\} ds \right| \\
\geq \;& \frac{1}{n} \left| \sum_{i=1}^{n} \frac{V_i \epsilon_2 I(X_i = \tau)}{\pi(W_i)} \int_{H_2(\tilde{t})}^{H_1(\tilde{t})} \exp(\beta^T Z_i) \lambda \{\exp(\beta^T Z_i) s\} ds \right| \\
\geq \;& \frac{1}{n} \sum_{i=1}^{n} \frac{V_i \epsilon_2 I(X_i = \tau) \exp(\beta^T Z_i)}{\pi(W_i)} \inf_{0<b<a<m, a-b=\epsilon_*} \left\{ \int_b^a \lambda \{\exp(\beta^T Z_i) s\} ds \right\},
\end{aligned}$$

where m is some constant. Note that the hazard function $\lambda(t)$ is positive. By the Law of Large Numbers, we obtain that there exists some $\tilde{\epsilon} > 0$ such that,

$$\sup_{t\in[0,\tau]} \mid A(H_1)(t) - A(H_2)(t) \mid \geq \tilde{\epsilon}. \tag{5.1}$$

Now, we show that $d(\widehat{H}_W, H_0) = o_p(1)$. If this is not true, then there exists $\epsilon > 0$ such that $d(\widehat{H}_W, H_0) > \epsilon$. By (5.1), we have

$$\sup_{t\in[0,\tau]} \mid A(\widehat{H}_W)(t) - A(H_0)(t) \mid = \sup_{t\in[0,\tau]} \mid A(H_0)(t) \mid \geq \tilde{\epsilon}. \tag{5.2}$$

However, by the Glivenko-Cantelli theorem (Van der Vaart, 1998), $A(H_0)(t) = o_p(1)$ uniformly in t, implying that $\sup_{t\in[0,\tau]} \mid A(H_0)(t) \mid = o_p(1)$, which contradicts to (5.2). Thus, $d(\widehat{H}_W, H_0) = o_p(1)$.

Part 2: Asymptotic expansions for $\widehat{H}_W$

By definition,

$$\begin{aligned}
d\widehat{H}_W(t) - dH_0(t) &= \frac{1}{n} \frac{\sum_{i=1}^n \Delta_i(t) dN_i(t)}{S_W^{(0)}(t;\beta,\widehat{H}_W,\pi)} - dH_0(t) \\
&= \left\{ \frac{1}{n} \frac{\sum_{i=1}^n \Delta_i(t) dN_i(t)}{S_W^{(0)}(t;\beta,H_0,\pi)} - dH_0(t) \right\} \\
&\quad + \frac{1}{n} \left\{ \frac{\sum_{i=1}^n \Delta_i(t) dN_i(t)}{S_W^{(0)}(t;\beta,\widehat{H}_W,\pi)} - \frac{\sum_{i=1}^n \Delta_i(t) dN_i(t)}{S_W^{(0)}(t;\beta,H_0,\pi)} \right\} \\
&\triangleq I_1 + I_2.
\end{aligned}$$

Note that, for $k = 0, 1, 2$,

$$S_W^{(k)}(t; \beta, H, \pi) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i(t) Y_i(t) \lambda_i(t; \beta, H) Z_i^{\otimes k} \exp(\beta^T Z_i),$$

and

$$s^{(k)}(t; \beta) = E\{B_i(t) Y_i(t) \lambda_i(t; \beta, H) Z_i^{\otimes k} \exp(\beta^T Z_i)\}, \quad k = 0, 1, 2.$$

By the Glivenko-Cantelli theorem, $S_W^{(k)}(t; \beta, H, \pi)$ converges uniformly over t and β to $s^{(k)}(t; \beta)$ and $R(t; \beta, H, \pi)$ converges uniformly over t and β to 0. Therefore,

$$\begin{aligned} I_1 &= \frac{1}{n} \frac{\sum_{i=1}^{n} \{\Delta_i(t) dN_i(t) - S_W^{(0)}(t; \beta, H_0, \pi) dH_0(t)\}}{S_W^{(0)}(t; \beta, H_0, \pi)} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i(t) dM_i(t)}{s^{(0)}(t; \beta)} + o_p(n^{-1/2}). \end{aligned}$$

Similarly, we can write

$$I_2 = -\frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i(t) dN_i(t)}{\{s^{(0)}(t; \beta)\}^2} \{S_W^{(0)}(t; \beta, \widehat{H}_W, \pi) - S_W^{(0)}(t; \beta, H_0, \pi)\} + o_p(n^{-1/2}). \quad (5.3)$$

Denote

$$v^{(k)}(t; \beta) = E\{B_i(t) Y_i(t) \dot{\lambda}_i(t; \beta, H) Z_i^{\otimes k} \exp(2\beta^T Z_i)\}, \quad k = 0, 1, 2.$$

By the consistency of $\widehat{H}_W$ and the Taylor series expansion,

$$S_W^{(0)}(t; \beta, \widehat{H}_W, \pi) - S_W^{(0)}(t; \beta, H_0, \pi) = v^{(0)}(t; \beta) \{\widehat{H}_W(t) - H_0(t)\} + o_p(n^{-1/2}). \quad (5.4)$$

Therefore, (5.3) and (5.4) yield that,

$$\begin{aligned} I_2 &= -\frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i(t) dN_i(t)}{\{s^{(0)}(t; \beta)\}^2} v^{(0)}(t; \beta) \{\widehat{H}_W(t) - H_0(t)\} + o_p(n^{-1/2}) \\ &= -\frac{v^{(0)}(t, \beta)}{s^{(0)}(t, \beta)} \{\widehat{H}_W(t) - H_0(t)\} dH_0(t) + o_p(n^{-1/2}). \end{aligned}$$

Combining expansions for I_1 and I_2, we obtain

$$d\widehat{H}_W(t) - dH_0(t) = -\frac{v^{(0)}(t, \beta)}{s^{(0)}(t, \beta)} \{\widehat{H}_W(t) - H_0(t)\} dH_0(t) + \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i(t) dM_i(t)}{s^{(0)}(t; \beta)} + o_p(n^{-1/2}),$$

which defines a first order ordinary differential equation for $\widehat{H}_W(t) - H_0(t)$. This equation has an explicit solution, given by

$$\widehat{H}_W(t) - H_0(t) = \frac{1}{n e(t, \beta)} \sum_{i=1}^{n} \int_0^t \frac{e(u, \beta)}{s^{(0)}(u, \beta)} \Delta_i(u) dM_i(u) + o_p(n^{-1/2}),$$

where

$$e(t,\beta)=\exp\left\{\int_0^t \frac{v^{(0)}(u,\beta)}{s^{(0)}(u,\beta)}dH_0(u)\right\}.$$

Part 3: Asymptotic normality of $U_W(\beta_0,\widehat{H}_W,\pi;B,D)$

Let

$$R_W^{(k)}(t;\beta,H,\pi)=\frac{1}{n}\sum_{j=1}^n \Delta_j(t)Y_j(t)\exp(\beta^T Z_j)Z_j^{\otimes k}k_j(t-;\beta,H),\quad k=0,1.$$

Denote $S^{(k)}(t)=S_W^{(k)}(t;\beta_0,H_0,\pi)$, $R^{(k)}(t)=R_W^{(k)}(t;\beta_0,H_0,\pi)$, $\widehat{S}^{(k)}(t)=S_W^{(k)}(t;\beta_0,\widehat{H}_W,\pi)$, $\widehat{R}^{(k)}(t)=R_W^{(k)}(t;\beta_0,\widehat{H}_W,\pi)$, $s^{(k)}(t)=s^{(k)}(t;\beta_0)$, $v^{(k)}(t)=v^{(k)}(t;\beta_0)$ and $e(t)=e(t;\beta_0)$. Note that $\lambda_i(t-;\beta,\widehat{H}_W)=\lambda_i(t;\beta,\widehat{H}_W)+o_p(n^{-1})$ and $k_i(t-;\beta,\widehat{H}_W)=k_i(t;\beta,\widehat{H}_W)+o_p(n^{-1})$. Thus, replacing $\lambda_i(t-;\beta,\widehat{H}_W)$ and $k_i(t-;\beta,\widehat{H}_W)$ respectively with $\lambda_i(t;\beta,\widehat{H}_W)$ and $k_i(t;\beta,\widehat{H}_W)$ does not alter asymptotic results. Note that

$$U_W(\beta_0,\widehat{H}_W,\pi;B,D)=\frac{1}{n}\sum_{i=1}^n\int_0^\tau \Delta_i(t)D(t)\Bigg[Z_i - \frac{\sum_{j=1}^n \Delta_j(t)Y_j(t)Z_j\exp(\beta_0^TZ_j)\{\lambda_j(t-;\beta_0,\widehat{H}_W)-k_j(t-;\beta_0,\widehat{H}_W)\}}{\sum_{j=1}^n \Delta_j(t)Y_j(t)\exp(\beta_0^TZ_j)\{\lambda_j(t-;\beta_0,\widehat{H}_W)-k_j(t-;\beta_0,\widehat{H}_W)\}}\Bigg]dN_i(t).$$

