

Statistica Sinica Preprint No: SS-2024-0003

Title	Rank-Based Inference for the Accelerated Failure Time Model with Partially Interval-Censored Data
Manuscript ID	SS-2024-0003
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0003
Complete List of Authors	Taehwa Choi, Sangbum Choi and Dipankar Bandyopadhyay
Corresponding Authors	Dipankar Bandyopadhyay
E-mails	bandyopd@gmail.com
Notice: Accepted author version.	

**RANK-BASED INFERENCE FOR THE
ACCELERATED FAILURE TIME MODEL WITH
PARTIALLY INTERVAL-CENSORED DATA**

Taehwa Choi, Sangbum Choi and Dipankar Bandyopadhyay

Sungshin Women's University, Korea University and Virginia Commonwealth University

Abstract: This paper presents a unified rank-based inferential procedure for fitting the accelerated failure time model to partially interval-censored data. A Gehan-type monotone estimating function is constructed based on the idea of the familiar weighted log-rank test, and an extension to a general class of rank-based estimating functions is suggested. The proposed estimators can be obtained via linear programming and are shown to be consistent and asymptotically normal via standard empirical process theory. Unlike common maximum likelihood-based estimators for partially interval-censored regression models, our approach can directly provide a regression coefficient estimator without involving a complex nonparametric estimation of the underlying residual distribution function. An efficient variance estimation procedure for the regression coefficient estimator is considered. Moreover, we extend the proposed rank-based procedure to the linear regression analysis of multivariate clustered partially interval-censored data. The finite-sample operating characteristics of our approach are examined via simulation studies. Data example from a colorectal cancer study illustrates the practical usefulness of the method.

Key words and phrases: Accelerated failure time, Clustered data, Empirical processes, Gehan statistics, Interval-censoring, Left-censoring, Survival analysis.

1. Introduction

Interval-censored (IC) responses are frequently encountered in various biomedical settings and health studies, requiring periodic follow-ups or inspections (Bogaerts et al., 2017), where the time to an event of interest is only known to fall within a particular interval. Instances of exact failure times observed for certain subjects within the same study leads to partially interval-censored data, which can be categorized into two groups; namely, partly interval-censored (PIC; Kim, 2003), and doubly-censored (DC; Cai and Cheng, 2004). In addition to a certain number of exact observations (failures/events), DC data are either left-censored or right-censored (e.g., HIV/AIDS studies), whereas PIC data contain additional interval-censoring (e.g., cancer trials).

The PIC situation commonly arises when studying disease progression, such as the time to progression-free survival in cancer studies, where the nature of periodic examinations for research participants makes it challenging to precisely pinpoint the disease onset or record events, such as time

of death, while the study is ongoing. Consequently, some subjects have well-documented information regarding the exact timing of disease onset or occurrence of specific events, while others are interval-censored due to their clinical visit schedules. On the other hand, for DC endpoints, the exact event status is identified only when the measurements are within a specific range. For example, in HIV/AIDS treatment trials, the HIV-1 RNA level is typically used to measure the efficacy of antiretroviral therapy. This measure is generally considered reliable only when the RNA levels fall within a specific range. If the measurements fall outside this range, they are left- or right-censored, leading to DC data. DC and PIC data are reduced to “Case-1” and “Case-2” interval-censored data, respectively, under absence of exact failure/event times (Bogaerts et al., 2017).

The presence of interval-censoring can introduce significant complexity to the inferential framework. Early research focused on nonparametric estimation of the underlying distribution function via self-consistency equations (Turnbull, 1974). Many authors delved into exploring the asymptotic behavior and developing computing algorithms for both PIC (Zhao et al., 2008), and DC (Wellner and Zhan, 1997) endpoints. For regression, several authors considered a class of asymptotically efficient, yet computationally challenging semiparametric transformation models, which includes propor-

tional hazards (PH) and proportional odds (PO) models, using generalized expectation-maximization (EM) algorithm (Pan et al., 2020), or direct maximum likelihood (ML) estimation (Choi and Huang, 2021). When the proportional hazards (PH) and proportional odds (PO) assumptions are violated, which is common in cancer trials, a more intuitive alternative is to use an accelerated failure time (AFT) model (Choi et al., 2021). In this model, the resulting regression estimators can be expressed as a mean ratio, specifically, the ratio of the mean values of the time to the event through a linear model. This approach offers a direct evaluation of the association between event time and covariates.

In this article, we propose a unified rank-based inferential procedure for fitting the AFT regression model to data with PIC and DC endpoints. Let T_i be the failure time, and X_i the p -vector of covariates. The AFT model specifies that

$$\log T_i = \beta' X_i + \varepsilon_i, \quad (i = 1, \dots, n), \quad (1.1)$$

where β is a p -vector of unknown regression parameters and ε_i is a random error with a common but unknown distribution. In contrast to traditional hazard-based approaches, such as the Cox proportional hazards (PH), or proportional odds (PO) models, our AFT specification in (1.1) directly quantifies the acceleration or deceleration of log-transformed survival times

with covariates, thereby offering a more intuitive linear interpretation while bypassing the often restrictive PH/PO assumptions. The classical PH and PO models constructed via the class of linear transformation models (Chen et al., 2002) are amenable to a linear interpretation only under specific parametric choices of the (random) error term, whereas, our AFT framework leaves the error term unspecified (and thus provide flexibility), yet maintaining the linear/direct relationship between the time-to-event response and covariates. This feature allows researchers to assess how different factors influence the timing of events, enhancing the practical application of survival models in various settings.

