Published in final edited form as:
Scand Stat Theory Appl. 2012 March ; 39(1): 34–52. doi:10.1111/j.1467-9469.2011.00746.x.

Estimation of Stratified Mark-Specific Proportional Hazards Models with Missing Marks

Yanqing Sun
Department of Mathematics and Statistics, The University of North Carolina at Charlotte, Charlotte, NC 28223, USA. yasun@uncc.edu

Peter B. Gilbert
Department of Biostatistics, University of Washington and Fred Hutchinson Cancer Research Center

Abstract

An objective of randomized placebo-controlled preventive HIV vaccine efficacy trials is to assess the relationship between the vaccine effect to prevent infection and the genetic distance of the exposing HIV to the HIV strain represented in the vaccine construct. Motivated by this objective, recently a mark-specific proportional hazards model with a continuum of competing risks has been studied, where the genetic distance of the transmitting strain is the continuous `mark' defined and observable only in failures. A high percentage of genetic marks of interest may be missing for a variety of reasons, predominantly due to rapid evolution of HIV sequences after transmission before a blood sample is drawn from which HIV sequences are measured. This research investigates the stratified mark-specific proportional hazards model with missing marks where the baseline functions may vary with strata. We develop two consistent estimation approaches, the first based on the inverse probability weighted complete-case (IPW) technique, and the second based on augmenting the IPW estimator by incorporating auxiliary information predictive of the mark. We investigate the asymptotic properties and finite-sample performance of the two estimators, and show that the augmented IPW estimator, which satisfies a double robustness property, is more efficient.

Keywords

Augmented inverse probability weighted complete-case estimator; auxiliary marks; competing risks; double robustness; failure time data; genetic data; HIV vaccine trial; mark-specific vaccine efficacy; missing at random; semiparametric model

1 Introduction

In many medical studies, evaluation of treatment efficacy is based on comparison of survival times between two treatment groups. This is often done through the ratio of two hazard functions of an endpoint. In a preventive HIV vaccine efficacy trial, the usual primary objective is to assess vaccine efficacy (VE) to prevent HIV infection, where typically VE is defined as one minus the hazard ratio (vaccine/placebo) of the failure event. However, it may be quite difficult to achieve an efficacious vaccine due to genetic variation of HIV (Fauci *et al.*, 2008). The study population is exposed to many genetic types of circulating

Supplementary Materials Web Appendices A and B referenced in this article are available at the website of the Scandinavian Journal of Statistics, http://onlinelibrary.wiley.com/journal.

HIVs but the vaccine only contains antigens based on one or a few strains, and the vaccine protection is likely to be lower against strains that are not in the vaccine or that have greater genetic distance from the strain(s) in the vaccine (Gilbert *et al.*, 1999). Assessment of this problem can be formulated with a competing risks failure time model where the endpoint is HIV infection and the mark variable is the genetic type or distance of the transmitting strain(s) from the strain(s) represented in the vaccine (A `mark' by definition is only meaningful/observable in failures). The goal is to evaluate mark-specific vaccine efficacy defined as one minus the mark-specific hazard ratio (vaccine/placebo) of infection.

In HIV vaccine efficacy trials, the mark variable of interest is some measure of genetic distance between an HIV sequence measured from an infected subject and an HIV sequence represented in the vaccine. For example, if the two HIV sequences are aligned, then the genetic distance may be defined as the weighted percent mismatch of amino acids (e.g., Wu *et al.*, 2001), with weights selected to make the distance scientifically/immunologically meaningful. Definitions of genetic distance between unaligned sequences are also possible, for example based on the difference in total counts of amino acids that possess certain physico-chemical properties. For many definitions of genetic distances of interest, the value of the distance may be unique for almost all infected subjects, making it natural to consider it as a continuous mark variable. Gilbert *et al.* (2004, 2008) developed procedures for estimation and testing of continuous mark-specific hazards rates and relative risks, and Sun *et al.* (2009) studied the continuous mark-specific proportional hazards model, which allows evaluating mark-specific vaccine efficacy with adjustment for covariate effects.

However, the previous work for a continuous mark did not account for missingness of the mark variable in failures, so that its valid implementation relies on a dubious missing completely at random assumption (MCAR). While these methods will perform satisfactorily for mark variables measured in almost all failures, for which violations of MCAR are inconsequential, they may be invalid and misleading under moderate-to-high rates of missingness. Moreover, arguably the marks of greatest scientific interest in HIV vaccine trials are subject to high rates of missingness. In particular, ideally the mark would be defined based on the HIV strain that is actually transmitted. However, because vaccine trials only periodically test for HIV infection, the earliest observable HIV genotype is measured using a subjects' earliest available blood sample, which is drawn days, weeks, or several months after the actual transmission event. Therefore the ideal mark definition is unachievable, and we must consider knowledge of HIV evolution and the HIV testing algorithm to develop a workable definition.

HIV vaccine trials test volunteers for anti-HIV antibodies at periodic intervals (e.g., every 3 or 6 months); these antibody-based tests have near-perfect sensitivity to detect infections that occurred at least 4 weeks ago but otherwise may miss the infection. For all subjects with an HIV antibody positive (Ab+) test, a “look-back” procedure is applied wherein earlier available blood samples are tested for HIV infection using a more sensitive antigen-based HIV-specific PCR assay, which has near-perfect sensitivity if the infection occurred at least one week ago; thus for example if blood samples are drawn 2 weeks and 6 weeks after HIV acquisition, and the Ab and PCR tests are applied to both samples, then with very high probability the 2-week sample will be Ab−, PCR+ and the 6-week sample will be Ab+, PCR +. Therefore, each infected subject is classified into one of two groups, defined by whether the earliest sample is Ab− and PCR+ (an `early' sample) or is Ab+ and PCR+ (a `later' sample). These groups approximately correspond to the first available sample being drawn less than or greater than 4 weeks after HIV acquisition. Extensive analysis of early viruses suggests that these viruses well-approximate the transmitting strain, whereas later viruses tend to have undergone significant evolution, with many genetic mutations (Keele *et al.*, 2008). If the mark is defined based on the earliest sampled virus from each infected subject,

then the fact that it is often significantly mutated relative to the transmitted virus may make the analysis misleading about the relationship between vaccine efficacy and the exposing virus. One remedy to this problem defines the mark as the early virus in the Ab− and PCR+ phase, which will likely be missing from many infected subjects, with missingness rate depending on the frequency of HIV testing. For the recent Step HIV vaccine trial with genetic data on 87 diagnosed HIV infections, with 6-monthly HIV antibody testing, 41 were Ab−/PCR+ viruses, for a missingness rate of 53% (unpublished data). The sequel trial that is ongoing (named HVTN 505), with 3-monthly HIV antibody testing, is expected to give a missingness rate of less than 50%. The `early' mark of interest may be missing for other reasons besides HIV evolution, for example because of a missing blood sample or a technical problem in measuring the HIV sequence; the developed methods accommodate this issue by allowing separate models of the different missingness types.

To analyze data with the mark defined as the genetic distance from the early virus, a `complete-case' application of the previously developed methods, which excludes infected subjects without a measured mark, may be highly biased and inefficient. Accordingly, this article develops valid and more efficient estimation methods for data-sets with missing at random continuous marks. The new methods restrict attention to a univariate mark variable, albeit if the vaccine contains multiple HIV sequences and/or if multiple sequences are transmitted to a study participant (which happens in a minority of cases, Keele *et al.*, 2008), then it would be of interest to study a multivariate mark variable defined as the set of genetic distances to the vaccine sequences. The current methods can accommodate multiple genetic distances by defining the univariate mark as the minimum of the multiple distances.

Early work in nonparametric estimation for failure time data with missingness of a discrete mark (failure type) include Dinse (1986), Racine-Poon & Hoel (1984) and Lagakos & Louis (1988). Goetghebeur & Ryan (1990) and Dewanji (1992) derived modified logrank tests to compare survival experience between two groups. Goetghebeur & Ryan (1995) and Lu & Tsiatis (2001) studied proportional cause-specific hazards models for one particular failure cause. Gao & Tsiatis (2005) studied linear transformation models of cause-specific hazard functions and Lu & Liang (2008) studied semiparametric additive models of cause-specific hazard functions.

We develop two estimation approaches for the stratified mark-specific proportional hazards model. The first uses inverse probability weighting (IPW) of the complete-case estimator, which leverages auxiliary predictors of whether the mark is observed. The second approach, adapting the theory of Robins *et al.* (1994), augments the IPW complete-case estimator with auxiliary predictors of the missing marks.

The article is organized as follows. Notation, assumptions, and the stratified mark-specific proportional hazards model are introduced in Section 2. The estimation procedures are proposed in Section 3. The asymptotic properties of the proposed estimators are derived in Section 4. Procedures for constructing confidence intervals for mark-specific vaccine efficacies are also given in Section 4. The finite-sample performance and robustness analysis of the estimators are studied in simulations in Section 5. Proofs of the main results are placed in the web Appendices.

2 Stratified mark-specific proportional hazards models and missing marks

Let T be the failure time, V a continuous mark variable, and $Z(t)$ a possibly time-dependent p-dimensional covariate. Under the competing risks model, the mark V is only defined and observable when T is observed, whereas if T is right-censored, the mark is undefined and meaningless. Suppose that the conditional mark-specific hazard function at time t given the

covariate history $Z(s)$, for $s \le t$, defined below only depends on the current value $Z(t)$. The conditional mark-specific hazard function given $Z(t) = z$ is defined as

$$\lambda(t, v|z) = \lim_{h_1, h_2 \to 0} P\{T \in [t, t+h_1), V \in [v, v+h_2) | T \ge t, Z(t) = z\} / h_1 h_2 \quad (1)$$

with t ranging over a fixed interval $[0, \tau]$. Assume that V is continuous and has a known bounded support; rescaling V if necessary, this support is taken to be [0, 1]. The function given in (1) is the natural analog of its discrete counterpart, with similar interpretation.