Then we derive

$$\begin{aligned}
& n^{1/2}U_W(\beta_0,\widehat{H}_W,\pi;B,D)\\
=\;& \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau \Delta_i(t)D(t)\left\{Z_i-\frac{\widehat{S}^{(1)}(t)-\widehat{R}^{(1)}(t)}{\widehat{S}^{(0)}(t)-\widehat{R}^{(0)}(t)}\right\}dN_i(t)\\
=\;& \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau \Delta_i(t)D(t)\left\{Z_i-\frac{\widehat{S}^{(1)}(t)-\widehat{R}^{(1)}(t)}{\widehat{S}^{(0)}(t)-\widehat{R}^{(0)}(t)}\right\}\{dM_i(t)+Y_i(t)\lambda_i(t;\beta_0,H)\exp(\beta_0^TZ_i)dH_0(t)\}\\
=\;& \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau \Delta_i(t)D(t)\left\{Z_i-\frac{\widehat{S}^{(1)}(t)-\widehat{R}^{(1)}(t)}{\widehat{S}^{(0)}(t)-\widehat{R}^{(0)}(t)}\right\}dM_i(t)\\
& + n^{1/2}\int_0^\tau D(t)\left\{S^{(1)}(t)-\frac{\widehat{S}^{(1)}(t)-\widehat{R}^{(1)}(t)}{\widehat{S}^{(0)}(t)-\widehat{R}^{(0)}(t)}S^{(0)}(t)\right\}dH_0(t)\\
\triangleq\;& J_1+J_2. \qquad (5.5)
\end{aligned}$$

Simple calculation shows that $E(R_W^{(k)}(t;\beta,H,\pi))=0$. By the consistency of $\widehat{H}(t)$ and the Glivenko-Cantelli theorem, we have $S^{(k)}(t)=s^{(k)}(t)+o_p(1)$, $R^{(k)}(t)=o_p(1)$, $\widehat{S}^{(k)}(t)=s^{(k)}(t)+o_p(1)$ and $\widehat{R}^{(k)}(t)=o_p(1)$ uniformly over t. Therefore,

$$J_1=\frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau \Delta_i(t)D(t)\left\{Z_i-\frac{s^{(1)}(t)}{s^{(0)}(t)}\right\}dM_i(t)+o_p(1). \qquad (5.6)$$

To closely examine J_2, we consider expansions of $\widehat{S}^{(k)}(t)$ and $\widehat{R}^{(k)}(t)$, for $k = 0, 1$. By the consistency of $\widehat{H}(t)$ and the Taylor series expansion, it is easy to show that

$$\widehat{S}^{(k)}(t) = S^{(k)}(t) + v^{(k)}(t)\{\widehat{H}_W(t) - H_0(t)\} + o_p(n^{-1/2}). \tag{5.7}$$

Note that

$$\eta_1^{(k)}(s,t) = E\left\{ B_i(t) Z_i^{\otimes k} \exp(3\beta_0^T Z_i) Y_i(s) \frac{\dot{\lambda}_i^2(s;\beta_0,H_0)}{\lambda_i(s;\beta_0,H_0)} \right\},$$
$$\eta_2^{(k)}(s,t) = E\{ B_i(t) Z_i^{\otimes k} \exp(2\beta_0^T Z_i) Y_i(s) \dot{\lambda}_i(s;\beta_0,H_0) \}.$$

After some algebra, we therefore obtain that

$$\begin{aligned}
&\widehat{R}^{(k)}(t) - R^{(k)}(t) \\
= & -\int_{t+}^{\tau} \left\{ \frac{1}{n}\sum_{i=1}^{n} \Delta_i(t) Y_i(s) Z_i^{\otimes k} \exp(3\beta_0^T Z_i) \frac{\dot{\lambda}_i^2(s;\beta_0,H_0)}{\lambda_i(s;\beta_0,H_0)} \right\} \{\widehat{H}_W(s) - H_0(s)\} dH_0(s) \\
& -\int_{t+}^{\tau} \left\{ \frac{1}{n}\sum_{i=1}^{n} \Delta_i(t) Y_i(s) Z_i^{\otimes k} \exp(2\beta_0^T Z_i) \dot{\lambda}_i(s;\beta_0,H_0) \right\} d\{\widehat{H}_W(s) - H_0(s)\} + o_p(n^{-1/2}) \\
= & -\int_{t+}^{\tau} \eta_1^{(k)}(s,t) \{\widehat{H}_W(s) - H_0(s)\} dH_0(s) - \int_{t+}^{\tau} \eta_2^{(k)}(s,t) d\{\widehat{H}_W(s) - H_0(s)\} + o_p(n^{-1/2}) \\
= & -\frac{1}{n}\sum_{i=1}^{n} \int_{t+}^{\tau} \frac{\eta_1^{(k)}(s,t)}{e(s)} \int_0^s \frac{e(u)}{s^{(0)}(u)} \Delta_i(u) dM_i(u) dH_0(s) \\
& + \frac{1}{n}\sum_{i=1}^{n} \int_{t+}^{\tau} \frac{\eta_2^{(k)}(s,t) v^{(0)}(s)}{s^{(0)}(s) e(s)} \int_0^s \frac{e(u)}{s^{(0)}(u)} \Delta_i(u) dM_i(u) dH_0(s) \\
& - \frac{1}{n}\sum_{i=1}^{n} \int_{t+}^{\tau} \frac{\eta_2^{(k)}(s,t)}{s^{(0)}(s)} \Delta_i(s) dM_i(s) + o_p(n^{-1/2}).
\end{aligned}$$

Interchanging the order of integration leads to

$$\begin{aligned}
&\widehat{R}^{(k)}(t) - R^{(k)}(t) \\
= & -\frac{1}{n}\sum_{i=1}^{n} \int_0^t \frac{e(u)\Delta_i(u)}{s^{(0)}(u)} \{\xi_1^{(k)}(t,t) - \xi_2^{(k)}(t,t)\} dM_i(u) \\
& - \frac{1}{n}\sum_{i=1}^{n} \int_{t+}^{\tau} \left[\frac{\eta_2^{(k)}(s,t)}{s^{(0)}(s)} + \frac{e(u)}{s^{(0)}(u)} \{\xi_1^{(k)}(u,t) - \xi_2^{(k)}(u,t)\} \right] \Delta_i(u) dM_i(u) \\
& + o_p(n^{-1/2}),
\end{aligned} \tag{5.8}$$

where

$$\xi_1^{(k)}(u,t) = \int_u^{\tau} \frac{\eta_1^{(k)}(s,t)}{e(s)} dH_0(s), \quad \xi_2^{(k)}(u,t) = \int_u^{\tau} \frac{\eta_2^{(k)}(s,t) v^{(0)}(s)}{e(s) s^{(0)}(s)} dH_0(s).$$

By definition of J_2 in (5.5), we obtain

$$\begin{aligned} J_2 &= n^{1/2}\int_0^\tau D(t)\frac{\widehat{S}^{(0)}(t)S^{(1)}(t)-\widehat{S}^{(1)}(t)S^{(0)}(t)}{\widehat{S}^{(0)}(t)-\widehat{R}^{(0)}(t)}dH_0(t) \\ &\quad - n^{1/2}\int_0^\tau D(t)\frac{\widehat{R}^{(0)}(t)S^{(1)}(t)-\widehat{R}^{(1)}(t)S^{(0)}(t)}{\widehat{S}^{(0)}(t)-\widehat{R}^{(0)}(t)}dH_0(t) \\ &\triangleq J_{21}-J_{22}. \end{aligned}$$

By (5.7),

$$\begin{aligned} J_{21} &= n^{1/2}\int_0^\tau D(t)\frac{s^{(1)}(t)v^{(0)}(t)-v^{(1)}(t)s^{(0)}(t)}{s^{(0)}(t)}\{\widehat{H}_W(t)-H_0(t)\}dH_0(t)+o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau D(t)\frac{s^{(1)}(t)v^{(0)}(t)-v^{(1)}(t)s^{(0)}(t)}{s^{(0)}(t)e(t)}\int_0^t\frac{e(u)}{s^{(0)}(u)}\Delta_i(u)dM_i(u)dH_0(t)+o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau\frac{e(u)}{s^{(0)}(u)}\Delta_i(u)q_1(u)dM_i(u)+o_p(1), \end{aligned}$$

where

$$q_1(u)=\int_u^\tau D(t)\frac{s^{(1)}(t)v^{(0)}(t)-v^{(1)}(t)s^{(0)}(t)}{s^{(0)}(t)e(t)}dH_0(t).$$

Similarly, by (5.8), after some algebra, we derive

$$\begin{aligned} J_{22} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau\frac{D(t)s^{(1)}(t)}{s^{(0)}(t)}\bigg(\int_0^t\frac{e(u)\Delta_i(u)}{s^{(0)}(u)}\{\xi_1^{(0)}(t,t)-\xi_2^{(0)}(t,t)\}dM_i(u) \\ &\quad +\int_{t+}^\tau\bigg[\frac{\eta_2^{(0)}(u,t)}{s^{(0)}(u)}+\frac{e(u)}{s^{(0)}(u)}\{\xi_1^{(0)}(u,t)-\xi_2^{(0)}(u,t)\}\bigg]\Delta_i(u)dM_i(u)\bigg)dH_0(t) \\ &\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau D(t)\bigg(\int_0^t\frac{e(u)\Delta_i(u)}{s^{(0)}(u)}\{\xi_1^{(1)}(t,t)-\xi_2^{(1)}(t,t)\}dM_i(u) \\ &\quad +\int_{t+}^\tau\bigg[\frac{\eta_2^{(1)}(u,t)}{s^{(0)}(u)}+\frac{e(u)}{s^{(0)}(u)}\{\xi_1^{(1)}(u,t)-\xi_2^{(1)}(u,t)\}\bigg]\Delta_i(u)dM_i(u)\bigg)dH_0(t)+o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau\frac{\Delta_i(u)}{s^{(0)}(u)}\bigg[e(u)\bigg\{q_5(u)+q_7(u)-q_2(u)-q_4(u)\bigg\}+q_6(u)-q_3(u)\bigg]dM_i(u), \end{aligned}$$