Various techniques have been developed for estimating regression parameters within the right-censored AFT framework (Jin et al., 2003; Choi and Choi, 2021). However, conducting statistical inference with general interval-censored data presents significant challenges, often requiring the simultaneous estimation of both the regression parameters and the underlying residual distribution. To address these challenges, researchers have employed the Buckley-James method (Buckley and James, 1979) to handle progressively interval-censored (PIC) data (Gao et al., 2017) and doubly censored (DC) data (Choi et al., 2021), where the error distribution is approximated using a modified self-consistency algorithm. Li and Pu (2003)

introduced a maximum correlation estimating method for AFT models under interval-censoring, but this approach is restricted to scenarios involving only a single covariate. Additional relevant work includes investigations into covariate analysis for Case-2 interval-censored data (Tian and Cai, 2006), although these methods are often ad-hoc and challenging to generalize. For Case-1 interval-censored data (commonly referred to as “current status” data), Groeneboom and Hendrickx (2018) explored nonparametric maximum likelihood (NPML) estimation methods for linear regression, utilizing a kernel-smoothing approach.

In this article, building on the work of Jin et al. (2003), we explore a class of weighted log-rank estimating functions for general interval-censored data. Our research shows that the proposed estimators are both consistent and asymptotically normal, a conclusion grounded in standard empirical process theory. A significant advantage of our estimator over existing approaches (Gao et al., 2017; Choi et al., 2021) is that it can directly estimate the regression parameter without needing to nonparametrically estimate the residual distribution function, a process that becomes particularly complex under interval-censoring. To facilitate variance estimation, we introduce an efficient resampling technique (Zeng and Lin, 2008) for estimating covariance matrices for a class of log-rank estimators. This approach elim-

inates the need to solve estimating equations for resampled data, thereby improving computational efficiency. Additionally, we extend our proposed methods to multivariate scenarios, where cluster sizes may influence event times (Kim, 2010).

The rest of the paper is organized as follows. Section 2 motivates our proposed rank-based approach using a simple data example, leading to the Gehan statistic. Section 3 develops the proposed rank regression framework, along with parameter and standard error estimation for PIC and DC responses. Section 4 studies the finite-sample properties of our proposal using simulated data. The methodology is illustrated in Section 5 through an application to a data example from a colorectal cancer study. Finally, Section 6 offers concluding remarks.

2. Motivation

To illustrate our method, let $\{1, 2-, 3, 4+, 5\}$ be a simple data of event times. Here, $\{1, 3, 5\}$ represent exact observations, $\{2-\}$ is a left-censored observation at 2, and $\{4+\}$ is a right-censored observation at 4. When conducting a linear rank test, our primary goal is to determine the order of two observations within all comparable pairs. For example, although $\{2-\}$ is left-censored and not fully observable, we can immediately know that

$2- < 3$, $2- < 4+$, and $2- < 5$, thereby fixing the order in these cases. However, the ordering between 1 and $2-$ remains indeterminate due to the censored nature of the observations. Assuming that censoring times are independent of event times, we can perform a Wilcoxon test on this dataset by considering only the comparable pairs, such as $(2-, 3)$, $(2-, 4+)$, and $(2-, 5)$, and determining their ranks within each group.

To be more specific, let us denote observations from two samples by $\{x_1, x_2, \dots, x_n\}$ and $\{y_1, y_2, \dots, y_m\}$, which are subject to double-censoring. After pooling the sample of $(n + m)$ observations into a single group, say $\{t_1, t_2, \dots, t_{n+m}\}$, we can compare each individual with the remaining $n + m - 1$. For comparing the i th individual with the j th one, define $G_i = \sum_{j=1}^{m+n} G_{ij}$, where

$$G_{ij} = \begin{cases} +1, & \text{if } x_i > y_j \text{ or } x_i \geq y_j - \text{ or } x_{i+} \geq y_j \text{ or } x_i + \geq y_{j-} \\ -1, & \text{if } x_i < y_j \text{ or } x_i - \leq y_j \text{ or } x_i \leq y_{j+} \text{ or } x_i - \leq y_{j+} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for the i th individual, G_i is the number of observations that are definitely less than t_i minus the number of observations that are definitely greater than t_i , i.e., $G_i = \#\{j : t_i > t_j\} - \#\{j : t_i < t_j\}$. Then, the Gehan

statistic (Gehan, 1965) can be defined as

$$G = \sum_{i=1}^{n+m} G_i I(i \text{ in group 1}). \quad (2.2)$$

When we consider the null hypothesis that there is no rank difference between two groups, the statistic G is expected to have a mean of 0. We can estimate its variance as $\widehat{\text{Var}}(G) = \frac{mn}{(m+n)(m+n-1)} \sum_{i=1}^{m+n} G_i^2$. This observation allows us to perform a hypothesis test for comparing two groups within a general interval-censoring framework. This approach can also be extended to address a linear regression problem involving multiple covariates. By utilizing this idea, we compare the ranks of residuals obtained from the linear model, making it a suitable approach for dealing with general interval-censored endpoints, which will be discussed in the subsequent sections.