Gilbert *et al.* (2008) defined mark-specific vaccine efficacy as $VE(t, v) = 1 - \lambda(t, v|z = 1)/\lambda(t, v|z = 0)$, with z the indicator of assignment to the vaccine group; they developed several nonparametric and semiparametric tests concerning $VE(t, v)$ as well as a nonparametric estimation method. Sun *et al.* (2009) studied the mark-specific proportional hazards (PH) model, $\lambda(t, v|z(t)) = \lambda_0(t, v)\exp\{\beta(v)^T z(t)\}$, where $\lambda_0(\cdot, v)$ is the unspecified baseline hazard function and $\beta(v)$ is the p-dimensional unknown regression coefficient function of v. For the HIV vaccine trial application, the covariate is partitioned as $z(t) = (z_1, z_2(t))^T$, where z_1 is the treatment (vaccine) group indicator and $z_2(t)$ is a vector of possibly time-dependent covariates. Then the covariate-adjusted mark-specific vaccine efficacy defined above takes the simpler form $VE(v) = 1 - \exp(\beta_1(v))$, without any dependence on t. Assuming the vaccine and placebo groups have the same distribution of the number of exposures to viruses with genetic distance marks in a neighborhood of v (by randomization and double-blinding), $VE(v)$ approximates the multiplicative effect of the vaccine to reduce the instantaneous rate of infection with this virus type (Gilbert *et al.*, 2008; Sun *et al.*, 2009). The statistical procedures developed by Sun *et al.* (2009) are based on observations of the random variables $(X, Z(\cdot), V)$ for $\delta = 1$ and $(X, Z(\cdot))$ for $\delta = 0$, where $X = \min\{T, C\}$, $\delta = I(T \le C)$, and C is a censoring random variable. Whereas the earlier work assumed V was observed whenever T is observed, here we allow V to be missing, and incorporate in the estimation procedures auxiliary covariates and/or auxiliary mark variables that inform about the probability V is observed and about the distribution of V. For our motivating example discussed above, genetic data for the `later' virus measured in almost all infected subjects is a natural auxiliary mark to help predict the genetic distance $V$ for an `early' virus that is frequently missing.

In practice, different key subgroups (e.g., men and women; subjects living in different geographic regions) typically have different baseline mark-specific hazards of HIV infection. The stratified mark-specific PH model postulates that the conditional mark-specific hazard function given covariate $z(t)$ for an individual in the kth stratum equals

$$\lambda_k(t, v|z(t)) = \lambda_{0k}(t, v)\exp\left\{\beta(v)^T z(t)\right\}, \quad k = 1, \ldots, K, \quad (2)$$

where $\beta(v)$ is the p-dimensional unknown regression coefficient function of v, $\lambda_{0k}(\cdot, v)$ is the unspecified baseline hazard function for the kth stratum and K is the number of strata. Model (2) allows different baseline functions for different strata. Similar generalizations of the Cox model was studied by Dabrowska (1997).

Next we introduce some notation and assumptions that are used throughout the article. Let n_k be the number of subjects in the kth stratum; the total sample size is $n = \sum_{k=1}^{K} n_k$. As above, the right-censored mark-specific failure time is represented by (X, δ, V) and $Z(\cdot)$ is the covariate process. Let R be the indicator of whether all possible data are observed for a subject; $R = 1$ if either $\delta = 0$ (right-censored) or if $\delta = 1$ and V is observed; and $R = 0$ otherwise. Auxiliary variables A may be helpful for predicting missing marks. Since the mark can only be missing for failures, supplemental information is potentially useful only for failures, for predicting missingness and for informing about the distribution of missing

marks. For example, the `later' HIV virus V^* can be considered a subset of A. In general, A could include multiple viral sequences per infected subject at multiple time-points along with the time points at which the viruses are sequenced, giving a picture of the intra-subject evolution of a subject's HIV infection. The relationship between A and V can be modelled to help predict V (see Section 5 for a simulation example).

We assume that the censoring time C is conditionally independent of (T, V) given $Z(\cdot)$ for an individual in the kth stratum. We also assume the mark V is missing at random (Rubin, 1976); that is, given $\delta = 1$ and $W = (T, Z(T), A)$ of an individual in the kth stratum, the probability that the mark V is missing depends only on the observed W, not on the value of V; this assumption is expressed as

$$r_k(W) \equiv P(R=1|\delta=1, W) = P(R=1|V, \delta=1, W). \quad (3)$$

Let $\pi_k(Q) = P(R = 1|Q)$ where $Q = (\delta, W)$. Then $\pi_k(Q) = \delta r_k(W) + (1 - \delta)$. The missing at random assumption (3) also implies that V is independent of R given Q:

$$\rho_k(v, W) \equiv P(V \le v|\delta=1, W) = P(V \le v|R=1, \delta=1, W). \quad (4)$$

For an observed value w of W of an individual in the kth stratum, we write $r_k(w) = P(R = 1|\delta = 1, W = w)$ and $\rho_k(v, w) = P(V \le v|\delta = 1, W = w)$. The stratum-specific definitions of $r_k(w)$ and $\rho_k(v, w)$ leave the options for the models of the probability of complete-case and mark distribution to be different for different strata.

Let for each k, $\{X_{ki}, Z_{ki}(\cdot), \delta_{ki}, R_{ki}, V_{ki}, A_{ki}; i = 1, \ldots, n_k\}$ be iid replicates of $\{X, Z(\cdot), \delta, R, V, A\}$ from that stratum. The observed data is denoted by $\{O_{ki}; i = 1 \ldots, n_k, k = 1, \ldots, K\}$, where $O_{ki} = \{X_{ki}, Z_{ki}(\cdot), R_{ki}, R_{ki}V_{ki}, A_{ki}\}$ for $\delta_{ki} = 1$ and $O_{ki} = \{X_{ki}, Z_{ki}(\cdot), R_{ki} = 1\}$ for $\delta_{ki} = 0$. We assume that $\{O_{ki}; i = 1 \ldots, n_k, k = 1, \ldots, K\}$ are independent for all subjects. Similarly, we denote $W_{ki} = (T_{ki}, Z_{ki}(T_{ki}), A_{ki})$ and $Q_{ki} = (\delta_{ki}, W_{ki})$.

We consider a parametric model $r_k(W_{ki}, \psi_k)$ for $r_k(W_{ki})$, where ψ_k is an unknown vector of parameters to be further discussed in the next section. Let $\pi_k(Q_{ki}, \psi_k) = \delta_{ki} r_k(W_{ki}, \psi_k) + (1 - \delta_{ki})$. Additional notation is introduced in the following. For $\beta \in \mathbb{R}^p$, $t \ge 0$, let $Y_{ki}(t) = I(X_{ki} \ge t)$,

$$S_k^{(j)}(t, \beta) = n_k^{-1} \sum_{i=1}^{n_k} Y_{ki}(t) \exp\left\{\beta^T Z_{ki}(t)\right\} Z_{ki}(t)^{\otimes j},$$

$$\tilde{S}_k^{(j)}(t, \beta, \psi_k) = n_k^{-1} \sum_{i=1}^{n_k} R_{ki}(\pi_k(Q_{ki}, \psi_k))^{-1} Y_{ki}(t) \exp\left\{\beta^T Z_{ki}(t)\right\} Z_{ki}(t)^{\otimes j},$$

where for any $z \in \mathbb{R}^p$, $z^{\otimes 0} = 1$, $z^{\otimes 1} = z$ and $z^{\otimes 2} = zz^T$. Define $s_k^{(j)}(t, \beta) = ES_k^{(j)}(t, \beta)$ and $\tilde{s}_k^{(j)}(t, \beta, \psi_k) = E\tilde{S}_k^{(j)}(t, \beta, \psi_k)$. Under the missing at random assumption (3), $s_k^{(j)}(t, \beta) = \tilde{s}_k^{(j)}(t, \beta, \psi_k)$ if the model $r_k(W_{ki}, \psi_k)$ is correctly specified. Let

$$J_k(t, \beta) = \frac{S_k^{(2)}(t, \beta)}{S_k^{(0)}(t, \beta)} - \left(\frac{S_k^{(1)}(t, \beta)}{S_k^{(0)}(t, \beta)}\right)^{\otimes 2},$$

$$\tilde{J}_k(t,\beta,\psi_k)=\frac{\tilde{S}_k^{(2)}(t,\beta,\psi_k)}{\tilde{S}_k^{(0)}(t,\beta,\psi_k)}-\left(\frac{\tilde{S}_k^{(1)}(t,\beta,\psi_k)}{\tilde{S}_k^{(0)}(t,\beta,\psi_k)}\right)^{\otimes 2},$$

$$\bar{Z}_k(t,\beta)=\frac{S_k^{(1)}(t,\beta)}{S_k^{(0)}(t,\beta)},\quad \tilde{Z}_k(t,\beta,\psi_k)=\frac{\tilde{S}_k^{(1)}(t,\beta,\psi_k)}{\tilde{S}_k^{(0)}(t,\beta,\psi_k)}.$$

Let $\bar{z}_k(t,\beta)=s_k^{(1)}(t,\beta)/s_k^{(0)}(t,\beta)$ and $I_k(t,\beta)=s_k^{(2)}(t,\beta)/s_k^{(0)}(t,\beta)-\left(\bar{z}_k(t,\beta)\right)^{\otimes 2}$.

3 Estimation procedures

When there are no missing marks, $\beta(v)$ in model (2) can be estimated by maximizing the following local log-partial likelihood function for $\beta = \beta(v)$ at a fixed v:

$$l(v,\beta)=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^1\int_0^\tau K_h(u-v)\left[\beta^T Z_{ki}(t)-\log\left(\sum_{j=1}^{n_k}Y_{kj}(t)\,e^{\beta^T Z_{kj}(t)}\right)\right]N_{ki}(dt,du), \quad (5)$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a kernel function with support $[-1, 1]$, τ is the end of the follow-up period and $h = h_n$ is a bandwidth. Here $N_{ki}(t, v) = I(X_{ki} \le t, \delta_{ki} = 1, V_{ki} \le v)$ is the marked counting process with a jump at the uncensored failure time X_{ki} and the associated mark V_{ki}. The log partial likelihood function (5) resembles that of Kalbfleisch & Prentice (1980) in the case of discrete marks, except that it borrows strength from observations having marks in the neighborhood of v. The kernel function is designed to give greater weight to observations with marks near v than those further away. As discussed by Silverman (1986, p.43), the choice of kernel function has little to do with the performance of the estimators. For example, the asymptotic relative efficiency of the Tukey kernel compared to the optimal Epanechnikov kernel is 0.9897 and the Gaussian kernel has a relative efficiency of 0.9512. It is common to assume compact support for technical simplicity. The bandwidth, on the other hand, is an important parameter in the estimation of $\beta(v)$. The cross-validation method is often used for bandwidth selection. This is further discussed in Section 5.

Taking the derivative of $l(v, \beta)$ with respect to β gives the score function

$$U(v,\beta)=l'_\beta(v,\beta)=\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^1\int_0^\tau K_h(u-v)\left(Z_{ki}(t)-\frac{S_k^{(1)}(t,\beta)}{S_k^{(0)}(t,\beta)}\right)N_{ki}(dt,du). \quad (6)$$

The maximum partial likelihood estimator is a solution to $U(v, \beta) = 0$. This estimator was studied by Sun *et al.* (2009) when there is no stratification, i.e. $K = 1$.