where

$$q_2(u)=\int_u^\tau\frac{D(t)s^{(1)}(t)}{s^{(0)}(t)}\{\xi_1^{(0)}(t,t)-\xi_2^{(0)}(t,t)\}dH_0(t),$$

$$q_3(u)=\int_0^u\frac{D(t)s^{(1)}(t)}{s^{(0)}(t)}\eta_2^{(0)}(u,t)dH_0(t),$$

$$q_4(u) = \int_0^u \frac{D(t)s^{(1)}(t)}{s^{(0)}(t)}\{\xi_1^{(0)}(u,t) - \xi_2^{(0)}(u,t)\}dH_0(t),$$

$$q_5(u) = \int_u^\tau D(t)\{\xi_1^{(1)}(t,t) - \xi_2^{(1)}(t,t)\}dH_0(t),$$

$$q_6(u) = \int_0^u D(t)\eta_2^{(1)}(u,t)dH_0(t),$$

and

$$q_7(u) = \int_0^u D(t)\{\xi_1^{(1)}(u,t) - \xi_2^{(1)}(u,t)\}dH_0(t).$$

As a result, we obtain that

$$\begin{aligned} J_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \Big[e(u)\{q_1(u) + q_2(u) + q_4(u) - q_5(u) - q_7(u)\} \\ &\quad + q_3(u) - q_6(u)\Big]\frac{\Delta_i(u)}{s^{(0)}(u)}dM_i(u) + o_p(1). \end{aligned} \tag{5.9}$$

Combining (5.6) and (5.9) gives us

$$\begin{aligned} & n^{1/2}U_W(\beta_0, \widehat{H}_W, \pi; B, D) \\ = \ & \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left[D(t)\left\{Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)}\right\} + \frac{Q(t)}{s^{(0)}(t)}\right]\Delta_i(t)dM_i(t) + o_p(1) \end{aligned} \tag{5.10}$$

where

$$Q(t) = e(t)\{q_1(t) + q_2(t) + q_4(t) - q_5(t) - q_7(t)\} + q_3(t) - q_6(t). \tag{5.11}$$

This implies that $n^{1/2}U_W(\beta_0, \widehat{H}_W, \pi; B, D)$ can be approximated by a sum of iid mean zero random variables. Therefore, by the central limit theorem,

$$n^{1/2}U_W(\beta, \widehat{H}_W, \pi; B, D) \xrightarrow{d} N(0, \Sigma_W),$$

where $\Sigma_W = E(M_i^{*\otimes 2}/\pi)$, and

$$M_i^* = \int_0^\tau B_i(t)\left[D(t)\left\{Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)}\right\} + \frac{Q(t)}{s^{(0)}(t)}\right]dM(t).$$

Part 4: Convergence of $-(\partial/\partial\beta)U_W(\beta_0, \widehat{H}_W, \pi; B, D)$

Let

$$\Psi^{(k)}(t;\beta) = \frac{1}{n}\sum_{j=1}^n \Delta_j(t)Y_j(t)Z_j^{\otimes k}\exp(\beta^T Z_j)\left\{\frac{\partial}{\partial\beta}\lambda_j(t-;\beta,\widehat{H}_W) - \frac{\partial}{\partial\beta}k_j(t-;\beta,\widehat{H}_W)\right\}, \ k = 0, 1.$$

It is easy to verify that

$$\frac{\partial}{\partial\beta}\lambda_j(t-;\beta,\widehat{H}_W)=\dot{\lambda}_j(t-;\beta,\widehat{H}_W)\exp(\beta^TZ_j)\bigg\{Z_j\widehat{H}_W(t)+\frac{\partial}{\partial\beta}\widehat{H}_W(t)\bigg\}.$$

After some algebra, we can show that

$$\begin{aligned}\Psi^{(k)}(t;\beta_0) &= v^{(k+1)}(t)H_0(t)+v^{(k)}(t)\frac{\partial}{\partial\beta}\widehat{H}_W(t)+\int_t^\tau\eta_2^{(k)}(u,t)d\bigg\{\frac{\partial}{\partial\beta}\widehat{H}_W(u)\bigg\}\\ &\quad+\int_t^\tau\bigg\{\eta_2^{(k+1)}(u,t)+\eta_1^{(k+1)}(u,t)H_0(u)+\eta_2^{(k)}(u,t)\frac{\partial}{\partial\beta}\widehat{H}_W(u)\bigg\}dH_0(u)+o_p(1).\end{aligned}$$

Next, we derive the asymptotic limits for $(\partial/\partial\beta)\widehat{H}_W$ and $d(\partial/\partial\beta)\widehat{H}_W$. Straightforward calculation shows that

$$d\bigg\{\frac{\partial}{\partial\beta}\widehat{H}_W(t)\bigg\} = -\frac{dH_0(t)}{s^{(0)}(t;\beta)}\{s^{(1)}(t;\beta)+v^{(1)}(t;\beta)H_0(t)+v^{(0)}(t;\beta)\frac{\partial}{\partial\beta}\widehat{H}_W(t)\}+o_p(1),$$

which again produces a first order ordinary differential equation for $(\partial/\partial\beta)\widehat{H}_W(t)$. Then

$$\frac{\partial}{\partial\beta}\widehat{H}_W(t)=-\frac{1}{e(t;\beta)}\int_0^t\frac{e(u;\beta)\{s^{(1)}(u;\beta)+v^{(1)}(u;\beta)H_0(u)\}}{s^{(0)}(u;\beta)}dH_0(u)+o_p(1).$$

Let

$$K(t)=-\frac{1}{e(t)}\int_0^t\frac{e(u)\{s^{(1)}(u)+v^{(1)}(u)H_0(u)\}}{s^{(0)}(u)}dH_0(u),$$

$$\begin{aligned}\psi^{(k)}(t;\beta_0) &= v^{(k+1)}(t)H_0(t)+v^{(k)}(t)K(t)\\ &\quad-\int_t^\tau\frac{\eta_2^{(k)}(u,t)\{s^{(1)}(u)+v^{(1)}(u)H_0(u)+v^{(0)}(u)K(u)\}}{s^{(0)}(u)}dH_0(u)\\ &\quad+\int_t^\tau\bigg\{\eta_2^{(k+1)}(u,t)+\eta_1^{(k+1)}(u,t)H_0(u)+\eta_2^{(k)}(u,t)K(u)\bigg\}dH_0(u),\end{aligned}$$

$$\Gamma(t;\beta_0)=\frac{S_W^{(2)}(t;\beta_0)+\Psi^{(1)}(t;\beta_0)}{S_W^{(0)}(t;\beta_0)}+\frac{S_W^{(1)}(t;\beta_0)\{S_W^{(1)}(t;\beta_0)+\Psi^{(0)}(t;\beta_0)\}^T}{\{S_W^{(0)}(t;\beta_0)\}^{\otimes2}},$$

and

$$\gamma(t;\beta_0)=\frac{s^{(2)}(t;\beta_0)+\psi^{(1)}(t;\beta_0)}{s^{(0)}(t;\beta_0)}+\frac{s^{(1)}(t;\beta_0)\{s^{(1)}(t;\beta_0)+\psi^{(0)}(t;\beta_0)\}^T}{\{s^{(0)}(t;\beta_0)\}^{\otimes2}}.$$

By the Glivenko-Cantelli theorem, $\Psi^{(k)}(t;\beta)=\psi^{(k)}(t;\beta)+o_p(1)$ and hence, $\Gamma(t;\beta)=\gamma(t;\beta)+$

$o_p(1)$ uniformly in β and t. After some algebra, we obtain

$$\begin{aligned} -\frac{\partial}{\partial\beta}U_W(\beta_0, \widehat{H}_W, \pi; B, D) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Delta_i(t)D(t)\Gamma(t;\beta_0)dN_i(t) + o_p(1) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Delta_i(t)D(t)\gamma(t;\beta_0)dN_i(t) + o_p(1) \\ &= \int_0^\tau D(t)\gamma(t;\beta_0)s^{(0)}(t)dH_0(t) + o_p(1) \stackrel{\triangle}{=} I_\beta + o_p(1), \end{aligned}$$

where the $o_p(1)$ terms are uniformly in β.

Part 5: Consistency of $\hat{\beta}_W$

In Part 3, we obtain that $U_W(\beta, \widehat{H}_W, \pi; B, D) \to 0$ in probability. By assumption (A4), Σ_W is positive definite. Then following the same arguments for the proof of Theorem 2 of Foutz (1977), we can show that $\hat{\beta}_W$ exists and is unique in a compact neighborhood of β_0 with probability converging to 1 as $n \to \infty$, and $\hat{\beta}_W$ is consistent for β_0.

Part 6: Asymptotic normality of $n^{1/2}(\hat{\beta}_W - \beta_0)$

By definition and the mean value theorem,

$$\begin{aligned} 0 &= U_W(\hat{\beta}_W, \widehat{H}_W, \pi; B, D) \\ &= U_W(\beta_0, \widehat{H}_W, \pi; B, D) + \{U_W(\hat{\beta}_W, \widehat{H}_W, \pi; B, D) - U_W(\beta_0, \widehat{H}_W, \pi; B, D)\} \\ &= U_W(\beta_0, \widehat{H}_W, \pi; B, D) + \frac{\partial}{\partial\beta}U_W(\beta^*, \widehat{H}_W, \pi; B, D)(\hat{\beta}_W - \beta_0), \end{aligned}$$

where β^* is between β_0 and $\hat{\beta}_W$. Then

$$n^{1/2}(\hat{\beta}_W - \beta_0) = -\left\{\frac{\partial}{\partial\beta}U_W(\beta^*, \widehat{H}_W, \pi; B, D)\right\}^{-1}\left\{n^{1/2}U_W(\beta_0, \widehat{H}_W, \pi; B, D)\right\}$$

By the results in Parts 3-5, we obtain that $n^{1/2}(\hat{\beta}_W - \beta_0)$ converges weakly to $N(0, I_\beta^{-1}\Sigma_W I_\beta^{-1})$.