3. Proposed Methods

3.1 Partly interval-censored (PIC) rank regression

We first consider our estimating procedure for PIC data. Suppose that there is a sequence of K examination times from periodic clinical visits and let $0 = W_0 < W_1 < W_2 < \dots < W_K < W_{K_i} = \infty$. We can identify the tightest interval of the examination times (U, V) containing T , i.e.,

3.1 Partly interval-censored (PIC) rank regression

$U = \max_k\{W_k : W_k \leq T\}$ and $V = \min_k\{W_k : W_k \geq T\}$. We use Δ to represent the censoring indicator, which takes on the value 1 for exact events and 0 for IC cases. When $\Delta_i = 1$, we set $U = V = T$. Thus, the PIC dataset is structured as follows: $\{(\Delta_i, \Delta_i T_i, (1 - \Delta_i)U_i, (1 - \Delta_i)V_i, X_i), i = 1, \dots, n\}$. Let $\max(a, b) = a \vee b$ and $\min(a, b) = a \wedge b$. This setup can also be summarized as $\{(\tilde{U}_i, \tilde{V}_i, \Delta_i, X_i), i = 1, \dots, n\}$, where $\tilde{U}_i = T_i \vee U_i = \Delta_i T_i + (1 - \Delta_i)U_i$ and $\tilde{V}_i = T_i \wedge V_i = \Delta_i T_i + (1 - \Delta_i)V_i$. Throughout the paper, we assume that the joint distribution of (W_1, \dots, W_K) is independent of T given X , which implies that the visit processes $\{W_k\}$ do not provide any further information about the distribution of T given X (Zhang and Heitjan, 2006), and that the proportion of observing exact observation is non-negligible, i.e., $P(\Delta = 1) > 0$.

The rank statistic for PIC data can be built by comparing $(\tilde{U}_i, \tilde{V}_i)$ and $(\tilde{U}_j, \tilde{V}_j)$ for all pairs of subjects with $i < j$. However, ambiguity arises when either the i th subject or the j th subject, or both, are interval-censored. Now, since $V_i \leq U_j$ implies $T_i \leq T_j$, it suffices to examine the following four rank identifiable inequalities: (i) $T_i \leq T_j$ if $(\Delta_i = 1, \Delta_j = 1)$, (ii) $T_i \leq U_j$ if $(\Delta_i = 1, \Delta_j = 0)$, (iii) $V_i \leq T_j$ if $(\Delta_i = 0, \Delta_j = 1)$, and (iv) $V_i \leq U_j$ if $(\Delta_i = 0, \Delta_j = 0)$. Note that this can be simply accomplished by checking $\tilde{V}_i \leq \tilde{U}_j$ for all $i < j$ pairs, i.e., whether or not the i th upper bound \tilde{V}_i is

3.1 Partly interval-censored (PIC) rank regression

less than the j th lower bound \tilde{U}_j . With covariates, let $u_i(\beta) = \log \tilde{U}_i - \beta' X_i$ and $v_i(\beta) = \log \tilde{V}_i - \beta' X_i$ denote the observed residuals, corresponding to \tilde{U}_i and \tilde{V}_i , respectively. Also, define $e_i(\beta) = \log T_i - \beta' X_i$ as the true error term that is not directly observable from the data.

By generalizing the Wilcoxon test (2.2) to our regression problem, we can formulate the Gehan estimating function as

$$S_n(\beta) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \eta_{2i} \eta_{1j} (X_i - X_j) I\{v_i(\beta) \leq u_j(\beta)\}, \quad (3.3)$$

where $\eta_{1i} = I(\tilde{U}_i > 0) = \Delta_i + (1 - \Delta_i)I(U_i > 0)$ and $\eta_{2i} = I(\tilde{V}_i < \infty) = \Delta_i + (1 - \Delta_i)I(V_i < \infty)$. Simply speaking, equation (3.3) takes averages of the covariate differences only when the ordering relationship between $e_i(\beta)$ and $e_j(\beta)$ is ensured for all possible combinations of (i, j) . Indeed, it is clear that $v_i(\beta) \leq u_j(\beta)$ implies $e_i(\beta) \leq e_j(\beta)$, since the relationship $e_i(\beta) \leq v_i(\beta) \leq u_j(\beta) \leq e_j(\beta)$ holds true in this case.

Nevertheless, as β resides within the indicator functions of $S_n(\beta)$ only, the function lacks differentiability, and typically, a unique solution to this estimating function does not exist. Instead, by recognizing the fact that

3.1 Partly interval-censored (PIC) rank regression

$S_n(\beta)$ is the negative gradient of the convex objective function

$$L_n(\beta) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \eta_{2i} \eta_{1j} \{v_i(\beta) - u_j(\beta)\}_-, \quad (3.4)$$

where $a_- = |a|I(a \leq 0)$, a regression parameter estimate can be obtained by minimizing $L_n(\beta)$ with respect to β . We define the Gehan estimator as

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{B}} L_n(\beta)$$

The optimizer of this minimization problem may not be unique, but the convexity of $L_n(\beta)$ implies that the set of minimizers is convex (Fyngenson and Ritov, 1994). The minimization of the equation (3.4) can be implemented by linear programming, however, efficiency will be lost as larger sample sizes can significantly affect the computational cost. Instead, since $L_n(\beta)$ can be equivalently expressed as

$$\sum_{i=1}^n \sum_{j=1}^n \eta_{2i} \eta_{1j} |v_i(\beta) - u_j(\beta)| + \left| M - \beta' \sum_{k=1}^n \sum_{l=1}^n \eta_{2k} \eta_{1l} (X_l - X_k) \right|, \quad (3.5)$$

where, $M > 0$ is a sufficiently large number, it can be easily solved by using a standard software, e.g., `rq()` function in R `quantreg` package (Koenker, 2008).