In the presence of missing marks, (5) can only be applied directly to estimate $\beta(v)$ through a complete-case analysis (e.g., excluding failures with a missing mark), which may be biased and inefficient. Our first approach to remedying this problem, following the idea of Horvitz & Thompson (1952), uses inverse probability weighting of complete-cases. However, this approach has been shown to be inefficient in several situations (Gao & Tsiatis, 2005; Lu & Liang, 2008); also see Scharfstein *et al.* (1999). Our second approach, adapting the idea of Robins *et al.* (1994), augments the inverse probability weighted complete-case estimation

equation with a consistent estimator of the conditional distribution of the mark that uses the estimator from the first approach and auxiliary data from subjects with missing marks.

Our model formulation is different from those in the existing literature on the discrete competing risks model, where only the cause-specific hazard function for one particular cause is modeled while the cause-specific hazard functions for the other causes are left unspecified. To evaluate mark-specific vaccine effects, model (2) specifies a proportional mark-specific hazards model for each mark v, $0 \le v \le 1$. This model induces restrictions on the conditional mark distribution $\rho_k(v, W_k)$ given in (4), so that arbitrary modeling may run into conflicts with the mark-specific PH model (2) and result in inconsistent estimation. This modeling restriction requires a different approach, as described below.

3.1 Inverse probability weighted complete-case estimator

Following Horvitz & Thompson (1952), inverse probability weighting of complete-cases has been commonly used in missing data problems. Let $r_k(W_{ki}, \psi_k)$ be the parametric model for the probability of complete-case $r_k(W_{ki})$ defined in (3), where ψ_k is a q-dimensional parameter. For example, one can assume the logistic model with $\text{logit}(r_k(W_{ki}, \psi_k)) = \psi_k^T W_{ki}$ for those with $\delta_{ki} = 1$. By (3), the maximum likelihood estimator $\widehat{\psi} = (\widehat{\psi}_1, \ldots, \widehat{\psi}_K)$ of $\psi = (\psi_1, \ldots, \psi_K)$ is obtained by maximizing the observed data likelihood,

$$\prod_{k,i} \{r_k(W_{ki}, \psi_k)\}^{R_{ki}\delta_{ki}} \{1 - r_k(W_{ki}, \psi_k)\}^{(1-R_{ki})\delta_{ki}}. \quad (7)$$

We propose the following inverse probability weighted (IPW) estimating equation for β:

$$U_{ipw}(v, \beta, \widehat{\psi}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau K_h(u - v) \left(Z_{ki}(t) - \tilde{Z}_k(t, \beta, \widehat{\psi}_k)\right) \frac{R_{ki}}{\pi_k(Q_{ki}, \widehat{\psi}_k)} N_{ki}(dt, du). \quad (8)$$

The IPW estimator of $\beta(v)$ solves the above equation and is denoted by $\widehat{\beta}^{ipw}(v)$.

The baseline function $\lambda_{0k}(t, v)$ can be estimated by $\widehat{\lambda}_{0k}^{ipw}(t, v)$, obtained by smoothing the increments of the following estimator of the doubly cumulative baseline function $\Lambda_{0k}(t, v) = \int_0^t \int_0^v \lambda_{0k}(s, u)\, ds\, du$:

$$\widehat{\Lambda}_{0k}^{ipw}(t, v) = \sum_{i=1}^{n_k} \int_0^t \int_0^v \frac{R_{ki}}{\pi_k(Q_{ki}, \widehat{\psi}_k)} \frac{N_{ki}(ds, du)}{n_k \tilde{S}_k^{(0)}(s, \widehat{\beta}^{ipw}(u), \widehat{\psi}_k)}. \quad (9)$$

For example, one can use kernel smoothing,

$$\widehat{\lambda}_{0k}^{ipw}(t, v) = \int_0^\tau \int_0^1 K_{b_1}^{(1)}(t - s)\, K_{b_2}^{(2)}(v - u)\, \widehat{\Lambda}_{0k}^{ipw}(ds, du), \quad (10)$$

where $K_{b_1}^{(1)}(x) = K^{(1)}(x/b_1)/b_1$ and $K_{b_2}^{(2)}(x) = K^{(2)}(x/b_2)/b_2$, with $K^{(1)}(\cdot)$ and $K^{(2)}(\cdot)$ the kernel functions and b_1 and b_2 the bandwidths.

3.2 Augmented inverse probability weighted complete-case estimator

The IPW estimator $\widehat{\beta}^{ipw}(v)$ uses only complete cases and is inefficient. Following Robins *et al.* (1994), Gao & Tsiatis (2005) proposed augmented inverse probability weighted complete-case estimators for the linear transformation model of cause-specific functions

subject to missing failure types. Lu & Liang (2008) studied a semiparametric additive model for cause-specific functions with missing failure types. In both papers, an additional model is supplied for the conditional probability of the failure type of interest $P(J = 2|W)$ to utilize the data from the failures with missing types as well as from the complete cases when there is only one stratum, where J is the type of failure in a typical competing risks set-up. Whereas Gao & Tsiatis (2005) and Lu & Liang (2008) had the flexibility to model $P(J = 2|W)$, our conditional distribution of the mark, $\rho_k(v, w)$, being specified for each $v \in [0, 1]$, has a built-in structure that must be considered. Nevertheless, we will show that the idea of Robins *et al.* (1994) can still be used for model (2) to improve efficiency. Specifically, more efficient estimation of (2) can be achieved by incorporating knowledge of $\rho_k(v, w)$ into the estimation procedure. Let $g_k(a|t, v, z) = P(A_{ki} = a|T_{ki} = t, V_{ki} = v, Z_{ki} = z, \delta_{ki} = 1)$. Then

$$\rho_k(v, w) = \int_0^v \lambda_k(t, u|z)\, g_k(a|t, u, z)\, du \Big/ \int_0^1 \lambda_k(t, u|z)\, g_k(a|t, u, z)\, du, \quad (11)$$

where $w = (t, z, a)$. If A_{ki} is independent of V_{ki} given $(T_{ki}, Z_{ki}, \delta_{ki})$, then $\rho_k(v, w) = \int_0^v \lambda_k(t, u|z)\, du \Big/ \int_0^1 \lambda_k(t, u|z)\, du$. In this case, $\rho_k(v, w)$ can be estimated by $\widehat{\rho}_k^{ipw}(v, w) = \int_0^v \widehat{\lambda}_k^{ipw}(t, u|z)\, du \Big/ \int_0^1 \widehat{\lambda}_k^{ipw}(t, u|z)\, du$, where $\widehat{\lambda}_k^{ipw}(t, u|z) = \widehat{\lambda}_{0k}^{ipw}(t, u) \exp\left\{\left(\widehat{\beta}^{ipw}(u)\right)^T z\right\}$. Routine kernel methods can be used to show that $\widehat{\rho}_k^{ipw}(v, W_{ki}) \xrightarrow{P} \rho_k^{ipw}(v, W_{ki})$ at the rate of $(nh)^{-1/2}$ when $b_1 \asymp b_2 \asymp h$, where $\asymp$ means convergence to zero at the same rate.

When the auxiliary marks A_{ki} are correlated with V_{ki}, the conditional distribution $\rho_k(v, w)$ involves the function $g_k(a|t, u, z)$, which can be modeled to capture the correlation. The stronger the correlation, the greater potential to improve efficiency. As considered in Section 1, if V_{ki} is the genetic distance of an early virus, then an auxiliary mark A_{ki} that is expected to predict V_{ki} is the genetic distances of all available later viruses and their sampling times.

Consider a parametric model $g_k(a|t, u, z, \theta_k)$ for $g_k(a|t, u, z)$. Let $\widehat{g}_k(a|t, u, z)$ be an estimator of $g_k(a|t, u, z)$. Then $\rho_k(v, w)$ can be estimated by

$$\widehat{\rho}_k^{ipw}(v, w) = \int_0^v \widehat{\lambda}_k^{ipw}(t, u|z)\, \widehat{g}_k(a|t, u, z)\, du \Big/ \int_0^1 \widehat{\lambda}_k^{ipw}(t, u|z)\, \widehat{g}_k(a|t, u, z)\, du. \quad (12)$$

Taking $b_1 \asymp b_2 \asymp h$, it is easy to show that $\widehat{\rho}_k^{ipw}(v, W_{ki}) \xrightarrow{P} \rho_k(v, W_{ki})$ at the rate of $(nh)^{-1/2}$.

Let $N_{ki}^x(t) = I(X_{ki} \le t, \delta_{ki} = 1)$, $N_{ki}^v(v) = I(V_{ki} \le v)$ and $\widehat{\rho}^{ipw}(\cdot) = \left(\widehat{\rho}_1^{ipw}(\cdot), \ldots, \widehat{\rho}_K^{ipw}(\cdot)\right)$. Following Robins *et al.* (1994), we obtain the following augmented (AUG) inverse probability weighted estimating equation for β:

$$U_{aug}\left(v, \beta, \widehat{\psi}, \widehat{\rho}^{ipw}(\cdot)\right) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau K_h(u - v)\left(Z_{ki}(t) - \bar{Z}_k(t, \beta)\right)\left\{\frac{R_{ki}}{\pi_k\left(Q_{ki}, \widehat{\psi}_k\right)} N_{ki}(dt, du) + \left(1 - \frac{R_{ki}}{\pi_k\left(Q_{ki}, \widehat{\psi}_k\right)}\right) N_{ki}^x(dt)\, d\left(\widehat{\rho}_k^{ipw}(u, W_{ki})\right)\right\}. \quad (13)$$

The AUG estimator of $\beta(v)$ solves the above equation and is denoted by $\widehat{\beta}^{aug}(v)$.