5.1. Proof of Theorem 4.1

Lemma 5.1. *Let $g_i(\beta, t) = g(X_i, \delta_i, Z_i, \beta, t)$ be a real-valued function satisfying $E|g_i(\beta, t)| \leq \infty$. Then under conditions (B1) and (D1), we have*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\tilde{\pi}(W_i)}g_i(\beta, t) = \frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\pi(W_i)}g_i(\beta, t) + o_p(1).$$

Proof:. We assume that W_i is continuous. For notational simplicity, let $g_i = g_i(\beta, t)$. Noting that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\tilde{\pi}(W_i)}g_i = \frac{1}{n}\sum_{i=1}^{n}\frac{V_i}{\pi(W_i)}g_i - \frac{1}{n}\sum_{i=1}^{n}\frac{V_i(\tilde{\pi}(W_i)-\pi(W_i))}{\tilde{\pi}(W_i)\pi(W_i)}g_i \triangleq I_1 + I_2,$$

it suffices to show that $I_2 = o_p(1)$. Indeed,

$$\begin{aligned}|I_2| \;\leq\; & \sup_w |\tilde{\pi}(w)-\pi(w)|\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{V_i|g_i|}{\tilde{\pi}(W_i)\pi(W_i)}I(\tilde{\pi}(W_i)\geq \pi(W_i)/2) \\ & + \frac{1}{n}\sum_{i=1}^{n}\frac{V_i|g_i|}{\tilde{\pi}(W_i)\pi(W_i)}I(\tilde{\pi}(W_i)< \pi(W_i)/2)\Big\}.\end{aligned}$$

Clearly, by condition (B1),

$$\frac{1}{n}\sum_{i=1}^{n}\frac{V_i|g_i|}{\tilde{\pi}(W_i)\pi(W_i)}I\{\tilde{\pi}(W_i)\geq \pi(W_i)/2\} \leq \frac{2}{n}\sum_{i=1}^{n}\frac{V_i|g_i|}{\pi^2(W_i)} = O_p(1).$$

For any $\epsilon > 0$, by conditions (B1) and (D1), we have,

$$\begin{aligned} & P\left[\frac{1}{n}\sum_{i=1}^{n}\frac{V_i|g_i|}{\tilde{\pi}(W_i)\pi(W_i)}I\{\tilde{\pi}(W_i)< \pi(W_i)/2\} > \epsilon\right] \\ \leq\; & P\Big(\bigcup_{i=1}^{n}\{I(\tilde{\pi}(W_i)< \pi(W_i)/2)=1\}\Big) \\ \leq\; & P\Big(\bigcup_{i=1}^{n}\{|\tilde{\pi}(W_i)-\pi(W_i)| > \inf_w \pi(w)/2\}\Big) \\ \leq\; & P\Big(\sup_w |\tilde{\pi}(w)-\pi(w)| > \inf_w \pi(w)/2\Big) \\ \leq\; & P\Big(\sup_w |\tilde{\pi}(w)-\pi(w)| > \epsilon_1/2\Big) \to 0, \quad \text{as } n\to\infty.\end{aligned}$$

By conditions (D1), we have $I_2 = O_p(n^{-c}) = o_p(1)$, which completes the proof. □

Proof of Theorem 4.1. Theorem 4.1 is obtained by modifying the proof of Theorem 3.1. We first show that $\widetilde{H}$ is consistent. To this end, one needs to show (8.1) holds. Using the same notations and arguments as in Part 1 of Appendix B, we have,

$$\begin{aligned} & \sup_{t\in[0,\tau]} |\, A(H_1)(t) - A(H_2)(t)\,| \\ \geq\; & \frac{1}{n}\sum_{i=1}^{n}\frac{V_i\epsilon_2 I(X_i=\tau)\exp(\beta^T Z_i)}{\tilde{\pi}(W_i)} \inf_{0<b<a<m, a-b=\epsilon_*}\left\{\int_b^a \lambda\{\exp(\beta^T Z_i)s\}ds\right\} \\ =\; & \frac{1}{n}\sum_{i=1}^{n}\frac{V_i\epsilon_2 I(X_i=\tau)\exp(\beta^T Z_i)}{\pi(W_i)} \inf_{0<b<a<m, a-b=\epsilon_*}\left\{\int_b^a \lambda\{\exp(\beta^T Z_i)s\}ds\right\} + o_p(1),\end{aligned}$$

where the last step follows from Lemma 5.1. This implies that (8.1) still holds. Thus, we can show that $\widetilde{H}$ is consistent.

Following the similar arguments to those in Part 2 of Appendix B, we obtain

$$\widetilde{H}(t) - H_0(t) = \frac{1}{ne(t,\beta)} \sum_{i=1}^{n} \int_0^t \frac{e(u,\beta)}{s^{(0)}(u,\beta)} \widetilde{\Delta}_i(u) dM_i(u) + o_p(n^{-1/2}),$$

where $\widetilde{\Delta}_i(u) = V_i B_i(u)/\tilde{\pi}(W_i)$. Similar arguments to those in Part 3 of Appendix B yield that

$$\begin{aligned}
& n^{1/2} U_W(\beta_0, \widetilde{H}, \tilde{\pi}; B, D) \\
= & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{V_i}{\tilde{\pi}(W_i)} M_i^* + o_p(1) \\
= & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{V_i}{\pi(W_i)} M_i^* + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1 - \frac{V_i}{\pi(W_i)}\right) m(W_i; \beta_0, H_0, \pi) + o_p(1) \\
\triangleq & F_n + o_p(1),
\end{aligned}$$

where the second equality follows from Condition (D2). It is easy to show that $var(F_n) = \widetilde{\Sigma}$. Finally, similar arguments to those in Part 4 of Appendix B yield that

$$\begin{aligned}
-\frac{\partial}{\partial \beta} U_W(\beta_0, \widetilde{H}, \tilde{\pi}; B, D) &= \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \tilde{\Delta}_i(t) D(t) \gamma(t; \beta_0) dN_i(t) + o_p(1) \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \Delta_i(t) D(t) \gamma(t; \beta_0) dN_i(t) + o_p(1) \\
&= \int_0^\tau D(t) \gamma(t; \beta_0) s^{(0)}(t) dH_0(t) + o_p(1),
\end{aligned}$$

where the second equality follows from Lemma 5.1, and the third equality follows from the Weak Law of Large Numbers.

By analogy with Parts 5 and 6 in Appendix B, we can show the consistency of $\tilde{\beta}$ and establish that

$$\sqrt{n}(\tilde{\beta} - \beta_0) \xrightarrow{d} N(0, I_\beta^{-1} \widetilde{\Sigma} I_\beta^{-1}),$$

where $I_\beta = \int_0^\tau D(t) \gamma(t; \beta_0) s^{(0)}(t) dH_0(t)$. This completes the proof of Theorem 4.1. □

5.2. Proof of Theorem 4.2

The asymptotic arguments in Appendix B can be applied to obtain an asymptotic expansion for $U_W(\beta_0, \widehat{H}_W, \pi(\alpha_0); B, D)$. By (8.10), we can write

$$n^{1/2} U_W(\beta_0, \widehat{H}_W, \pi(\alpha_0); B, D) = n^{-1/2} \sum_{i=1}^{n} U_{\beta,i}(\alpha_0, \beta_0) + o_p(1),$$

where

$$U_{\beta,i}(\alpha_0, \beta_0) = \int_0^\tau \left[D(t) \left\{ Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)} \right\} + \frac{Q(t)}{s^{(0)}(t)} \right] \Delta_i(t) dM_i(t),$$

and $Q(t)$ is defined in (8.11). Note that the score function for α is given by $U_\alpha(\alpha) = n^{-1} \sum_{i=1}^{n} U_{\alpha,i}(\alpha)$, where

$$U_{\alpha,i}(\alpha) = \frac{V_i - \pi_i(\alpha)}{\pi_i(\alpha)\{1 - \pi_i(\alpha)\}} \dot{\pi}_i(\alpha).$$

Thus, $n^{1/2}\{U_W^T(\beta_0, \widehat{H}_W, \pi(\alpha_0); B, D), U_\alpha^T(\alpha_0)\}$ can be asymptotically written as a sum of n iid random vectors. By the multivariate central limit theorem,

$$n^{1/2}\{U_W^T(\beta_0, \widehat{H}_W, \pi(\alpha_0); B, D), U_\alpha^T(\alpha_0)\} \xrightarrow{d} N(0, \Sigma_{SWF})$$

where Σ_{SWF} is

$$\begin{pmatrix} \Sigma_W & E\{U_{\beta,i}(\alpha_0, \beta_0) U_{\alpha,i}^T(\alpha_0)\} \\ E\{U_{\alpha,i}(\alpha_0) U_{\beta,i}^T(\alpha_0, \beta_0)\} & I_\alpha \end{pmatrix}.$$