Furthermore, following Jin et al. (2003), we can generalize the Gehan

3.1 Partly interval-censored (PIC) rank regression

function to the weighted log-rank estimating function as

$$S_\phi(\beta) = n^{-1} \sum_{i=1}^n \eta_{2i} \phi_i(\beta) \left\{ X_i - \frac{\sum_{j=1}^n \eta_{1j} X_j I\{v_i(\beta) \leq u_j(\beta)\}}{\sum_{j=1}^n \eta_{1j} I\{v_i(\beta) \leq u_j(\beta)\}} \right\}, \quad (3.6)$$

where $\phi_i(\cdot)$ is a data-dependent nonnegative weight for the i th subject. The choices of $\phi_i(\beta) = \sum_{j=1}^n \eta_{1j} I\{u_j(\beta) \geq v_i(\beta)\}$ and $\phi_i(\beta) = 1$ correspond to the Gehan estimating function (3.3) and log-rank function, respectively. The lack of smoothness and monotonicity still imparts computational challenges in solving $S_\phi(\beta) = 0$, particularly with multiple covariates. However, the results of Jin et al. (2003) imply that there exist a sequence of solutions that is strongly consistent for β .

Specifically, consider a class of monotone weighted estimating functions

$$S_\phi(\beta, b) = n^{-1} \sum_{i=1}^n \eta_{2i} \eta_{1j} w_i\{b, v_i(b)\} (X_i - X_j) I\{v_i(\beta) \leq u_j(\beta)\}, \quad (3.7)$$

with the corresponding objective function

$$L_\phi(\beta, b) = n^{-1} \sum_{i=1}^n \eta_{2i} \eta_{1j} w_i\{b, v_i(b)\} \{v_i(\beta) - u_j(\beta)\}_-, \quad (3.8)$$

where $w_i(b, t) = \phi_i(b) / \sum_{j=1}^n I\{u_j(b) \geq t\}$. The minimizer of (3.8) can be obtained iteratively as $\hat{\beta}_{(k)} = \arg \min_{\beta \in \mathbb{B}} L_\phi(\beta, \hat{\beta}_{(k-1)})$, $k \geq 1$, where, the initial value is set to be the Gehan estimator, i.e., $\hat{\beta}_{(0)} = \hat{\beta}$. We define the

3.1 Partly interval-censored (PIC) rank regression

log-rank estimator by $\hat{\beta}_\phi = \lim_{k \rightarrow \infty} \hat{\beta}_{(k)}$. In our experience, these estimators converge in about 5–10 iterative steps, which is not computationally burdensome, and easy to implement. In Section 3.2, we demonstrate that $\hat{\beta}_{(k)}$ can be represented as a weighted average of the solutions to $S_n(\beta)$ and $S_\phi(\beta)$. This shows that $\hat{\beta}_{(k)}$ is consistent for any $k \geq 1$ and converges to the solution of $S_\phi(\beta)$ as $k \rightarrow \infty$.

The methods developed for PIC data can be adapted to DC data with only minor adjustments. For the i th subject, let (T_i, L_i, R_i) represent the exact event time, left-censoring time, and right-censoring time, respectively. In the context of DC data, we observe the data $\{(\tilde{T}_i, \delta_i, X_i) \mid i = 1, \dots, n\}$, where $\tilde{T}_i = L_i \vee (T_i \wedge R_i)$ and $\delta_i = (\delta_{1i}, \delta_{2i}, \delta_{3i})$ with $\delta_{1i} = I(L_i < T_i \leq R_i)$, $\delta_{2i} = I(T_i > R_i)$, and $\delta_{3i} = I(T_i \leq L_i)$. Here, $\tilde{T}_i = T_i$ only if $\delta_{1i} = 1$; otherwise, it is either right-censored ($\delta_{2i} = 1$) or left-censored ($\delta_{3i} = 1$). DC data can thus be viewed as a special case of PIC data by defining $(U_i, V_i) = (0, L_i)$ for left-censoring and $(U_i, V_i) = (R_i, \infty)$ for right-censoring. This similarity allows for a unified estimation approach for both PIC and DC data structures. Detailed procedures for estimation under the DC setup are provided in the Supplementary Material S2.

3.2 Asymptotic properties

In this section, we establish the asymptotic properties of the proposed rank estimators $\hat{\beta}$ and $\hat{\beta}_\phi$. All technical details (proofs of the Theorems and related Lemmas) are relegated to the Supplementary Material S1. We first impose the following regularity conditions.

- (C1) The true value of β , denoted by β_0 , lies in the interior of a known compact set $\mathbb{B} \subset \mathbb{R}^p$. The covariate X is uniformly bounded, i.e., $\sup_i \|X_i\| < \infty$ for $i = 1, \dots, n$.
- (C2) The residual distribution $F_0 \in \mathcal{F}$ is uniformly bounded away from 0, and has a density with continuous derivative bounded away from 0 on their support.
- (C3) The distribution of Δ depends only on the observed data $\{\Delta, \Delta T, (1-\Delta)U, (1-\Delta)V, X\}$. There exists a positive constant c_0 such that $P(\Delta = 1|X) > c_0$ with probability 1.
- (C4) The joint density of the examination times (W_1, \dots, W_K) given $\Delta = 0$ is continuous and differentiable in their support with respect to some dominating measure. There exists a positive constant τ_0 such that $P(\min_{0 \leq k \leq K-1} (W_{k+1} - W_k) > \tau_0 | X, K, \Delta = 0) = 1$.

Note that, (C1) is a standard assumption in survival analysis implying

the compactness of the parameter space with Euclidean norm and boundedness of covariates, while (C2), (C3) and (C4) are essential assumptions that guarantee identifiability of the regression parameters, and strong consistency of their estimators. In particular, (C2) and (C4) state the smoothness conditions for the underlying distribution function. Condition (C3) implies that the IC variables (U, V) do not convey any additional information on the law of T apart from assuming T to be bracketed by U and V . This implies that the visit process that generates the (U, V) is independent of T given X , and the U and V are constructed from this visit process and from T . In addition, the proportion of observing exact time is non-ignorable to ensure the PIC setup.