The doubly cumulative baseline function $\Lambda_{0k}(t, v) = \int_0^t \int_0^v \lambda_{0k}(s, u)\, ds du$ is estimated by

$$\widehat{\Lambda}_{0k}^{aug}(t,v)=\sum_{i=1}^{n_k}\int_0^t\int_0^v\frac{R_{ki}}{\pi_k\left(Q_{ki},\widehat{\psi}_k\right)}\frac{N_{ki}(ds,du)}{n_kS_k^{(0)}\left(s,\widehat{\beta}^{aug}(u)\right)}+\left(1-\frac{R_{ki}}{\pi_k\left(Q_{ki},\widehat{\psi}_k\right)}\right)\frac{N_{ki}^x(ds)\,d\left(\widehat{\rho}_k^{ipw}(u,W_{ki})\right)}{n_kS_k^{(0)}\left(s,\widehat{\beta}^{aug}(u)\right)}. \quad (14)$$

The baseline hazard function $\lambda_{0k}(t, v)$ can be estimated by the kernel estimator

$$\widehat{\lambda}_{0k}^{aug}(t,v)=\int_0^\tau\int_0^1K_{b1}^{(1)}(t-s)\,K_{k2}^{(2)}(v-u)\,\widehat{\Lambda}_{0k}^{aug}(ds,du). \quad (15)$$

4 Asymptotic properties

Let $\mathscr{F}_t=\sigma\{I(X_{ki}\le s,\delta_{ki}=1), I(X_{ki}\le s,\delta_{ki}=0), V_{ki}I(X_{ki}\le s,\delta_{ki}=1), Z_{ki}(s);0\le s\le t, i=1,\ldots,n_k, k=1,\ldots,K\}$, be the (right-continuous) filtration generated by the full data processes $\{N_{ki}(s, v), Y_{ki}(s), Z_{ki}(s); 0\le s\le t, 0\le v\le 1, i=1,\ldots,n_k, k=1,\ldots,K\}$. Assume $E(N_{ki}(dt,dv)|\mathscr{F}_{t-})=E(N_{ki}(dt,dv)|Y_{ki}(t),Z_{ki}(t))$, that is, the mark-specific instantaneous failure rate at time t given the observed information up to time t only depends on the failure status and the current covariate value. By the definition (1), $E(N_{ki}(dt,dv)|\mathscr{F}_{t-})=Y_{ki}(t)\lambda_k(t,v|Z_{ki}(t))\,dtdv$. Hence, the mark-specific intensity of $N_{ki}(t, v)$ with respect to $\mathscr{F}_t$ equals $Y_{ki}(t)\lambda_{ki}(t, v|Z_{ki}(t))$. Let $M_{ki}(t,u)=\int_0^t\int_0^u[N_{ki}(ds,dx)-Y_{ki}(s)\lambda_k(s,x|Z_{ki}(s))\,dsdx]$. By Aalen & Johansen (1978), $M_{ki}(\cdot, v_1)$ and $M_{ki}(\cdot, v_2)-M_{ki}(\cdot, v_1)$ are orthogonal square integrable martingales with respect to $\mathscr{F}_t$ for any $0\le v_1\le v_2\le 1$.

Let $\mathscr{F}_t^*=\mathscr{F}_t\cup\{R_{ki},\delta_{ki}A_{ki}; i=1,\ldots,n_k,k=1,\ldots,K\}$ be the right continuous filtration obtained by adding R_{ki} and $\delta_{ki}A_{ki}$ to $\mathscr{F}_t$. Let $\lambda_{ki}^*(t,v)=P(T_{ki}=t,V_{ki}=v|X_{ki}\ge t,Z_{ki}(t),R_{ki},\delta_{ki}A_{ki})$. Then

$$E(N_{ki}(dt,dv)|\mathscr{F}_{t-}^*)=Y_{ki}(t)\lambda_{ki}^*(t,v)\,dtdu, \quad (16)$$

where $Y_{ki}(t)\lambda_{ki}^*(t,v)$ is the intensity of $N_{ki}(t, v)$ with respect to $\mathscr{F}_t^*$. Assume that $\lambda_{ki}^*(t,v)$ is continuous in (t, v). Let $M_{ki}^*(t,v)=N_{ki}(t,v)-\int_0^t\int_0^v\lambda_{ki}^*(s,u)\,Y_{ki}(s)\,dsdu$. By Aalen & Johansen (1978), for any $0\le v_1\le v_2\le 1$, the processes $M_{ki}^*(t,v_1)$ and $M_{ki}^*(t,v_2)-M_{ki}^*(t,v_1)$, $0\le t\le\tau$ are orthogonal square integrable martingales.

The following regularity conditions are assumed throughout the rest of the paper. Most of the notation can be found at the end of Section 2.

Condition A

(A.1) $\beta(v)$ has componentwise continuous second derivatives on [0, 1]. For each $k = 1, \ldots, K$, the second partial derivative of $\lambda_{0k}(t, v)$ with respect to v exists and is continuous on $[0, \tau]\times[0, 1]$. The covariate process $Z_{ki}(t)$ has paths that are left continuous and of bounded variation, and satisfies the moment condition $E[\|Z_{ki}(t)\|^4\exp(2M\|Z_{ki}(t)\|)]<\infty$, where M is a constant such that $(v,\beta(v))\in[0,1]\times(-M, M)^p$ for all v and $\|A\|=\max_{k,l}|a_{kl}|$ for a matrix $A=(a_{kl})$.

(A.2) Each component of $s_k^{(j)}(t,\theta)$ is continuous on $[0,\tau]\times[-M, M]^p$, $\tilde{s}_k^{(j)}(t,\theta\psi_k)$ is continuous on $[0,\tau]\times[-M, M]^p\times[-L, L]^q$ for some $M, L>0$ and $j=0,1,2$. $\sup_{t\in[0,\tau],\theta\in[-M,M]^p}\|S_k^{(j)}(t,\theta)-s_k^{(j)}(t,\theta)\|=O_p\left(n^{-1/2}\right)$, and $\sup_{t\in[0,\tau],\theta\in[-M,M]^p,\psi_k\in[-L,L]^q}\|\tilde{S}_k^{(j)}(t,\theta,\psi_k)-\tilde{s}_k^{(j)}(t,\theta,\psi_k)\|=O_p\left(n^{-1/2}\right)$.

(A.3) The limit $p_k = \lim_{n\to\infty} n_k/n$ exists and $0 < p_k < 1$. $s_k^{(0)}(t,\theta)\rangle 0$ on $[0, \tau] \times [-M, M]^p$ and the matrix $\Sigma(v) = \sum_{k=1}^{K} p_k \Sigma_k(v)$ is positive definite, where $\Sigma_k(v) = \sum_{k=1}^{K} \int_0^\tau I_k(t, \beta(v))\, \lambda_{0k}(t, v)\, s_k^{(0)}(t, \beta(v))\, dt$.

(A.4) The kernel functions $K(\cdot)$, $K^{(1)}(\cdot)$ and $K^{(2)}(\cdot)$ are symmetric with support $[-1, 1]$ and have bounded variation. The bandwidths satisfy $b_1 \asymp b_2 \asymp h$ and $nh^2 \to \infty$ and $nh^5 = O(1)$ as $n \to \infty$.

(A.5) There is a $\varepsilon > 0$ such that $r_k(W_{ki}) \geq \varepsilon$ for all k, i with $\delta_{ki} = 1$.

Discussion of some of these conditions can be found in Sun *et al.* (2009). To avoid the problems at the boundaries $v = 0, 1$, we study the asymptotic properties of $\widehat{\beta}(v)$ for interior values of $v \in [a, b] \subset (0, 1)$.

4.1 Asymptotic results of inverse probability weighted complete-case estimator

The consistency and asymptotic normality of $\widehat{\beta}^{ipw}(v)$ are established in the next two theorems.

Theorem 1. Under Condition A, if the model for $r_k(W_{ki})$ is correctly specified, then $\widehat{\beta}^{ipw}(v) \xrightarrow{P} \beta(v)$ uniformly in $v \in [a, b] \subset (0, 1)$ as $n \to \infty$.

Let $\nu_0 = \int K^2(x)\, dx$, $\mu_2 = \int x^2 K(x) dx$,

$$\vartheta(v) = \sum_{k=1}^{K} p_k E\left\{ \int_0^\tau \left[Z_{ki}(t) - \bar{z}_k(t, \beta) \right] \frac{\partial^2 \lambda_k(t, v | Z_{ki}(t))}{\partial v^2} Y_{ki}(t)\, dt \right\},$$

$$\Sigma_k^*(v) = E\left\{ \int_0^\tau \left[Z_{ki}(t) - \bar{z}_k(t, \beta) \right]^{\otimes 2} \frac{R_{ki}}{(\pi_k(t, Z_{ki}, \delta_{ki} A_{ki}))^2} \lambda_{ki}^*(t, v)\, Y_{ki}(t)\, dt \right\}.$$

Theorem 2. Under Condition A, if the model for $r_k(W_{ki})$ is correctly specified, then $(nh)^{1/2}\left[\widehat{\beta}^{ipw}(v) - \beta(v) + \frac{1}{2}\mu_2 h^2 \Sigma^{-1}(v)\, \vartheta(v)\right] \xrightarrow{\mathcal{D}} N\left(0, \nu_0 \Sigma^{-1}(v)\, \Sigma^*(v)\, \Sigma^{-1}(v)\right)$ for $v \in [a, b]$ as $n \to \infty$, where $\Sigma^*(v) = \sum_{k=1}^{K} p_k \Sigma_k^*(v)$ and $\Sigma(v)$ is given in condition (A.3).

By (16) and by the property of expectation that $E(\cdot) = E\{E(\cdot | W_{ki})\}$,

$$\int_0^v \Sigma^*(u)\, du = \sum_{k=1}^{K} p_k E\left\{ \int_0^v \int_0^\tau \left[Z_{ki}(t) - \bar{z}_k(t, \beta(u)) \right]^{\otimes 2} \frac{1}{\pi_k(W_{ki})} N_{ki}(dt, du) \right\}.$$

$\nu_0 \Sigma^*(v)$ can be consistently estimated by

$$\tilde{\Sigma}_{ipw}^*(v) = n^{-1} h \sum_{k=1}^{K} \sum_{i=1}^{n_k} \int_0^1 \int_0^\tau (K_h(u - v))^2 \left[Z_{ki}(t) - \tilde{Z}_k\left(t, \widehat{\beta}(u), \widehat{\psi}_k\right) \right]^{2\otimes} \frac{R_{ki}}{\left(\pi_k\left(Q_{ki}, \widehat{\psi}_k\right)\right)^2} N_{ki}(dt, du).$$

The proof of Theorem 2 uses a Taylor expansion of the score function, leading to $\widehat{\beta}^{ipw}(v)-\beta(v)=-\left(U'_{ipw}\left(v,\beta^*(v),\widehat{\psi}\right)\right)^{-1}U_{ipw}\left(v,\beta(v),\widehat{\psi}\right)$, where $\beta^*(v)$ is on the line segment between $\widehat{\beta}^{ipw}(v)$ and $\beta(v)$, and

$$U'_{ipw}\left(v,\beta,\widehat{\psi}\right)=-\sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^1\int_0^\tau K_h(u-v)\frac{R_{ki}}{\pi_k\left(Q_{ki},\widehat{\psi}_k\right)}\tilde{J}_k\left(t,\beta,\widehat{\psi}_k\right)N_{ki}(dt,dt)$$

is the derivative of $U_{ipw}\left(v,\beta,\widehat{\psi}\right)$ with respect to β. Here $\tilde{J}_k\left(t,\beta,\widehat{\psi}_k\right)$ is defined at the end of Section 2. It can be shown that $\widehat{\Sigma}_{ipw}(v)\equiv -n^{-1}U'_{ipw}\left(v,\widehat{\beta}(v),\widehat{\psi}\right)\xrightarrow{P}\Sigma(v)$ as $n\to\infty$. Thus, the asymptotic variance $\nu_0\Sigma^{-1}(v)\Sigma^*(v)\Sigma^{-1}(v)$ can be estimated by $\left(\widehat{\Sigma}_{ipw}(v)\right)^{-1}\tilde{\Sigma}^*_{ipw}(v)\left(\widehat{\Sigma}_{ipw}(v)\right)^{-1}$.