Note that

$$\begin{aligned} E\{U_{\beta,i}(\alpha_0, \beta_0) U_{\alpha,i}^T(\alpha_0)\} &= E\left\{ M_i^* \frac{V_i}{\pi_i(\alpha_0)} \frac{V_i - \pi_i(\alpha_0)}{\pi_i(\alpha_0)\{1 - \pi_i(\alpha_0)\}} \frac{\partial \pi_i(\alpha)}{\partial \alpha} \right\} \\ &= E\left\{ M_i^* \frac{\dot{\pi}_i(\alpha_0)}{\pi_i(\alpha_0)} \right\} = I_{\alpha\beta}, \end{aligned}$$

and $I_\alpha = E\{U_{\alpha,i}(\alpha_0)\}^{\otimes 2}$. We have $-(\partial/\partial\alpha) U_\alpha(\alpha) = I_\alpha + o_p(1)$ uniformly over α. Since $U_W(\hat{\beta}_{SW}, \widehat{H}_W, \pi(\hat{\alpha}); B, D) = 0$, the Taylor series expansion yields,

$$\begin{aligned} n^{1/2}(\hat{\beta}_{SW} - \beta_0) &= I_\beta^{-1} \Bigg[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{\beta,i}(\alpha_0, \beta_0) \\ &\qquad + \left\{ \frac{\partial}{\partial \alpha} U_W(\beta_0, \widehat{H}_W, \pi(\alpha); B, D) \right\} n^{1/2}(\hat{\alpha} - \alpha_0) \Bigg] + o_p(1) \\ &= I_\beta^{-1} \Bigg[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{\beta,i}(\alpha_0, \beta_0) \\ &\qquad + \left\{ \frac{\partial}{\partial \alpha} U_W(\beta_0, \widehat{H}_W, \pi(\alpha); B, D) \right\} I_\alpha^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{\alpha,i}(\alpha_0) \Bigg] + o_p(1). \end{aligned}$$

Applying the same arguments as that in Part 2 in Appendix B, we can show that

$$d\left\{\frac{\partial}{\partial\alpha}\widehat{H}_W(t)\right\} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{s^{(0)}(t;\beta)}\frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)}dM_i(t) - \frac{v^{(0)}(t;\beta)}{s^{(0)}(t;\beta)}\frac{\partial}{\partial\alpha}\widehat{H}_W(t)dH_0(t) + o_p(1),$$

which again produces a first order ordinary differential equation for $(\partial/\partial\alpha)\widehat{H}_W(t)$. Therefore,

$$\begin{aligned}\frac{\partial}{\partial\alpha}\widehat{H}_W(t) &= -\frac{1}{ne(t;\beta)}\sum_{i=1}^{n}\int_0^t\frac{e(u;\beta)}{s^{(0)}(u;\beta)}\frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)}dM_i(t) + o_p(1)\\ &= -E\left\{\frac{\dot{\pi}_i(\alpha)}{\pi_i(\alpha)}\frac{1}{e(t;\beta)}\int_0^t\frac{e(u;\beta)}{s^{(0)}(u;\beta)}dM_i(t)\right\} + o_p(1).\end{aligned}$$

For $k = 0, 1, 2$, let

$$\rho^{(k)}(t) = E\left\{\frac{\dot{\pi}_i(\alpha)}{\pi_i(\alpha)}B_i(t)Y_i(t)\dot{\lambda}_i(t;\beta,H)Z_i^{\otimes}\exp(2\beta^T Z_i)\right\}.$$

By differentiating $\widehat{S}^{(k)}(t)$ with respect to α, after some algebra, we obtain,

$$\frac{\partial}{\partial\alpha}\widehat{S}^{(k)}(t) = -\rho^{(k)}(t) + \frac{v^{(1)}(t;\beta)}{e(t;\beta)}E\left\{\frac{\dot{\pi}_i(\alpha)}{\pi_i(\alpha)}\int_0^t\frac{e(u;\beta)}{s^{(0)}(u;\beta)}dM_i(t)\right\} + o_p(1).$$

Thus,

$$\begin{aligned}-\frac{\partial}{\partial\alpha}U_W(\beta_0,\widehat{H}_W,\pi(\alpha);B,D) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{\pi}_i(\alpha)}{\pi_i(\alpha)}\int_0^\tau\Delta_i(t)D(t)\left\{Z_i - \frac{\widehat{S}^{(1)}(t) - \widehat{R}^{(1)}(t)}{\widehat{S}^{(0)}(t) - \widehat{R}^{(0)}(t)}\right\}dN_i(t)\\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Delta_i(t)D(t)\frac{\partial}{\partial\alpha}\left\{\frac{\widehat{S}^{(1)}(t) - \widehat{R}^{(1)}(t)}{\widehat{S}^{(0)}(t) - \widehat{R}^{(0)}(t)}\right\}dN_i(t)\\ &= E\left\{\frac{\dot{\pi}_i(\alpha_0)}{\pi_i(\alpha_0)}\int_0^\tau B_i(t)\left[D(t)\left\{Z_i - \frac{s^{(1)}(t)}{s^{(0)}(t)}\right\}\right]dM_i(t)\right\}\\ &\quad + E\left\{\frac{\dot{\pi}_i(\alpha_0)}{\pi_i(\alpha_0)}\int_0^\tau B_i(t)\frac{Q(t)}{s^{(0)}(t)}dM_i(t)\right\} + o_p(1)\\ &= I_{\alpha\beta} + o_p(1).\end{aligned}$$

Therefore, $n^{1/2}(\hat{\beta}_{SW} - \beta_0) \xrightarrow{d} N\{I_\beta^{-1}(\Sigma_W - I_{\alpha\beta}^{-1}I_\alpha I_{\alpha\beta}^{-1})I_\beta^{-1}\}$, as $n \to \infty$.

5.3. Proof of Theorem 4.3

Lemma 5.2. *Under conditions (B1) – (B3) and (C1) – (C3), we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{V_i}{\hat{\pi}(W_i)}M_i^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{V_i}{\pi(W_i)}M_i^* + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1 - \frac{V_i}{\pi(W_i)}\right)E(M_i^* \mid W_i) + o_p(1).$$

Proof of Lemma 5.2. Lemma 5.2 can be established along the same lines as for Lemma A3 of Huang and Wang (2010). Therefore, we omit the details. □

Proof of Theorem 4.3. Note that the kernel estimator $\hat{\pi}(w)$ satisfies $\sup_w |\hat{\pi}(w) - \pi(w)| = O_p(h^r + (nh^d)^{-1/2})$ (Van der Vaart, 1998). Therefore, condition (D1) holds. By Lemma 5.2, Condition (D2) holds with $m(W_i; \beta_0, H_0, \pi) = E(M_i^* \mid W_i)$. As a result, Theorem 4.3 follows directly from Theorem 4.1. □

5.4. Proof of Corollary 4.2

To show Corollary 4.2, it suffices to show that

$$E\left(\frac{1-\pi(W_i)}{\pi(W_i)} m(W_i;\beta_0,H_0,\pi) m^T(W_i;\beta_0,H_0,\pi)\right) - E\left(\frac{1-\pi(W_i)}{\pi(W_i)} M_i^{*T} m(W_i;\beta_0,H_0,\pi)\right) + E\left(\frac{1-\pi(W_i)}{\pi(W_i)} M_i^{*o\otimes 2}\right) \tag{5.12}$$

is semi-positive definite for any function $m(W_i; \beta_0, H_0, \pi)$. Indeed, it is easily seen that (5.12) is equivalent to

$$\begin{aligned} & E\left(\frac{1-\pi(W_i)}{\pi(W_i)} (M_i^{*o} - m(W_i;\beta_0,H_0,\pi))^{\otimes 2}\right) + 2E\left(\frac{1-\pi(W_i)}{\pi(W_i)} (M_i^{*o} - M_i^*)\right) \\ \triangleq\ & J_1 + J_2. \end{aligned}$$

Note that

$$J_2 = 2E\left\{E\left(\frac{1-\pi(W_i)}{\pi(W_i)} (M_i^{*o} - M_i^*) | W_i\right)\right\} = 0,$$

and J_1 is semi-positive definite. Therefore, (5.12) is semi-positive definite, which completes the proof.

5.5. Proof of Theorem 5.1

The proof of Theorem 5.1 is the extension of that of Theorem 3.1. Here, we only sketch the key steps. By the Glivenko-Cantelli theorem, we have $S_A^{(k)}(t; \beta, H, \pi)$ and $R_A^{(k)}(t; \beta, H, \pi)$ converge uniformly in t and β to $s^{(k)}(t; \beta)$ and 0, respectively. Similar arguments as in Part 2 of Appendix B can be used to show

$$\begin{aligned} \widehat{H}_A(t) - H_0(t) \ =\ & \frac{1}{ne(t,\beta)} \sum_{i=1}^n \int_0^t \frac{e(u,\beta)}{s^{(0)}(u,\beta)} \Big[\Delta_i(u) dM_i(u) \\ & - \left(\frac{V_i}{\pi(W_i)} - 1\right) E\{B_i(u) dM_i(u) \mid W_i\}\Big] + o_p(n^{-1/2}). \end{aligned}$$

Similar arguments as in Part 3 of Appendix B can be used to show

$$\begin{aligned} & n^{1/2}U_{FA}(\beta_0, \widehat{H}_A, \pi; B, D) \\ = \; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{V_i}{\pi(W_i)} M_i^* - \frac{V_i - \pi(W_i)}{\pi(W_i)} E(M_i^* \mid W_i) \right\} + o_p(1). \end{aligned} \tag{5.13}$$