Theorem 1. *Under conditions (C1)–(C4), the proposed regression estimator $\hat{\beta}$ is strongly consistent for β_0 , and $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a zero-mean normal distribution with covariance matrix $\Gamma = A^{-1}\Omega(A^{-1})'$.*

Theorem 2. *Under conditions (C1)–(C4), for any $k \geq 1$, an iterative estimator $\hat{\beta}_{(k)}$ is strongly consistent for β_0 , and $n^{1/2}(\hat{\beta}_{(k)} - \beta_0)$ converges to the same distribution of $n^{1/2}(\hat{\beta}_\phi - \beta_0)$ as $k \rightarrow \infty$.*

Theorem 1 states the asymptotic behavior of $\hat{\beta}$. Here, $A = E\{\partial\psi_\beta/(\partial\beta)|_{\beta=\beta_0}\}$ and $\Omega = E\{\psi'_{\beta_0}\psi_{\beta_0}\}$, where we let ψ_β denote the influence function of $S_n(\beta)$.

The consistency result can be simply proved by using standard arguments of

the Glivenko-Cantelli theorem by defining a proper function class composed of indicator functions. For asymptotic normality, we apply the standard asymptotic results of the Z -estimator to our estimating functions. Theorem 2 shows that the asymptotic behavior of $\hat{\beta} = \lim_{k \rightarrow \infty} \hat{\beta}_{(k)}$ is equivalent to that of $\hat{\beta}_\phi$ as the iteration proceeds.

3.3 Variance estimation

A traditional way to estimate standard error for β is to employ resampling-based methods, such as bootstrap and multiplier sampling (Jin et al., 2003, 2006a). However, such multiple resampling-based procedures are usually inefficient, with enhanced computational cost. Instead, we utilize an efficient resampling procedure (Zeng and Lin, 2008) for variance estimation. Note that we can asymptotically write $S_n(\beta)$ as follows:

$$n^{1/2}S_n(\beta) = n^{1/2} \sum_{i=1}^n \psi_{\beta_0, i} + n^{1/2}(\beta - \beta_0)A + o_p(1 + n^{1/2}\|\beta - \beta_0\|),$$

where A is the $p \times p$ invertible matrix (as in Theorem 1), equivalent to the asymptotic slope of $S_n(\beta)$ at β_0 . To approximate Ω , we use perturbed resampling approach (Jin et al., 2006b), which provides more reliable results than standard bootstrapping under multivariate clustered data. This is partly because bootstrapping under informatively clustered data possibly

increases the imbalance in the bootstrap data. Let R be a fixed number of the perturbed resamplings and $S_n^*(\beta)$ be a perturbed estimating function as

$$S_n^*(\beta) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n Z_i \eta_{2i} \eta_{1j} (X_i - X_j) I\{v_i(\beta) \leq u_j(\beta)\},$$

where, $Z_i \sim$ i.i.d. $\text{Exp}(1)$. Given the sample, $n^{1/2}S_n^*(\hat{\beta})$ asymptotically has a zero-mean normal distribution with the covariance matrix Ω (van der Vaart and Wellner, 1996). Our variance estimation proceeds as follows:

Step 1. Calculate $R \times p$ matrix $n^{1/2}S_n^*(\hat{\beta})$ and approximate Ω by $\hat{\Omega} = \widehat{\text{Cov}}\{n^{1/2}S_n^*(\hat{\beta})\}$.

Step 2. Calculate $R \times p$ matrix $n^{1/2}S_n(\hat{\beta} + n^{-1/2}K_r)$, where K_r ($r = 1, \dots, R$) is a p -dimensional zero-mean independent random vector, such as $N(0, 1)$.

Step 3. Regress the j th row of $n^{1/2}S_n(\hat{\beta} + n^{-1/2}K_r)$ on K_r for $j = 1, \dots, p$ and $r = 1, \dots, R$, and let \hat{A} be the matrix whose j th row is the j th least squares estimate.

Step 4. Estimate the covariance matrix of $n^{1/2}(\hat{\beta} - \beta_0)$ by $\hat{\Gamma} = \hat{A}^{-1}\hat{\Omega}(\hat{A}^{-1})'$.

We use a similar inferential method for rank estimation with general weight function. Our numerical studies show that this procedure can pro-

3.4 Extension to multivariate partly interval-censored data

duce variance estimators very reliably, achieving nominal coverage probabilities under both PIC and DC settings.

3.4 Extension to multivariate partly interval-censored data

Next, we extend our methods to handle clustered interval-censored data. This scenario often occurs when subjects are sampled within clusters, leading to correlated failure times among subjects within the same cluster. Suppose that there are n clusters, with the i th cluster having m_i members ($m_i \ll n$). For the k th member in the i th cluster, let (T_{ik}, U_{ik}, V_{ik}) denote a tuple of the exact observation, left and right examination times, and $\Delta_{ik} = 1 - I(U_{ik} \leq T_{ik} < V_{ik})$, the censoring indicator, such that $\Delta_{ik} = 1$ when T_{ik} is exactly observed. The observed multivariate event data under PIC can be represented as $\{(\Delta_{ik}, \Delta_{ik}T_{ik}, (1 - \Delta_{ik})U_{ik}, (1 - \Delta_{ik})V_{ik}, X_{ik}), k = 1, \dots, m_i, i = 1, \dots, n\}$. Also, define $\tilde{U}_{ik} = T_{ik} \vee U_{ik} = \Delta_{ik}T_{ik} + (1 - \Delta_{ik})U_{ik}$ and $\tilde{V}_{ik} = T_{ik} \wedge V_{ik} = \Delta_{ik}T_{ik} + (1 - \Delta_{ik})V_{ik}$.