4.2 Asymptotic results of augmented inverse probability weighted complete-case estimator

Let S^{ψ}_{ki} and I^{ψ}_{k} be the score vector and information matrix for $\widehat{\psi}_k$ under (7). Then

$$S^{\psi}_{ki}=\frac{\delta_{ki}\left(R_{ki}-r_k\left(W_{ki},\psi_{k0}\right)\right)}{r_k\left(W_{ki},\psi_{k0}\right)\left(1-r_k\left(W_{ki},\psi_{k0}\right)\right)}\frac{\partial r_k\left(W_{ki},\psi_{k0}\right)}{\partial\psi_k},\quad(17)$$

$$I^{\psi}_{ki}=E\left\{\frac{\delta_{ki}}{r_k\left(W_{ki},\psi_{k0}\right)\left(1-r_k\left(W_{ki},\psi_{k0}\right)\right)}\frac{\partial r_k\left(W_{ki},\psi_{k0}\right)}{\partial\psi_k}\left(\frac{\partial r_k\left(W_{ki},\psi_{k0}\right)}{\partial\psi_k}\right)^T\right\},\quad(18)$$

and $\widehat{\psi}_k-\psi_{k0}=n_k^{-1}\sum_{i=1}^{n_k}\left(I^{\psi}_k\right)^{-1}S^{\psi}_{ki}+o_p\left(n_k^{-1/2}\right)$.

We also introduce the following notation:

$$\begin{aligned}
\mathscr{A}_{ki} &= \int_0^1\int_0^\tau K_h(u-v)\left(Z_{ki}(t)-\bar{z}_k(t,\beta(u))\right)\frac{R_{ki}}{\pi_k(Q_{ki},\psi_{k0})}M_{ki}(dt,du),\\
\mathscr{B}_{ki} &= \int_0^1\int_0^\tau K_h(u-v)\left(Z_{ki}(t)-\bar{z}_k(t,\beta(u))\right)\left(1-\frac{R_{ki}}{\pi_k(Q_{ki},\psi_{k0})}\right)E\left\{M_{ki}(dt,du)\,|Q_{ki}\right\},\\
\mathscr{D}_k &= n_k^{-1}\sum_{i=1}^{n_k}\int_0^1\int_0^\tau K_h(u-v)\left(Z_{ki}(t)-\bar{z}_k(t,\beta(u))\right)\otimes\left\{\frac{-R_{ki}}{(\pi_k(Q_{ki},\psi_{k0}))^2}\frac{\partial\pi_k(Q_{ki},\psi_{k0})}{\partial\psi_k}M_{ki}(dt,du)\right\},\\
\mathscr{O}_{ki} &= \mathscr{D}_k\left(I^{\psi}_k\right)^{-1}S^{\psi}_{ki}.
\end{aligned}\quad(19)$$

The next theorem shows that the AUG estimator $\widehat{\beta}^{aug}(t,v)$ is consistent if either $r_k(w,\psi_k)$ or $g_k(a|t,v,z,\theta_k)$ is correctly specified, a double robustness property.

Theorem 3. Assuming Condition A, $\widehat{\beta}^{aug}(v)\xrightarrow{P}\beta(v)$ uniformly in v ∈ [a, b] ⊂ (0, 1) as n → ∞. This consistency holds if either rk(w; ψ_k) or $g_k(a|t,v,z,\theta_k)$ is correctly specified.

Following the proofs of Theorem 2 and 3, it is easy to show that $(nh)^{1/2}\left[\widehat{\beta}^{aug}(v)-\beta(v)+\frac{1}{2}\mu_2h^2\Sigma^{-1}(v)\,\vartheta^*(v)\right]$ is asymptotically normal for $v\in[a,b]\subset(0,1)$ for some function $\vartheta^*(v)$ if either $r_k(w,\psi_k)$ or $g_k(a|t,v,z,\theta_k)$ is correctly specified. When

both $r_k(w, \psi_k)$ and $g_k(a|t, v, z, \theta_k)$ are correctly specified, Theorem 4 below shows that $\widehat{\beta}^{aug(v)}$ is more efficient than $\widehat{\beta}^{ipw(v)}$.

Theorem 4. Assuming Condition A, if both $r_k(w, \psi_k)$ and $g_k(a|t, v, z, \theta_k)$ are correctly specified, we have $(nh)^{1/2}\left[\widehat{\beta}^{aug}(v) - \beta(v) + \frac{1}{2}\mu_2 h^2 \Sigma^{-1}(v)\vartheta(v)\right] \underset{\rightarrow}{\mathscr{D}} N\left(0, \nu_0 \Sigma^{-1}(v)\Sigma^*(v)\Sigma^{-1}(v)\right)$ for $v \in [a, b]$ as $n \to \infty$ where $\Sigma^*(v) = \sum_{k=1}^{K} pk\Sigma_k^*(v)$ and $\Sigma(v)$ is given in the condition (A. 3). The estimator $\widehat{\beta}^{aug(v)}$ is more efficient than $\widehat{\beta}^{ipw(v)}$ in the sense that

$$\begin{aligned} Cov\Big\{(nh)^{1/2}\Big[\widehat{\beta}^{ipw}(v) &- \beta(v) + (1/2)\mu_2 h^2 \Sigma^{-1}(v)\vartheta(v) = Cov\Big\{(nh)^{1/2}\Big[\widehat{\beta}^{aug}(v) - \beta(v) + (1/2)\mu_2 h^2 \Sigma^{-1}(v)\vartheta(v)\Big]\Big\} \\ &+ h\Sigma^{-1}(v)\left(\sum_{k=1}^{K}(n_k/n)\, Cov\{\mathscr{O}_{k1} - \mathscr{B}_{k1}\}\right)\Sigma^{-1}(v) + o_p(h). \end{aligned} \tag{20}$$

Equation (20) shows that the asymptotic covariance $Cov\left\{(nh)^{1/2}\left[\widehat{\beta}^{aug}(v) - \beta(v) + (1/2)_{\mu_2} h^2 \Sigma^{-1}(v)\vartheta(v)\right]\right\}$ is smaller than $Cov\left\{(nh)^{1/2}\left[\widehat{\beta}^{ipw}(v) - \beta(v) + (1/2)_{\mu_2} h^2 \Sigma^{-1}(v)\vartheta(v)\right]\right\}$ and the difference between the two covariances is at the order of h. The demonstrated efficiency gain for $\widehat{\beta}^{aug}(v)$ is reasonable since the estimation procedures for both $\widehat{\beta}^{aug}(v)$ and $\widehat{\beta}^{ipw}(v)$ are based on the local partial likelihood with h as its bandwidth.

Let $\widehat{\rho}_k^{aug}(\cdot)$ be defined similarly to $\widehat{\rho}_k^{ipw}(\cdot)$ given in (12) with the AUG estimator $\widehat{\lambda}_k^{aug}(t, u|z) = \widehat{\lambda}_{0k}^{aug}(t, u)\exp\left\{\left(\widehat{\beta}^{aug}(u)\right)^T z\right\}$. Let $\widehat{\rho}^{aug}(\cdot) = \left(\widehat{\rho}_1^{aug}(\cdot), \ldots, \widehat{\rho}_K^{aug}(\cdot)\right)$, $\widehat{\Sigma}_{aug}(v) = -n^{-1}U'_{aug}\left(v, \widehat{\beta}^{aug}(v), \widehat{\Psi}, \widehat{\rho}(\cdot)\right)$ and

$$\tilde{\Sigma}_{aug}^*(v) = n^{-1}h\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\{\int_0^1\int_0^\tau K_h(u - v)\left(Z_{ki}(t) - \bar{Z}_k\left(t, \widehat{\beta}^{aug}(u)\right)\right)\left[\frac{R_{ki}}{\pi_k\left(Q_{ki}, \widehat{\psi}_k\right)}N_{ki}(dt, du) + \left(1 - \frac{R_{ki}}{\pi_k\left(Q_{ki}, \widehat{\psi}_k\right)}\right)N_{ki}^x(dt)\, d\left(\widehat{\rho}_k^{aug}(u, W_{ki})\right)\right]\right\}^{\otimes 2}.$$

The asymptotic variance of $(nh)^{1/2}\left(\widehat{\beta}^{aug}(v) - \beta(v)\right)$ can be consistently estimated by $\left(\widehat{\Sigma}_{aug}(v)\right)^{-1}\tilde{\Sigma}_{aug}^*(v)\left(\widehat{\Sigma}_{aug}(v)\right)^{-1}$.

Let $\beta(v) = \left(\beta_1(v), \beta_2^T(v)\right)^T$ where $\beta_1(v)$ is the coefficient for vaccination status and $\beta_2(v)$ for other covariates. Let $\sigma_1^2(v)$ be the first element on the diagonal of the matrix $\nu_0\Sigma^{-1}(v)\Sigma^*(v)\Sigma^{-1}(v)$ and $\widehat{\sigma}_1^2(v)$ its consistent estimator, which is the first element on the diagonal of $\left(\widehat{\Sigma}_{aug}(v)\right)^{-1}\tilde{\Sigma}_{aug}^*(v)\left(\widehat{\beta}_{aug}(v)\right)^{-1}$. By Theorem 4, a large sample 100(1−α% pointwise confidence interval for $\beta_1(v)$ is given by $\widehat{\beta}_1^{aug}(v) \pm (nh)^{1/2} z_{\alpha/2}\widehat{\sigma}_1(v)$, $0 < v < 1$,

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. The confidence intervals for the other coefficient functions can be constructed similarly.