It is easily verified that $U_{FA}(\beta, \widehat{H}_A, \pi; B, D) \to 0$ in probability, provided either $\pi(W_i)$ or $f(Z_i^m \mid W_i)$ is correctly specified. By the regularity condition (A6) and similar arguments in Part 5 of Appendix B, $\hat{\beta}_{FA}$ is consistent, if either $\pi(W_i)$ or $f(Z_i^m \mid W_i)$ is correctly specified. When $f(Z_i^m \mid W_i)$ and $\pi(W_i)$ are both correct, then the central limit theorem implies,

$$n^{1/2}U_{FA}(\beta_0, \widehat{H}_A, \pi; B, D) \xrightarrow{d} N\left\{0, \Sigma_W - E\left(\frac{1-\pi(W_i)}{\pi(W_i)} M_i^{*o\otimes 2}\right)\right\},$$

where $M_i^{*o} = E(M_i^* \mid W_i)$. By definition and the fact that $E(V_i \mid W_i) = \pi(W_i)$, we can show that

$$\begin{aligned} \frac{\partial}{\partial\beta} S_A^{(k)}(t; \beta, \widehat{H}_A, \pi) &= \frac{\partial}{\partial\beta} S^{(k)}(t; \beta, \widehat{H}_A, \pi) + o_p(1), \\ \frac{\partial}{\partial\beta} R_A^{(k)}(t; \beta, \widehat{H}_A, \pi) &= \frac{\partial}{\partial\beta} R^{(k)}(t; \beta, \widehat{H}_A, \pi) + o_p(1), \end{aligned}$$

and $(\partial/\partial\beta)\widehat{H}_A(t) = (\partial/\partial\beta)\widehat{H}_W(t) + o_p(1)$. As a result, we obtain

$$\begin{aligned} & -\frac{\partial}{\partial\beta} U_{FA}(\beta, \widehat{H}_A, \pi; B, D) \\ = \; & \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \Delta_i(t) D(t) \frac{\partial}{\partial\beta} \left\{ \frac{S_A^{(1)}(t; \beta, \widehat{H}_A, \pi) + R_A^{(1)}(t; \beta, \widehat{H}_A, \pi)}{S_A^{(0)}(t; \beta, \widehat{H}_A, \pi) + R_A^{(0)}(t; \beta, \widehat{H}_A, \pi)} \right\} dN_i(t) \\ & + \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{V_i}{\pi(W_i)}\right) \int_0^\tau \frac{\partial}{\partial\beta} \left\{ \frac{S_A^{(1)}(t; \beta, \widehat{H}_A, \pi) + R_A^{(1)}(t; \beta, \widehat{H}_A, \pi)}{S_A^{(0)}(t; \beta, \widehat{H}_A, \pi) + R_A^{(0)}(t; \beta, \widehat{H}_A, \pi)} E\{B_i(t) D(t) dN_i(t) \mid W_i\} \right\} \\ = \; & I_\beta + o_p(1), \end{aligned}$$

where the last equality follows from Part 4 of Appendix B. The asymptotic normality of $\hat{\beta}_{FA}$ can then be established as in Part 6 of Appendix B.

5.6. Proof of Theorem 5.2

Let $U(\theta) = \{U_{FA}^T(\beta, \widehat{H}_A, \pi; B, D), U_\phi^T(\alpha, \phi), U_\alpha^T(\alpha)\}^T$. First, we calculate $(-\partial/\partial\theta)U(\theta)$. Apparently,

$$(-\partial/\partial\beta)U_\phi(\alpha, \phi) = 0 \;\text{ and }\; (-\partial/\partial\beta)U_\alpha(\alpha) = 0. \tag{5.14}$$

In addition,

$$-\frac{\partial}{\partial\alpha}U_\phi(\alpha,\phi)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)}\frac{\partial}{\partial\phi}f(Z_i^m\mid W_i;\phi)-\frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)}E\left\{\frac{\partial}{\partial\phi}f(Z_i^m\mid W_i;\phi)\mid W_i;\phi\right\}\right],$$

and

$$-\frac{\partial}{\partial\phi}U_\phi(\alpha,\phi)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{V_i}{\pi_i(\alpha)}\frac{\partial^2}{\partial\phi^2}f(Z_i^m\mid W_i;\phi)-\frac{V_i-\pi_i(\alpha)}{\pi_i(\alpha)}\frac{\partial}{\partial\phi}E\left\{\frac{\partial}{\partial\phi}f(Z_i^m\mid W_i;\phi)\mid W_i;\phi\right\}\right].$$

By the Law of Large Numbers, if $f(Z^m\mid W;\phi)$ is correctly specified,

$$-\frac{\partial}{\partial\alpha}U_\phi(\alpha,\phi)=o_p(1)\ \text{ and }\ -\frac{\partial}{\partial\phi}U_\phi(\alpha,\phi)=I_\phi+o_p(1),\tag{5.15}$$

where $I_\phi=E\{(\partial^2/\partial\phi^2)f(Z_i^m\mid W_i;\phi)\}$. It is easily shown that, $(-\partial/\partial\alpha)U_\alpha(\alpha)=I_\alpha+o_p(1)$. As shown in Appendix B,

$$(-\partial/\partial\beta)U_{FA}(\beta,\widehat{H}_A,\pi;B,D)=I_\beta+o_p(1).\tag{5.16}$$

By definition, it is easy to show that $(\partial/\partial\phi)\widehat{H}_A(t)=o_p(1)$. We then have

$$\begin{aligned}&\frac{\partial}{\partial\phi}S_A^{(k)}(t;\beta,\widehat{H}_A,\pi)\\ =\ &\frac{1}{n}\sum_{j=1}^{n}\left(1-\frac{V_j}{\pi(W_j)}\right)Y_j(t)\frac{\partial}{\partial\phi}E\{B_j(t)Z_j^{\otimes k}\exp(\beta^TZ_j)\lambda_j(t-;\beta,\widehat{H}_A)\mid W_j\}+o_p(1)\\ =\ &o_p(1),\end{aligned}\tag{5.17}$$

where the last step follows from $E(V_j\mid W_j)=\pi(W_j)$. Similarly, we obtain

$$\frac{\partial}{\partial\phi}R_A^{(k)}(t;\beta,\widehat{H}_A,\pi)=o_p(1).\tag{5.18}$$

Therefore, by (5.18) and (5.17), we obtain that

$$\begin{aligned}&-\frac{\partial}{\partial\phi}U_{FA}(\beta,\widehat{H}_A,\pi;B,D)\\ =\ &\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Delta_i(t)D(t)\frac{\partial}{\partial\phi}\left\{\frac{S_A^{(1)}(t;\beta,\widehat{H}_A,\pi)+R_A^{(1)}(t;\beta,\widehat{H}_A,\pi)}{S_A^{(0)}(t;\beta,\widehat{H}_A,\pi)+R_A^{(0)}(t;\beta,\widehat{H}_A,\pi)}\right\}dN_i(t)\\ &+\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{V_i}{\pi(W_i)}\right)\int_0^\tau\frac{\partial}{\partial\phi}\left\{\frac{S_A^{(1)}(t;\beta,\widehat{H}_A,\pi)+R_A^{(1)}(t;\beta,\widehat{H}_A,\pi)}{S_A^{(0)}(t;\beta,\widehat{H}_A,\pi)+R_A^{(0)}(t;\beta,\widehat{H}_A,\pi)}\right.\\ &\left.\times E\{B_i(t)D(t)dN_i(t)\mid W_i\}\right\}+o_p(1)\\ =\ &o_p(1).\end{aligned}$$

Similarly, we can show that $(\partial/\partial\alpha)\widehat{H}_A(t) = o_p(1)$, and

$$\begin{aligned}
& -\frac{\partial}{\partial\alpha} S_A^{(k)}(t;\beta,\widehat{H}_A,\pi) \\
= & \frac{1}{n}\sum_{j=1}^{n} \frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)} Y_j(t) Z_j^{\otimes k} \exp(\beta^T Z_j)\lambda_j(t-;\beta,H) \\
& - \frac{1}{n}\sum_{j=1}^{n} \frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)} Y_j(t) E\{B_i(t) Z_j^{\otimes k} \exp(\beta^T Z_j)\lambda_j(t-;\beta,H) \mid W_i\} + o_p(1) \\
= & o_p(1), \qquad (5.19)
\end{aligned}$$

under the assumption that the model $f(Z_i^m \mid W_i;\phi)$ is correct. Similarly,

$$\begin{aligned}
& -\frac{\partial}{\partial\alpha} R_A^{(k)}(t;\beta,\widehat{H}_A,\pi) \\
= & \frac{1}{n}\sum_{j=1}^{n} \frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)} Y_j(t) Z_j^{\otimes k} \exp(\beta^T Z_j) k_j(t-;\beta,H) \\
& - \frac{1}{n}\sum_{j=1}^{n} \frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)} Y_j(t) E\{B_i(t) Z_j^{\otimes k} \exp(\beta^T Z_j) k_j(t-;\beta,H) \mid W_i\} + o_p(1) \\
= & o_p(1). \qquad (5.20)
\end{aligned}$$