We assume the marginal distribution of T_{ik} follows the AFT model

$$\log T_{ik} = \beta' X_{ik} + \varepsilon_{ik}, \quad (i = 1, \dots, n; k = 1, \dots, m_i), \quad (3.9)$$

where $(\varepsilon_{i1}, \dots, \varepsilon_{im_i})$, $i = 1, \dots, n$ are independent random vectors. Within the i th cluster, the error terms, $\varepsilon_{i1}, \dots, \varepsilon_{im_i}$, are assumed to be exchangeable

3.4 Extension to multivariate partly interval-censored data

with a common marginal distribution F . Let $u_{ik}(\beta) = \log \tilde{U}_{ik} - \beta' X_{ik}$ and $v_{ik}(\beta) = \log \tilde{V}_{ik} - \beta' X_{ik}$ denote the observed residuals under model (3.9). To estimate the regression parameter β , we can solve the generalized log-rank estimating function

$$S_{\phi}^{\dagger}(\beta) = n^{-1} \sum_{i=1}^n \varphi_i \sum_{k=1}^{m_i} \eta_{2ik} \phi_{ik}(\beta) \left[X_{ik} - \frac{\sum_{j=1}^n \varphi_j \sum_{l=1}^{m_j} \eta_{1jl} X_{jl} I\{v_{ik}(\beta) \leq u_{jl}(\beta)\}}{\sum_{j=1}^n \varphi_j \sum_{l=1}^{m_j} \eta_{1jl} I\{v_{ik}(\beta) \leq u_{jl}(\beta)\}} \right], \tag{3.10}$$

where $\eta_{1ik} = \Delta_{ik} + (1 - \Delta_{ik})I(U_{ik} > 0)$, $\eta_{2ik} = \Delta_{ik} + (1 - \Delta_{ik})I(V_{ik} < \infty)$, and φ_i is a known weight to adjust for possible informative cluster sizes (ICS), a setup where the cluster size can be correlated to the survival time of interest (Lam et al., 2021). By convention, we may use $\varphi_i = 1$, which tends to overweight the large clusters, because each individual contributes equally in the estimating equation. However, when cluster sizes are informative to the outcome of interest, we can incorporate the inverse of cluster sizes as a weight, for example, $\varphi_i = 1/m_i$ or $\varphi_i = 1/m_i^{\alpha}$ for some $0 \leq \alpha \leq 1$, to relieve the cluster-size effect. This adjustment is also known to increase statistical efficiency (Wang et al., 2008). As before, $\phi_{ik}(\beta)$ is a weight function, leading to the Gehan estimator and the log-rank estimator, respectively, if $\phi_{ik}(\beta) = \sum_{j=1}^n \varphi_j \sum_{l=1}^{m_j} \eta_{1jl} I\{v_{ik}(\beta) \leq u_{jl}(\beta)\}$ and $\phi_{ik}(\beta) = 1$.

3.4 Extension to multivariate partly interval-censored data

For estimation, we consider the monotone modification of $S_\phi^\dagger(\beta)$ as

$$S_\phi^\dagger(\beta, b) = n^{-1} \sum_{i=1}^n \sum_{k=1}^{m_i} \sum_{j=1}^n \sum_{l=1}^{m_j} \varphi_i \varphi_j w_{ik} \{b, v_{ik}(b)\} \eta_{2ik} \eta_{1jl} (X_{ik} - X_{jl}) I\{v_{ik}(\beta) \leq u_{jl}(\beta)\}, \quad (3.11)$$

where $w_{ik}(b, t) = \phi_{ik}(b) / \sum_{j=1}^n \sum_{l=1}^{m_j} \varphi_j \eta_{2jl} I\{u_{jl}(b) \geq t\}$. Note that $S_\phi^\dagger(\beta, b)$ is componentwise monotone in β , with the gradient of the convex function given by

$$L_\phi^\dagger(\beta, b) = n^{-1} \sum_{i=1}^n \sum_{k=1}^{m_i} \sum_{j=1}^n \sum_{l=1}^{m_j} \varphi_i \varphi_j w_{ik} \{b, v_{ik}(b)\} \eta_{2ik} \eta_{1jl} \{v_{ik}(\beta) \leq u_{jl}(\beta)\}_-, \quad (3.12)$$

which can be minimized again via linear programming or minimizing ℓ_1 -type convex objective function. Let $\hat{\beta}_{\phi, (0)}^\dagger = \hat{\beta}^\dagger$. The Gehan estimator $\hat{\beta}^\dagger$ is easy to implement since $w_{ik} \equiv 1$. The minimization for the log-rank estimator is carried out iteratively, i.e., $\hat{\beta}_{\phi, (k)}^\dagger = \arg \min_{\beta \in \mathbb{B}} L_\phi^\dagger(\beta, \hat{\beta}_{\phi, (k-1)}^\dagger)$ ($k \geq 1$). If the iterative algorithm converges as $k \rightarrow \infty$, the limit, say $\hat{\beta}_\phi^\dagger$, satisfies the original estimating equation $S_\phi^\dagger(\beta)$. The consistency and asymptotic normality of $\hat{\beta}^\dagger$ and $\hat{\beta}_\phi^\dagger$ can be established using similar methods to those outlined for the asymptotic results in Section 3.2.