The mark-specific vaccine efficacy $VE(v) = 1-\exp(\beta_1(v))$ can be estimated by $\widehat{VE}(v) = 1 - \exp\left(\widehat{\beta}_1^{aug}(v)\right)$. By Theorem 4 and the delta method, $(nh)^{1/2}\left(\widehat{VE}(v) - \mathrm{VE}(v)\right) \xrightarrow{\mathcal{D}} N\left(0, \sigma_1^2(v)\exp(2\beta_1(v))\right)$ for $v \in (0,1)$. A large sample $100(1-\alpha)$ % pointwise confidence interval for $\mathrm{VE}(v)$ is then given by $\widehat{\mathrm{VE}}(v) \pm (nh)^{1/2} z_{\alpha/2}\widehat{\sigma}_1(v)\exp\left(\widehat{\beta}_1^{aug}(v)\right)$, $0 < v < 1$.

5 Simulation study

5.1 Assessment of estimation procedures under correctly specified models

First, we conduct a simulation study to check the finite sample performance of the two proposed estimation procedures when both $r_k(w)$ and $g_k(a|t, v, z)$ are correctly specified. The augmented inverse probability weighted complete-case estimator (AUG) for (2) with missing marks is compared to the *complete-case estimator* (CC) wherein the failures with missing marks are deleted for the analysis, and to the inverse probability weighted complete-case estimator (IPW). These estimators are compared to the unachievable benchmark, the *full data likelihood estimator* (Full), which analyzes the full data-set before some simulated marks are deleted. We consider the case with $K = 1$ stratum, in which case the CC and Full methods are that of Sun *et al.* (2009).

With $k = 1$ throughout this section, let Z_{ki} be the treatment indicator taking value 0 or 1 with probability of 0.5 for each value. The (T_{ki}, V_{ki}) are generated from the following mark-specific proportional hazards model:

$$\lambda(t, v|z) = \exp\{\gamma v + (\alpha + \beta v) z\}, \quad t \geq 0, 0 \leq v \leq 1, \quad (21)$$

where α, β and γ are constants. Under model (21), the mark-specific baseline function is $\lambda_0(t, v) = \exp(\gamma v)$ and the mark-specific vaccine efficacy is $\mathrm{VE}(v) = 1 - \exp(\alpha + \beta v)$. The vaccine efficacy $\mathrm{VE}(v) = 0$ for $\alpha = 0$ and $\beta = 0$, indicating no vaccine efficacy. The vaccine efficacy does not depend on the type of infecting strain if $\beta = 0$, while the $\mathrm{VE}(v)$ decreases in v if $\beta > 0$. We examine the estimation procedures for the following specific models:

- **(M1)** $(\alpha, \beta, \gamma) = (0, 0, 0.3)$, where $\mathrm{VE}(v) = 0$;
- **(M2)** $(\alpha, \beta, \gamma) = (-0.69, 0, 0.3)$, where $\mathrm{VE}(v)$ does not depend on v;
- **(M3)** $(\alpha, \beta, \gamma) = (-0.6, 0.6, 0.3)$, where $\mathrm{VE}(v)$ decreases;
- **(M4)** $(\alpha, \beta, \gamma) = (-1.2, 1.2, 0.3)$, where $\mathrm{VE}(v)$ decreases.

All failure times greater than $\tau = 2.0$ are considered as censored. In addition, we generate random censoring times from an exponential distribution, independent of (T, V), with parameter adjusted so that the overall censoring rates range from 25% to 35%. The complete-case indicator R_{ki} is generated with conditional probability $r_k(W_{ki}) = P(R_{ki} = 1|\delta_{ki} = 1, W_{ki})$, where

$$\mathrm{logit}(r_k(W_{ki})) = \psi_{k0} + \psi_{k1} Z_{ki}, \quad i = 1, \ldots, n_k. \quad (22)$$

Here $\psi_{k0} = 0.2$ and $\psi_{k1} = -0.2$. The proportion of missing marks is about 50%, similar to the rate observed in current HIV vaccine efficacy trials.

For the proposed AUG estimator, conditional on (T_{ki}, Z_{ki}, V_{ki}), we assume that a single auxiliary mark follows the model

$$A_{ki}=(\theta+1)^{-1}(V_{ki}+\theta U_{ki}),\quad \theta>0, \quad (23)$$

for $i = 1, \ldots, n_k$, where V_{ki} is the possibly missing mark generated from model (21), U_{ki} is uniformly distributed on [0, 1] independent of V_{ki}, and θ > 0 measures the association between A_{ki} and V_{ki}. The correlation coefficient ρ between A_{ki} and V_{ki} is 1 for θ = 0. Since A_{ki} is observed for all observed failure times, the AUG estimator in this case is the Full estimator. The A_{ki} and V_{ki} are independent for θ = ∞(ρ = 0), in which case the AUG estimator is denoted by AUG-A0. The AUG estimator is denoted by AUG-A1 for θ = 0.8 yielding ρ ∞ 0.78, by AUG-A2 for θ = 0.4 with ρ ∞ 0.92 and by AUG-A3 for θ = 0.2 with ρ ∞ 0.98. Thus the simulation compares the AUG estimators at four levels of association between A_{ki} and V_{ki}.

We study settings with relatively high correlations because these occur for real data-sets due to the fact that inter-subject HIV sequence diversity is considerably larger than intra-subject HIV sequence diversity (Keele *et al.*, 2008). We have found linear correlations between early and later sequence distances as high as 0.98.

Under model (23), the conditional density of A_{ki} given (T_{ki}, Z_{ki}, V_{ki}) is

$$g_k(a|t,v,z;\theta)=\frac{1+\theta}{\theta}I\left\{\frac{v}{1+\theta}\le a\le\frac{v+\theta}{1+\theta}\right\},0\le a\le 1,0\le v\le 1. \quad (24)$$

The likelihood for θ is

$$L(\theta)=\prod_{\delta_{ki}=1,R_{ki}=1}\left(\frac{1+\theta}{\theta}I\left\{\frac{V_{ki}}{1+\theta}\le A_{ki}\le\frac{V_{ki}+\theta}{1+\theta}\right\}\right)\text{ for }\theta>0.$$

It is easy to show that the maximum likelihood estimator is $\widehat{\theta}=\max_{\delta_{ki}=1,R_{ki}=1}\{V_{ki}/A_{ki},(1-V_{ki})/(1-A_{ki})\}-1$. Plugging in the density estimator $g_k\left(a|t,v,z;\widehat{\theta}\right)$ into (12) yields $\widehat{\rho}_k^{ipw}(v,w)$, which is used to construct the AUG estimator of β in (13).

Performance of the proposed estimation procedures is evaluated for the models described in (21), (22) and (24) through simulations. We use the Epanechnikov kernel $K(x) = .75(1 - x^2)I\{|x| \le 1\}$. The same kernel is also used for $K^{(1)}(x)$ and $K^{(2)}(x)$. For sample size $n = 500$ and bandwidths $b_1 = 0.1$ and $h = b_2 = 0.15$, Figure 1 shows the average estimates of $\beta(v)$ under the models (M1)–(M4) based on 500 simulations for four different procedures: Full, CC, IPW and AUG. Figure 2 shows the standard errors of these estimates and Figure 3 shows the pointwise coverage probabilities of 95% confidence intervals for $\beta(v)$ constructed using the AUG estimators based on 500 simulations.

The simulation results show that the bias of both estimators is very small at a level comparable to that of the full data likelihood estimator (Figure 1). However, the complete-case estimator yields large bias especially for models (M2)–(M4). The simulations also validate that the AUG estimator is more efficient than the IPW estimator (Figure 2). In particular, the AUG estimator without auxiliary marks (AUG-A0) has smaller standard errors than the IPW estimator, and the AUG estimator with the auxiliary mark has smaller standard errors than the AUG estimator without auxiliary marks. Moreover, the standard errors of the AUG estimator decrease as the correlation between the auxiliary mark and the mark of interest increases, with the standard errors for AUG-A1, AUG-A2 and AUG-A3 getting closer to that of the Full estimator. The pointwise coverage probabilities displayed in

Figures 3 are almost all in the range of 92.5% and 97.5% for $v \in (0, 1)$, indicating adequate performance of the proposed variance estimator for the AUG estimator. The deviation from the 95% nominal level is attributed to both the limited sample size (at $n = 500$) and the limited number of simulation runs (also at 500). The pointwise coverage probabilities are expected to be closer to 95% when both the sample size and the number of simulation runs increase.

The bandwidth selection is important in nonparametric smoothing procedures. The bias tends to be larger and the standard errors smaller for larger bandwidths. A reasonable bandwidth choice trades-off between bias and standard error. We have tried simulations with different bandwidths, taking bandwidth b_2 equal to 0.05 and 0.1 and b_1 equal to 0.15, and found our procedures relatively insensitive to the choice. The standard errors are larger at the boundary points than at the inner points, $v \in (b_2, 1 - b_2)$. In practice, the appropriate bandwidth selection can be based on a $\mathcal{K}$-fold cross-validation method. This approach is widely used in the nonparametric function estimation literature and has been investigated by Efron & Tibshirani (1993) and Tian *et al.* (2005) among others. The simulations also show satisfactory performance of the AUG estimator of $VE(v)$; see Figures 1–3 in the web Appendices.

5.2 Robustness analysis of the estimation procedures under mis-specified models

This subsection considers robustness of the proposed estimators to mis-specifications of $r_k(w)$ and/or $g_k(a|t, v, z)$, and to violation of the missing at random assumption. As above Z_{ki} is a binary random variable taking value 0 or 1 with probability 0.5 for each value. The (T_{ki}, V_{ki}) are generated from the mark-specific proportional hazards model (21) with (α, β, γ) given in (M3). The censoring times C_{ki} for model (M3) remain the same as before yielding approximately 30% censoring and $X_{ki} = \min(T_{ki}, C_{ki})$.

Robustness of the estimators to mis-specification of $r_k(w)$ is examined by assuming model (22) while the actual complete-case indicator R_{ki} is generated with the conditional probability $r_k(W_{ki}) = P(R_{ki} = 1|\delta_{ki} = 1, W_{ki})$, where

$$\text{logit}\,(r_k\,(W_{ki})) = 1.1 + Z_{ki} - X_{ki}, \quad i = 1, \ldots, n_k. \quad (25)$$

Robustness of the AUG estimator is also examined when $g_k(a|t, v, z)$ is mis-specified. This is carried out by assuming the model (23) for the auxiliary mark or equivalently model (24) for $g_k(a|t, v, z)$ while the actual mark is generated from

$$A_{ki} = (1.2 + 2\tau)^{-1}\,(V_{ki} + 0.2U_{ki} + 2X_{ki}), \quad (26)$$

for $i = 1, \ldots, n_k$, where U_{ki} is uniformly distributed on [0, 1] independent of V_{ki}.