As a result, we obtain that

$$\begin{aligned}
& -\frac{\partial}{\partial\alpha} U_{FA}(\beta,\widehat{H}_A,\pi;B,D) \\
= & \frac{1}{n}\sum_{i=1}^{n} \frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)} \int_0^\tau B_i(t)D(t)\{Z_i - T(t;\beta,\widehat{H}_A,\pi)\} dN_i(t) \\
& - \frac{1}{n}\sum_{i=1}^{n} \frac{V_i\dot{\pi}_i(\alpha)}{\pi_i^2(\alpha)} \int_0^\tau \left[E\{B_i(t)D(t)Z_i dN_i(t) \mid W_i\} - T(t;\beta,\widehat{H}_A,\pi) E\{B_i(t)D(t)dN_i(t) \mid W_i\} \right] \\
& + \frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{\pi_i(\alpha)} \int_0^\tau B_i(t)D(t) \frac{\partial}{\partial\alpha} T(t;\beta,\widehat{H}_A,\pi) dN_i(t) \\
& + \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_i(\alpha) - V_i}{\pi_i(\alpha)} \int_0^\tau \frac{\partial}{\partial\alpha} \left[T(t;\beta,\widehat{H}_A,\pi) E\{B_i(t)D(t)dN_i(t) \mid W_i\} \right] \\
\triangleq & I_1 - I_2 + I_3 + I_4,
\end{aligned}$$

where $T(t;\beta,H,\pi) = \{S_A^{(1)}(t;\beta,H,\pi) - R_A^{(1)}(t;\beta,H,\pi)\}/\{S_A^{(0)}(t;\beta,H,\pi) - R_A^{(0)}(t;\beta,H,\pi)\}$. Note that $I_1 - I_2 = o_p(1)$, if the model $f(Z_i^m \mid W_i;\phi)$ is correctly specified. Since $E(V_i \mid W_i) = \pi_i(\alpha)$, we know that $I_4 = o_p(1)$. By (5.19) and (5.20), $(\partial/\partial\alpha)T(t;\beta,\widehat{H}_A,\pi) = o_p(1)$,

implying that $I_3 = o_p(1)$. Therefore,

$$(\partial/\partial\alpha)U_{FA}(\beta, \widehat{H}_A, \pi; B, D) = o_p(1). \tag{5.21}$$

Combining (5.14), (5.15), (5.16), (5.21), we have

$$(-\partial/\partial\theta)U(\theta) = I^{FA} + o_p(1), \tag{5.22}$$

where $I^{FA} = \text{diag}(I_\beta, I_\phi, I_\alpha)$. By the Glivenko-Cantelli theorem, the convergence could be strengthened to the uniform convergence over θ.

Let $I^{\beta\beta}$, $I^{\beta\phi}$ and $I^{\beta\alpha}$ be the corresponding partitions of $(I^{FA})^{-1}$. By the block matrix inverse formula, we have $I^{\beta\beta} = (I_\beta)^{-1}$, $I^{\beta\phi} = 0$ and $I^{\beta\alpha} = 0$. The standard asymptotic expansion yields

$$\begin{aligned} n^{1/2}(\hat{\beta}_{FA}(\hat{\alpha}, \hat{\phi}) - \beta_0) &= n^{1/2} I^{\beta\beta} U_{FA}(\beta_0, \widehat{H}_A, \pi; B, D) + n^{1/2} I^{\beta\phi} U_\phi + n^{1/2} I^{\beta\alpha} U_\alpha + o_p(1) \\ &= n^{1/2} I_\beta^{-1} U_{FA}(\beta_0, \widehat{H}_A, \pi; B, D) + o_p(1). \end{aligned}$$

Following the same arguments, $n^{1/2}(\hat{\beta}_{FA}(\alpha, \phi) - \beta_0)$, $n^{1/2}(\hat{\beta}_{FA}(\hat{\alpha}, \phi) - \beta_0)$ and $n^{1/2}(\hat{\beta}_{FA}(\alpha, \hat{\phi}) - \beta_0)$ can be shown to have the same asymptotic expansions, suggesting that $\hat{\beta}_{FA}(\alpha, \phi)$, $\hat{\beta}_{FA}(\hat{\alpha}, \phi)$ and $\hat{\beta}_{FA}(\alpha, \hat{\phi})$ are asymptotically equivalent to $\hat{\beta}_{FA}(\hat{\alpha}, \hat{\phi})$. The proof of Theorem 5.2 is complete.

6. Further Empirical Studies

6.1. Performance of the Proposed Methods

We adopt the same simulation designs as in Section 7 in the main draft. Two weighting schemes (W1) and (W2) are considered in the main draft. In this supplementary material, we consider another two weighting schemes (W3), $B_i(t) = \pi(t, Z_i^c, 1)$, $D(t) = 1$, and (W4), $B_i(t) = \pi(t, Z_i^c, 1)$, $D(t) = \{\sum_{i=1}^n V_i Y_i(t)\}/\{\sum_{i=1}^n Y_i(t)\}$.

In terms of the treatment of $k_i(t; \beta, H)$, we first consider Huang and Wang's type estimators (2010) with $k_i(t; \beta, H) = 0$. We calculate four estimators respectively based on the full cohort (Full-A), complete-case only (Complete-A), the weighted estimator with true missingness probabilities (WE-A) and the weighted estimator with estimated missingness probabilities under a parametric model (WE-$\hat{\alpha}$-A). Second, based on our estimation

method, we compute estimators respectively based on the full cohort (Full-B), complete-case only (Complete-B), the weighted estimator with true missingness probabilities (WE-B), the weighted estimator with estimated missingness probabilities under a parametric model (WE-$\hat{\alpha}$-B), the weighted estimator with estimated missingness probabilities under a non-parametric model (WE-$\hat{\pi}$-B), the fully augmented weighted estimator with true missingness probabilities (FAW-B) and the fully augmented weighted estimator with estimated missingness probabilities under a parametric model (FAW-$\hat{\alpha}$-B). As in Section 6 of the main draft, the same logistic model and the kernel function are employed for WE-$\hat{\alpha}$-B and WE-$\hat{\pi}$-B.

Tables 1-3 show the bias, empirical standard error, model-based standard error and 95% coverage rate for various estimators, with the sample size $n = 100$.

Table 4 shows the bias, empirical standard error, model-based standard error and 95% coverage rate for various estimators, under the first simulation scenario with $n = 200$. In this setting, the model-based standard error (MSE) is very close to the empirical standard error (ESE), with a percentage bias of less than 10%, except for the complete case analysis. Moreover, the 95% coverage rate for our estimator is all between 93%-95% when the missingness probability is known or estimated parametrically. The estimator WE-$\hat{\pi}$-B sometimes may be unstable, due to the nonparametric estimation of the missingness probability. But its 95% coverage rate is still reasonable (between 91%-93%). This suggests that our variance estimator is accurate, for moderate sample sizes.

6.2. Sensitivity Analysis under Misspecification of $\pi_i(\alpha)$

We consider the same simulation scenario as in Section 7.2 of the main draft to evaluate the sensitivity of various estimators under misspecification of $\pi_i(\alpha)$.

Figure 1 shows the relative bias of estimators using weight (W3) (WE-$\hat{\alpha}$-A-W3, WE-$\hat{\alpha}$-B-W3) and weight (W4) (WE-$\hat{\alpha}$-A-W4, WE-$\hat{\alpha}$-B-W4) averaged over 100 replications. All estimators have little bias when η is close to 0. As expected, the magnitude of the bias of estimators increases as η further departs from 0. The estimators of β_2 seem less sensitive with respect to the specification of $\pi_i(\alpha)$ than those of β_1. In particular, the estimator WE-$\hat{\alpha}$-A-W4 shows largest bias for estimating β_1 and β_2 among the estimators we consider.

References

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

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

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association* **72** 147–148.

Huang, B. and Wang, Q. (2010). Semiparametric analysis based on weighted estimating equations for transformation models with missing covariates. *Journal of Multivariate Analysis* **101** 2078–2090.

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

Van der Vaart, A. V. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

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

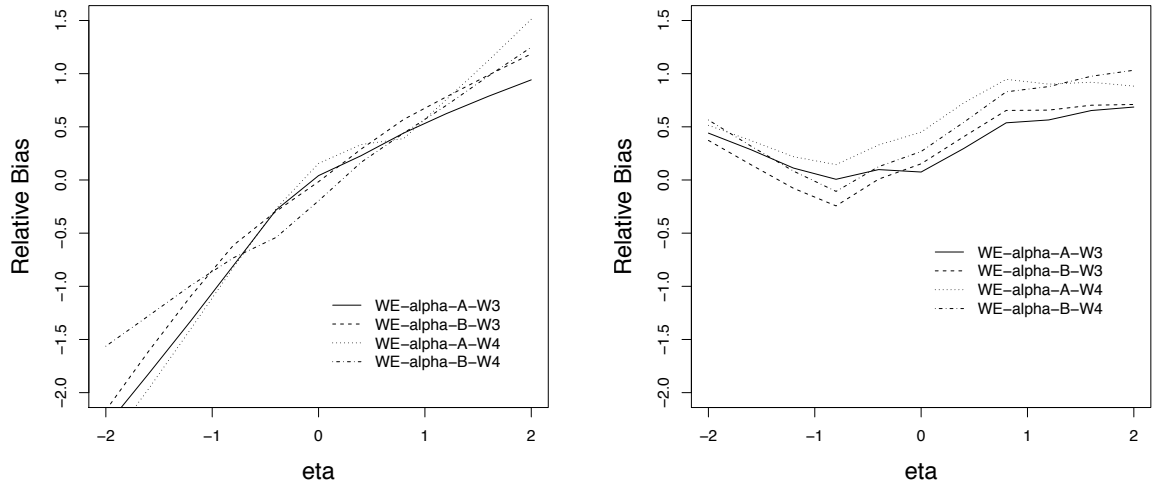

FIG 1. *Effects of misspecifying $\pi_i(\alpha)$ on the proposed estimators for β_1 and β_2 under weights (W3) and (W4). The left panel is the averaged relative bias for estimation of β_1 and the right panel is for β_2.*

TABLE 1

Simulation results for the first scenario with $n = 100$: Bias, empirical standard error (ESE), model-based standard error (MSE) and 95% coverage rate (CR, in percent) of the proposed estimators using weights (W3) and (W4)