To account for the cluster structure in the variance estimation of the marginal model (Jin et al., 2006b; Xu et al., 2023), we propose using the

following perturbed estimating function:

$$S_n^{\dagger*}(\beta) = n^{-1} \sum_{i=1}^n \sum_{k=1}^{m_i} \sum_{j=1}^n \sum_{l=1}^{m_j} Z_i Z_j \varphi_i \varphi_j \eta_{2ik} \eta_{1jl} (X_{ik} - X_{jl}) I\{v_{ik}(\beta) \leq u_{jl}(\beta)\},$$

where perturbation variables Z_i and Z_j are included to reflect the cluster structure (Jin et al., 2006b). To approximate Ω^\dagger , we use $\hat{\Omega}^\dagger = \widehat{\text{Cov}}\{n^{1/2}S_n^{\dagger*}(\hat{\beta})\}$.

We then update the j th row of the matrix A^\dagger with \hat{A}^\dagger , obtained by regressing $n^{1/2}S_n^{\dagger*}(\hat{\beta}^\dagger + n^{-1/2}K_r)$ on K_r (for $j = 1, \dots, p$). Finally, the covariance matrix of $n^{1/2}(\hat{\beta}^\dagger - \beta_0)$ can be estimated by $\hat{\Gamma}^\dagger = \hat{A}^{\dagger-1}\hat{\Omega}^\dagger(\hat{A}^{\dagger-1})'$.

4. Simulation Studies

In this section, we present several simulation results using synthetic data to assess the finite-sample performance of our estimates for PIC and DC data, in both univariate and multivariate scenarios. Our method is implemented in the R package `rankIC`, which includes a detailed vignette. The package is available at the following link: <https://github.com/taehwa015/rankIC>.

4.1 Scenario 1: Univariate data

Here, we generate failure times from the AFT model, $\log T = 2 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where $X_1 \sim N(0, 1)$ and $X_2 \sim \text{Bernoulli}(0.5)$, and the true regression parameter is set to $(\beta_1, \beta_2) = (1, 1)$. The residual ε is generated

4.1 Scenario 1: Univariate data

Table 1: Simulation results for PIC data. Table entries are the average bias (Bias), empirical standard error (ESE), asymptotic standard error (ASE), and coverage probability (CP) of the 95% Wald-type confidence intervals for the parameter estimates obtained from the Gehan and log-rank methods, under $n = 200$ and 400 , censoring rates of 30% (right-censoring: 6% and interval-censoring: 24%) and 60% (left-censoring: 1%, right-censoring: 7% and interval-censoring: 52%), and error distributions that follow Normal(0,1) denoted as $N(0, 1)$, Extreme Value (EV), and Exponential(1), denoted as Exp(1).

Cens	Error	n	Par	Gehan				Log-rank			
				Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
30%	$N(0, 1)$	200	β_1	-0.005	0.076	0.074	0.942	0.007	0.081	0.079	0.941
			β_2	0.000	0.146	0.146	0.946	0.003	0.161	0.156	0.943
		400	β_1	0.000	0.053	0.052	0.938	0.010	0.060	0.056	0.919
			β_2	0.001	0.103	0.102	0.948	0.007	0.110	0.110	0.944
	EV	200	β_1	0.001	0.084	0.083	0.936	0.019	0.216	0.109	0.942
			β_2	0.001	0.162	0.165	0.951	0.000	0.238	0.216	0.941
		400	β_1	0.001	0.059	0.058	0.947	0.008	0.078	0.077	0.952
			β_2	0.005	0.113	0.115	0.948	0.010	0.151	0.152	0.948
	Exp(1)	200	β_1	-0.001	0.046	0.048	0.965	0.003	0.075	0.076	0.962
			β_2	0.003	0.089	0.091	0.960	0.013	0.150	0.148	0.947
		400	β_1	-0.001	0.031	0.032	0.951	0.005	0.053	0.053	0.952
			β_2	-0.003	0.062	0.062	0.954	-0.002	0.104	0.103	0.957
60%	$N(0, 1)$	200	β_1	0.000	0.074	0.075	0.948	0.023	0.085	0.083	0.938
			β_2	0.003	0.155	0.148	0.935	0.016	0.173	0.160	0.929
		400	β_1	-0.003	0.053	0.053	0.945	0.018	0.058	0.058	0.930
			β_2	0.000	0.098	0.104	0.960	0.019	0.109	0.113	0.952
	EV	200	β_1	-0.004	0.084	0.083	0.946	0.015	0.112	0.109	0.941
			β_2	0.005	0.163	0.164	0.945	0.014	0.219	0.213	0.939
		400	β_1	-0.002	0.058	0.058	0.945	0.016	0.081	0.076	0.931
			β_2	-0.008	0.117	0.115	0.944	0.008	0.159	0.150	0.936
	Exp(1)	200	β_1	-0.002	0.046	0.048	0.958	0.011	0.076	0.076	0.952
			β_2	0.000	0.084	0.090	0.967	0.012	0.137	0.146	0.963
		400	β_1	-0.003	0.031	0.032	0.965	0.011	0.050	0.053	0.965
			β_2	0.000	0.062	0.062	0.957	0.009	0.101	0.102	0.961

from one of the following three underlying distributions: (i) standard normal distribution (“ $N(0, 1)$ ”), (ii) extreme value (EV) distribution (“EV”), and (iii) exponential distribution (“Exp(1)”). All simulation results are obtained based on 1000 data replications with sample sizes $n = 200$ or 400 and $R = 200$ perturbations.