Robustness of the estimators to violation of the missing at random assumption (3) is examined by assuming model (22) while the actual R_{ki} depends on V_{ki} through the model

$$\text{logit}\,(r_k\,(W_{ki})) = 0.6 + Z_{ki} - 2V_{ki}, \quad i = 1, \ldots, n_k. \quad (27)$$

The proportion of missing marks among the observed failure times is kept around 50% in all of the cases.

For sample size $n = 500$ and bandwidths $b_1 = 0.1$ and $h = b_2 = 0.15$, Figure 4 shows the bias of the estimators of $\beta(v)$ under model (M3) based on 500 simulations for four different procedures: Full, CC, IPW and AUG. Figure 5 shows the standard errors of these estimators. In both figures, panel (a) shows the plots when $r_k(w)$ is mis-specified following (25) and

$g_k(a|t, v, z)$ is correctly specified by (24) with $\theta = 0.2$; panel (b) shows the plots when $g_k(a|t, v, z)$ is mis-specified following (26) and $r_k(w)$ is correctly specified by (22) with $\psi_{k1} = 0.2$ and $\psi_{k1} = -0.2$; and panel (c) shows the plots when $r_k(w)$ is mis-specified following (25) and $g_k(a|t, v, z)$ is mis-specified following (26); and panel (d) shows the plots when $r_k(w)$ depends on V_{ki} following (27) and $g_k(a|t, v, z)$ is correctly specified by (24) with $\theta = 0.2$.

The bias of the IPW estimator is large as seen in Figure 4(a) when $r_k(w)$ is mis-specified. Figures 4(a) and (b) show that the AUG estimator has very little bias tracking closely the full data likelihood estimator when one of $r_k(w)$ and $g_k(a|t, v, z)$ is mis-specified, reflecting the double robustness property of the AUG estimator. When both $r_k(w)$ and $g_k(a|t, v, z)$ are mis-specified, the AUG estimator still has smaller bias compared to the IPW estimator, as seen in Figure 4(c). Surprisingly, the AUG estimator yields as little bias as the full data likelihood estimator when the missing at random assumption is violated in Figure 4(d) while the IPW estimator shows very large bias. In all the scenarios, the complete-case estimator has largest bias. Figure 5 shows that the AUG estimator is more efficient than the IPW estimator with smaller standard error when both $r_k(w)$ and $g_k(a|t, v, z)$ are correctly specified, and this efficiency gain diminishes when one of them is mis-specified. Again the AUG estimator seems to have smaller standard error when the missing at random assumption is violated as shown in Figure 5(d).

6 Discussion

To address the fact that mark variables are subject to missingness in HIV vaccine efficacy trials and other applications, this article extends previous work on the continuous mark-specific proportional hazards model to accommodate marks missing at random. We developed two approaches based on inverse probability weighting (IPW) of complete-cases, wherein failures in the score equation are weighted by the reciprocal of the probability the mark is observed. As always for IPW-based methods, the methods provide unstable estimation if some failures have outlying estimated probabilities near zero; therefore effective implementation in practice requires careful definition of the mark and covariates in the missingness model. Moreover, performance of the IPW-based methods depends on the ability to build a good model for the probability of observing the mark in failures. For HIV vaccine efficacy trials with the mark defined as the genetic distance of an `early' virus (i.e., measured on an HIV antibody negative blood sample), subject characteristics that could help predict whether the `early' virus is observed include viral load (very high viral load predicts the virus is early, and very low/undetectable viral load makes the sequencing technology prone to fail to measure the sequence), the timing of the subject's HIV tests including any extra tests beyond those scheduled, whether the infection was diagnosed during the series of injections (when HIV testing is more frequent), and whether any antiretroviral pre-exposure prophylaxis was received.

Following Robins *et al.* (1994), our second approach augments the IPW estimating equation with a term for failures with a missing mark, which recovers information through modeling and estimation of the mark distribution. The simulation study showed that this approach effectively gains back efficiency compared to the IPW approach, and that the quality of predictive modeling heavily influences the extent of efficiency improvement. In our formulation the conditional density of the auxiliaries given the mark of interest is modeled and estimated. Therefore in practice a one-dimensional (or at most two-dimensional) auxiliary may be all that the data can support, given the need to either use nonparametric density estimation or to specify a reasonable parametric model. Fortunately for HIV vaccine efficacy trials with mark the genetic distance of the `early' virus, the genetic distance of a `later' virus may be a well-predicting univariate mark, as supported by examination of HIV sequence databases. Recently completed and ongoing efficacy trials are measuring distances

from early and later viruses, and it will be of interest to apply the new methods to the forthcoming data-sets.

This article focused on point and interval estimation, leaving the problem of inference for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research of Yanqing Sun was partially supported by NSF grants DMS-0604576 and DMS-0905777, NIH grant R37 AI054165-09 and a fund provided by UNC Charlotte. The research of Peter Gilbert was partially supported by NIH grant R37 AI054165-09. The authors thank the associate editor and two referees for their constructive suggestions.

REFERENCES

Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics. 1978; 5:141–150.

Dabrowska DM. Smoothed Cox regression. The Annals of Statistics. 1997; 25:1510–1540.

Dewanji A. A note on a test for competing risks with missing failure type. Biometrika. 1992; 79:855–857.

Dinse GE. Nonparametric prevalenace and mortality estimators for animal experiments with incomplete cause-of-death data. Journal of the American Statistical Association. 1986; 81:328–336.

Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman & Hall; New York: 1993.

Fauci AS, Johnston MI, Dieffenbach CW, Burton DR, Hammer SM, Hoxie JA, Martin M, Overbaugh J, Watkins DI, Mahmoud A, Greene WC. HIV vaccine research: the way forward. Science. 2008; 321:530–532. [PubMed: 18653883]

Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, Para MF, Gilbert PB, The rgp120 HIV Vaccine Study Group. Placebo-controlled phase 3 trial of recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. Journal of Infectious Diseases. 2005; 191:654–665. [PubMed: 15688278]

Gao G, Tsiatis AA. Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failure. Biometrika. 2005; 92:875–891.

Gilbert PB, Lele S, Vardi Y. Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. Biometrika. 1999; 86:27–43.

Gilbert PB, McKeague IW, Sun Y. Tests for comparing mark-specific hazards and cumulative incidence functions. Lifetime Data Analysis. 2004; 10:5–28. [PubMed: 15130048]

Gilbert PB, McKeague IW, Sun Y. The Two-Sample Problem for Failure Rates Depending on a Continuous Mark: An Application to Vaccine Efficacy. Biostatistics. 2008; 9:263–276. [PubMed: 17704528]

Goetghebeur E, Ryan L. A modified logrank test for competing risks with missing failure type. Biometrika. 1990; 77:207–211.

Goetghebeur E, Ryan L. Analysis of competing risks survival data when some failure types are missing. Biometrika. 1995; 82:821–834.

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association. 1952; 47:663–685.

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. Wiley; New York: 1980.

Keele B, Giorgi E, Salazar-Gonzalez J, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping L-H, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH, Shaw GM. Identification and characterization of

transmitted and early founder virus envelopes in primary HIV-1 infection. Proceedings of the National Academy of Sciences. 2008; 105:7552–7557.

Lagakos SW, Louis TA. Use of tumour lethality to interpret tumorigenicity experiments lacking cause of death data. Journal of Applied Statistics. 1988; 37:169–179.

Lu W, Liang Y. Analysis of competing risks data with missing cause of failure under additive hazards model. Statistica Sinica. 2008; 19:219–234.

Lu K, Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics. 2001; 57:1191–1197. [PubMed: 11764260]

Racine-Poon AH, Hoel DG. Nonparametric estimation of survival function when cause of death is uncertain. Biometrics. 1984; 40:1151–1158. [PubMed: 6534414]

Rubin DB. Inference and missing data. Biometrika. 1976; 63:581–592.

Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.

Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models: rejoinder. Journal of the American Statistical Association. 1999; 94:1135–1146.

Silverman, BW. Density Estimation for Statistics and Data Analysis. Chapman and Hall; London: 1986.

Sun Y, Gilbert PB, McKeague IW. Proportional hazards models with continuous marks. The Annals of Statistics. 2009; 37:394–426.

Tian L, Zucker D, Wei LJ. On the Cox model with time-varying regression coefficients. Journal of the American Statistical Association. 2005; 100:172–183.

Wu TJ, Hsieh YC, Li LA. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics. 2001; 57:441–448. [PubMed: 11414568]

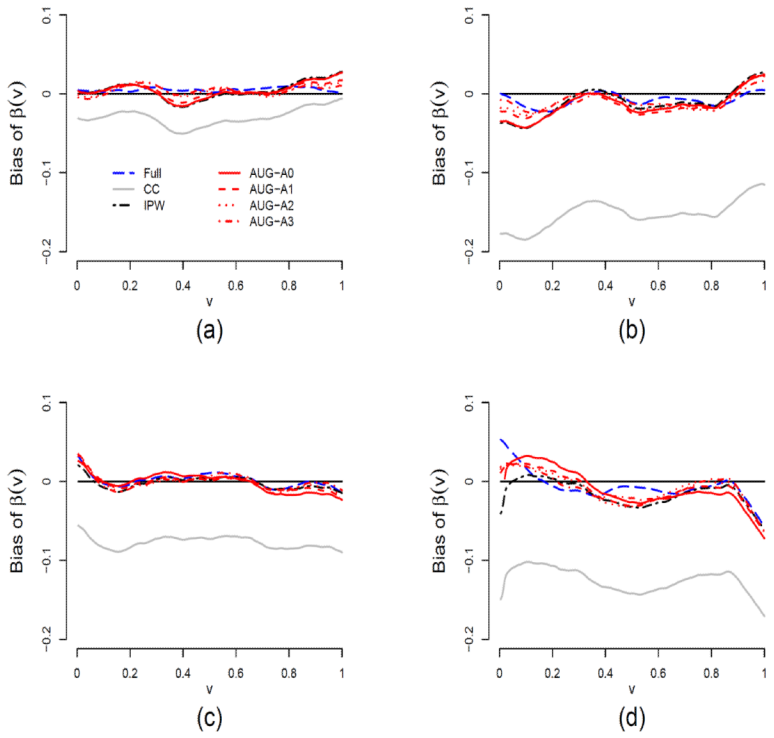

Figure 1.

Bias of estimation for β(*v*) under four procedures: Full, CC, IPW and AUG based on 500 simulations for $n = 500$, $b_1 = 0.1$ and $h = b_2 = 0.15$; (a) for model (M1), (b) for model (M2), (c) for model (M3) and (d) for model (M4). AUG-A0 is the AUG estimator corresponding to ρ = 0, AUG-A1 for ρ ≈ 0.78, AUG-A2 for ρ ≈ 0.92 and AUG-A3 for ρ ≈ 0.98, where ρ is the correlation coefficient between A_{ki} and V_{ki}.