		β_1				β_2			
Weight	Method	Bias	ESE	MSE	CR	Bias	ESE	MSE	CR
(W3)	Full-A	-0.048	0.38	0.35	92	0.032	0.25	0.22	93
	Complete-A	-0.085	0.46	0.42	88	0.048	0.28	0.25	90
	WE-A	-0.059	0.52	0.48	91	0.034	0.32	0.28	92
	WE-$\hat{\alpha}$-A	-0.052	0.50	0.48	93	0.035	0.29	0.26	92
	Full-B	-0.039	0.37	0.34	93	0.035	0.22	0.19	94
	Complete-B	-0.071	0.43	0.38	87	0.052	0.26	0.23	88
	WE-B	-0.052	0.44	0.41	92	0.031	0.28	0.25	93
	WE-$\hat{\alpha}$-B	-0.043	0.42	0.40	92	0.028	0.26	0.24	92
	WE-$\hat{\pi}$-B	-0.048	0.42	0.38	91	0.034	0.25	0.24	92
	FAW-B	-0.039	0.41	0.39	93	0.030	0.25	0.23	92
	FAW-$\hat{\alpha}$-B	-0.042	0.42	0.39	93	0.032	0.25	0.22	92
(W4)	Full-A	-0.056	0.38	0.36	94	0.054	0.26	0.23	93
	Complete-A	-0.088	0.46	0.37	74	0.075	0.31	0.25	84
	WE-A	-0.067	0.49	0.46	92	0.052	0.35	0.31	92
	WE-$\hat{\alpha}$-A	-0.061	0.49	0.45	91	0.057	0.32	0.28	91
	Full-B	-0.042	0.38	0.35	94	0.041	0.23	0.20	94
	Complete-B	-0.074	0.43	0.35	88	0.053	0.27	0.22	85
	WE-B	-0.052	0.45	0.42	92	0.040	0.29	0.26	92
	WE-$\hat{\alpha}$-B	-0.050	0.43	0.40	92	0.042	0.27	0.24	92
	WE-$\hat{\pi}$-B	-0.051	0.44	0.38	91	0.044	0.29	0.23	90
	FAW-B	-0.050	0.43	0.41	94	0.043	0.26	0.23	94
	FAW-$\hat{\alpha}$-B	-0.053	0.42	0.39	93	0.041	0.27	0.25	95

TABLE 2

Simulation results for the second scenario with $n = 100$: Bias, empirical standard error (ESE), model-based standard error (MSE) and 95% coverage rate (CR, in percent) of proposed estimators using weights (W3) and (W4)

		β_1				β_2			
Weight	Method	Bias	ESE	MSE	CR	Bias	ESE	MSE	CR
(W3)	Full-A	-0.065	0.53	0.49	92	-0.024	0.28	0.26	94
	Complete-A	-0.090	0.64	0.55	81	-0.043	0.35	0.29	90
	WE-A	-0.058	0.57	0.53	92	0.027	0.31	0.27	92
	WE-$\hat{\alpha}$-A	-0.055	0.56	0.52	93	0.030	0.27	0.24	93
	Full-B	-0.065	0.44	0.41	94	0.021	0.22	0.21	95
	Complete-B	-0.088	0.55	0.48	88	0.050	0.29	0.24	84
	WE-B	-0.052	0.51	0.48	93	0.018	0.26	0.25	94
	WE-$\hat{\alpha}$-B	-0.045	0.49	0.48	94	0.020	0.24	0.22	93
	WE-$\hat{\pi}$-B	-0.042	0.50	0.45	91	0.026	0.25	0.22	92
	FAW-B	-0.046	0.48	0.45	94	0.022	0.24	0.21	95
	FAW-$\hat{\alpha}$-B	-0.042	0.48	0.47	95	0.023	0.23	0.23	94
(W4)	Full-A	-0.067	0.56	0.52	91	0.034	0.33	0.31	94
	Complete-A	-0.088	0.68	0.59	83	0.056	0.42	0.30	84
	WE-A	-0.051	0.59	0.55	92	0.026	0.36	0.33	93
	WE-$\hat{\alpha}$-A	-0.053	0.56	0.54	94	0.034	0.32	0.30	93
	Full-B	-0.055	0.43	0.42	95	0.027	0.25	0.23	94
	Complete-B	-0.078	0.58	0.53	90	0.062	0.35	0.30	88
	WE-B	-0.043	0.52	0.50	94	0.028	0.29	0.26	92
	WE-$\hat{\alpha}$-B	-0.038	0.49	0.48	95	0.025	0.26	0.23	93
	WE-$\hat{\pi}$-B	-0.040	0.50	0.46	92	0.024	0.25	0.23	93
	FAW-B	-0.041	0.48	0.47	95	0.025	0.24	0.22	94
	FAW-$\hat{\alpha}$-B	-0.037	0.48	0.45	93	0.026	0.24	0.23	95

TABLE 3

Simulation results for the third scenario with $n = 100$: Bias, empirical standard error (ESE), model-based standard error (MSE) and 95% coverage rate (CR, in percent) of the proposed estimators using weights (W3) and (W4)

		β_1				β_2			
Weight	Method	Bias	ESE	MSE	CR	Bias	ESE	MSE	CR
(W3)	Full-A	0.023	0.56	0.52	94	0.038	0.29	0.26	94
	Complete-A	0.091	0.65	0.59	79	0.058	0.32	0.25	85
	WE-A	0.029	0.64	0.62	92	0.053	0.42	0.39	93
	WE-$\hat{\alpha}$-A	0.031	0.60	0.56	92	0.035	0.37	0.35	93
	Full-B	0.025	0.44	0.40	92	0.044	0.22	0.21	94
	Complete-B	0.086	0.52	0.47	86	0.068	0.33	0.27	88
	WE-B	0.038	0.58	0.56	93	0.034	0.32	0.28	92
	WE-$\hat{\alpha}$-B	0.036	0.54	0.52	93	0.035	0.30	0.27	93
	WE-$\hat{\pi}$-B	0.032	0.53	0.50	92	0.032	0.32	0.28	91
	FAW-B	0.035	0.53	0.51	93	0.036	0.31	0.30	94
	FAW-$\hat{\alpha}$-B	0.034	0.52	0.51	95	0.034	0.32	0.29	93
(W4)	Full-A	0.027	0.57	0.53	93	0.034	0.29	0.27	94
	Complete-A	0.061	0.58	0.52	88	0.060	0.36	0.31	87
	WE-A	0.032	0.68	0.63	91	0.038	0.40	0.36	92
	WE-$\hat{\alpha}$-A	0.036	0.62	0.59	92	0.040	0.37	0.36	94
	Full-B	0.039	0.40	0.38	94	0.047	0.24	0.22	93
	Complete-B	0.068	0.46	0.39	85	0.078	0.34	0.26	87
	WE-B	0.036	0.56	0.53	93	0.048	0.38	0.35	93
	WE-$\hat{\alpha}$-B	0.045	0.51	0.48	93	0.048	0.35	0.33	93
	WE-$\hat{\pi}$-B	0.038	0.50	0.44	90	0.052	0.34	0.30	91
	FAW-B	0.043	0.52	0.49	93	0.046	0.34	0.32	93
	FAW-$\hat{\alpha}$-B	0.042	0.52	0.50	93	0.050	0.34	0.33	94

TABLE 4

Simulation results for the first scenario with $n = 200$: Bias, empirical standard error (ESE), model-based standard error (MSE) and 95% coverage rate (CR, in percent) of the proposed estimators using weights (W1) and (W2)

		β_1				β_2			
Weight	Method	Bias	ESE	MSE	CR	Bias	ESE	MSE	CR
(W1)	Full-A	-0.023	0.31	0.30	94	0.014	0.20	0.19	94
	Complete-A	-0.048	0.43	0.39	90	0.037	0.24	0.20	88
	WE-A	-0.031	0.37	0.35	93	0.020	0.26	0.24	93
	WE-$\hat{\alpha}$-A	-0.024	0.35	0.36	93	0.023	0.24	0.22	92
	Full-B	-0.020	0.26	0.25	94	0.018	0.18	0.17	95
	Complete-B	-0.055	0.35	0.30	89	0.041	0.28	0.22	86
	WE-B	-0.025	0.33	0.32	94	0.019	0.22	0.23	93
	WE-$\hat{\alpha}$-B	-0.028	0.30	0.28	93	0.017	0.21	0.20	93
	WE-$\hat{\pi}$-B	-0.032	0.32	0.29	91	0.021	0.20	0.22	92
	FAW-B	-0.025	0.31	0.29	94	0.016	0.20	0.21	94
	FAW-$\hat{\alpha}$-B	-0.028	0.31	0.30	94	0.018	0.19	0.20	94
(W2)	Full-A	-0.032	0.35	0.33	93	0.041	0.21	0.20	94
	Complete-A	-0.088	0.41	0.37	88	0.063	0.23	0.20	90
	WE-A	-0.043	0.44	0.41	92	0.029	0.29	0.27	93
	WE-$\hat{\alpha}$-A	-0.035	0.41	0.39	93	0.032	0.27	0.25	93
	Full-B	-0.041	0.31	0.32	94	0.038	0.19	0.18	94
	Complete-B	-0.084	0.36	0.32	89	0.067	0.23	0.19	85
	WE-B	-0.035	0.39	0.37	93	0.035	0.24	0.23	94
	WE-$\hat{\alpha}$-B	-0.034	0.36	0.36	95	0.039	0.21	0.20	94
	WE-$\hat{\pi}$-B	-0.038	0.37	0.34	92	0.041	0.21	0.19	93
	FAW-B	-0.034	0.37	0.36	94	0.035	0.21	0.21	94
	FAW-$\hat{\alpha}$-B	-0.032	0.36	0.35	94	0.036	0.21	0.20	94