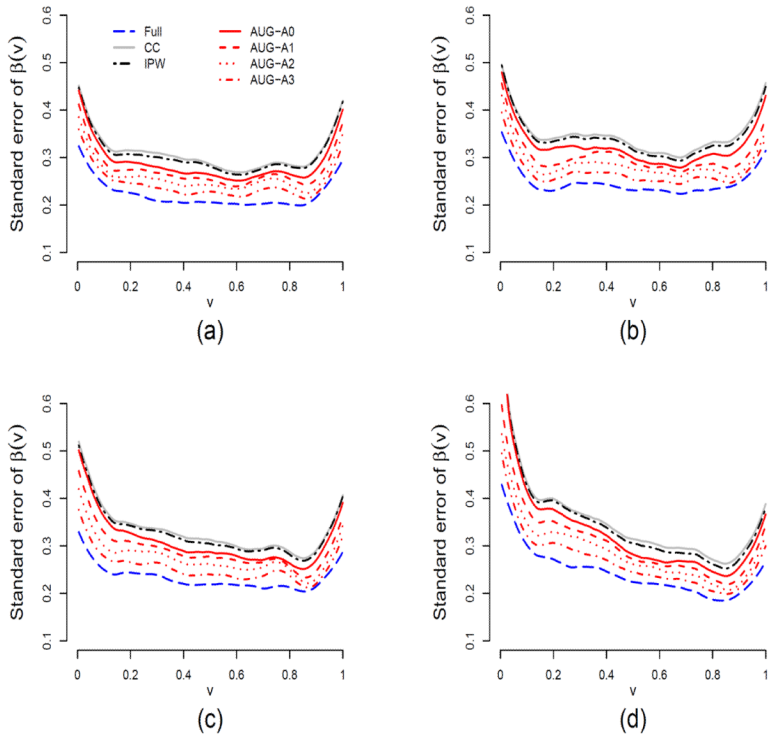

Figure 2.

Average standard errors of the estimates for β(*v*) under four procedures: Full, CC, IPW and AUG based on 500 simulations for $n = 500$, $b_1 = 0.1$ and $h = b_2 = 0.15$; (a) for model (M1), (b) for model (M2), (c) for model (M3) and (d) for model (M4). AUG-A0 is the AUG estimator corresponding to $\rho = 0$, AUG-A1 for $\rho \approx 0.78$, AUG-A2 for $\rho \approx 0.92$ and AUG-A3 for $\rho \approx 0.98$, where ρ is the correlation coefficient between A_{ki} and V_{ki}.

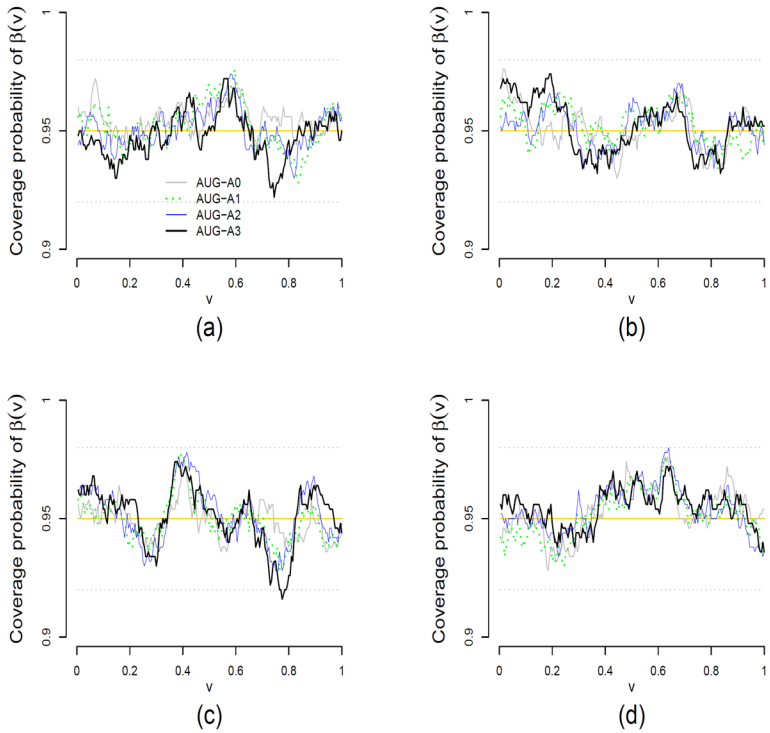

Figure 3.

Estimated pointwise coverage probabilities of 95% confidence intervals for β(*v*) constructed using the AUG estimator based on 500 simulations for $n = 500$, $b_1 = 0.1$ and $h = b_2 = 0.15$; (a) for model (M1), (b) for model (M2), (c) for model (M3) and (d) for model (M4). AUG-A0 is the AUG estimator corresponding to ρ = 0, AUG-A1 for ρ ≈ 0.78, AUG-A2 for ρ ≈ 0.92 and AUG-A3 for ρ ≈ 0.98, where ρ is the correlation coefficient between A_{ki} and V_{ki}.

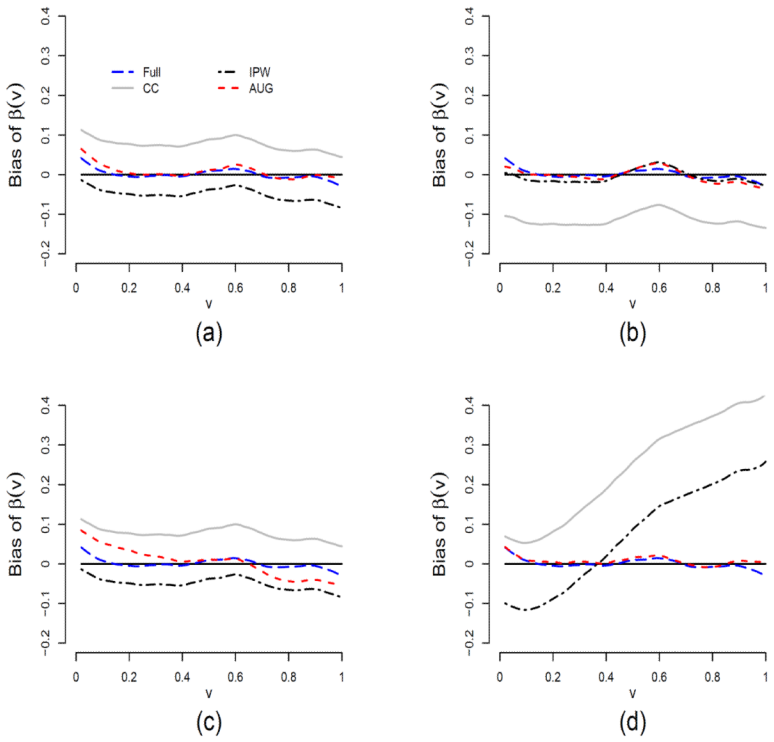

Figure 4.

Bias of estimation for $\beta(v)$ with the estimators, Full, CC, IPW and AUG, under model (M3) with $n = 500$, $b_1 = 0.1$ and $h = b_2 = 0.15$ based on 500 simulations. (a) shows the plots when $r_k(w)$ is mis-specified following (25) and $g_k(a|t, v, z)$ is correctly specified by (24) with $\theta = 0.2$. (b) shows the plots when $g_k(a|t, v, z)$ is mis-specified following (26) and $r_k(w)$ is correctly specified by (22) with $\psi_{k1} = 0.2$ and $\psi_{k1} = -0.2$. (c) shows the plots when $r_k(w)$ is mis-specified following (25) and $g_k(a|t, v, z)$ is mis-specified following (26). (d) shows the plots when $r_k(w)$ depends on V_{ki} following (27) and $g_k(a|t, v, z)$ is correctly specified by (24) with $\theta = 0.2$.

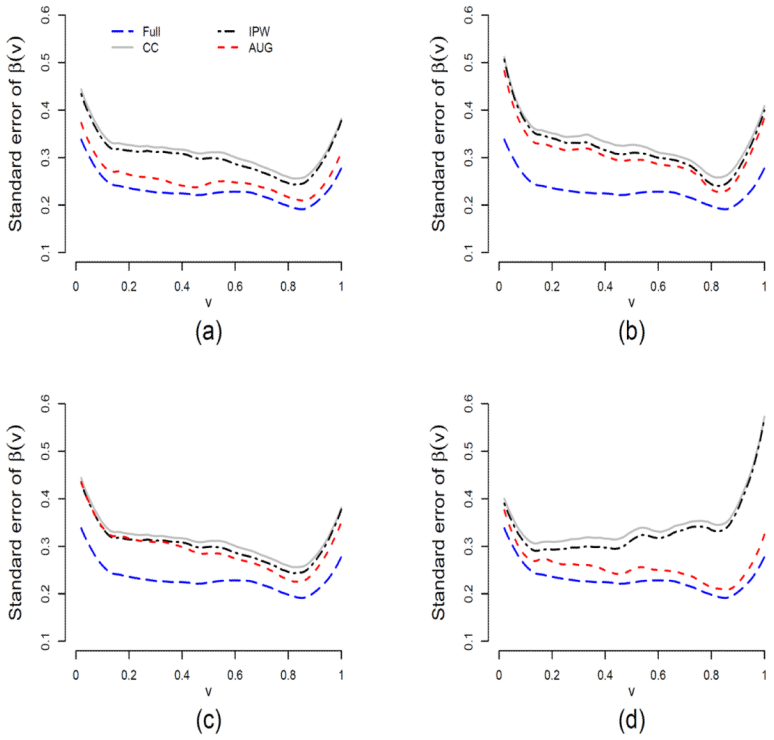

Figure 5.

Standard errors of the estimators, Full, CC, IPW and AUG, for $\beta(v)$ under model (M3) with $n = 500$, $b_1 = 0.1$ and $h = b_2 = 0.15$ based on 500 simulations. (a) shows the plots when $r_k(w)$ is mis-specified following (25) and $g_k(a|t, v, z)$ is correctly specified by (24) with $\theta = 0.2$. (b) shows the plots when $g_k(a|t, v, z)$ is mis-specified following (26) and $r_k(w)$ is correctly specified by (22) with $\psi_{k1} = 0.2$ and $\psi_{k1} = -0.2$. (c) shows the plots when $r_k(w)$ is mis-specified following (25) and $g_k(a|t, v, z)$ is mis-specified following (26). (d) shows the plots when $r_k(w)$ depends on V_{ki} following (27) and $g_k(a|t, v, z)$ is correctly specified by (24) with $\theta = 0.2$.