We first examine the PIC setup, where the proportion of exact obser-

4.1 Scenario 1: Univariate data

Table 2: Simulation results for DC data. Table entries are the average bias (Bias), empirical standard error (ESE), asymptotic standard error (ASE), and coverage probability (CP) of the 95% Wald-type confidence intervals for the parameter estimates obtained from the Gehan and log-rank methods, under $n = 200$ and 400 , $(\pi_L, \pi_R) = (15\%, 15\%)$ and $(30\%, 30\%)$, with π_L and π_R denoting proportions of left- and right-censoring, respectively.

(π_L, π_R)	Error	n	Par	Gehan				Log-rank			
				Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
(15%,15%)	$N(0,1)$	200	β_1	-0.003	0.086	0.084	0.948	0.025	0.093	0.088	0.925
			β_2	0.003	0.168	0.163	0.934	0.005	0.180	0.173	0.930
		400	β_1	-0.004	0.060	0.059	0.942	0.024	0.062	0.062	0.930
			β_2	-0.002	0.111	0.115	0.958	0.001	0.119	0.122	0.963
	EV	200	β_1	-0.004	0.095	0.095	0.945	0.020	0.086	0.085	0.931
			β_2	-0.019	0.178	0.183	0.953	-0.012	0.163	0.168	0.953
		400	β_1	-0.005	0.064	0.066	0.953	0.019	0.059	0.060	0.944
			β_2	-0.011	0.129	0.129	0.953	-0.002	0.117	0.118	0.948
	Exp(1)	200	β_1	-0.002	0.055	0.058	0.965	0.028	0.085	0.086	0.947
			β_2	-0.005	0.106	0.111	0.970	-0.008	0.162	0.168	0.959
		400	β_1	-0.002	0.037	0.040	0.968	0.028	0.057	0.060	0.939
			β_2	-0.002	0.075	0.076	0.965	0.001	0.116	0.117	0.952
(30%,30%)	$N(0,1)$	200	β_1	-0.004	0.106	0.102	0.932	0.006	0.111	0.105	0.925
			β_2	0.003	0.206	0.205	0.938	-0.026	0.211	0.209	0.939
		400	β_1	-0.005	0.071	0.072	0.945	0.004	0.072	0.073	0.950
			β_2	-0.001	0.142	0.144	0.959	-0.028	0.148	0.148	0.942
	EV	200	β_1	-0.006	0.118	0.116	0.942	0.005	0.110	0.106	0.941
			β_2	-0.024	0.232	0.233	0.945	-0.045	0.214	0.213	0.935
		400	β_1	-0.008	0.082	0.082	0.943	0.002	0.075	0.075	0.945
			β_2	-0.022	0.160	0.164	0.957	-0.042	0.146	0.151	0.941
	Exp(1)	200	β_1	-0.006	0.073	0.074	0.954	0.003	0.099	0.096	0.948
			β_2	-0.004	0.145	0.151	0.962	-0.034	0.198	0.201	0.952
		400	β_1	-0.006	0.050	0.051	0.950	0.002	0.067	0.068	0.956
			β_2	0.003	0.101	0.103	0.965	-0.023	0.136	0.139	0.958

variations is given by $p^* = p_0 - 0.1 \times I(X_2 = 1)$, with $p_0 \in (0, 1)$ chosen to achieve the desired censoring rate. Conversely, with probability $1 - p^*$, the data are subject to interval-censoring. In this scenario, a sequence of random examination times $\{W_k \mid k = 1, \dots, m\}$ is generated by setting $W_k - W_{k-1} \sim \text{Uniform}(0.1, 1)$, ensuring that $0 = W_0 < W_1 < \dots < W_m < \tau$, where τ is the maximum follow-up time, set to 100. Then, we can create the IC data $(U = W_k, V = W_{k+1})$ if $W_k < T < W_{k+1}$ for $k \leq m - 1$. If

4.1 Scenario 1: Univariate data

Table 3: Comparing average Bias (Bias), mean squared error (MSE) and relative efficiency (RE) of the parameter estimates obtained from the Gehan and log-rank methods to the Buckley-James (BJ) type estimators, for data generated under the PIC and DC settings, where $n = 400$, with errors distributed as $N(0, 1)$, Extreme Value (EV), and $\text{Exp}(1)$, and under censoring rates of 20% and 40%. The reported MSEs of the estimators are multiplied by 100. The Buckley-James (BJ) approach for PIC was implemented via the Gao et al. (2017)'s method

Type	Error	Cens	Par	Buckley-James		Gehan			Log-rank			
				Bias	MSE	Bias	MSE	RE	Bias	MSE	RE	
PIC	$N(0, 1)$	20%	β_1	-0.001	0.3	0.000	0.3	0.840	0.006	0.3	0.840	
			β_2	0.001	1.0	0.002	1.0	0.996	0.005	1.2	0.830	
		40%	β_1	-0.001	0.3	-0.004	0.3	0.843	0.010	0.3	0.843	
			β_2	0.001	1.0	0.000	1.1	0.923	0.009	1.3	0.781	
		EV	20%	β_1	-0.001	0.4	-0.001	0.4	1.075	0.003	0.6	0.717
				β_2	0.005	1.5	0.000	1.3	1.155	0.004	2.2	0.683
	Exp(1)	40%	β_1	0.000	0.4	-0.002	0.3	1.407	0.008	0.6	0.703	
			β_2	0.008	1.5	-0.003	1.3	1.172	0.002	2.3	0.662	
		20%	β_1	0.001	0.2	-0.001	0.1	2.490	0.003	0.3	0.830	
			β_2	0.005	1.0	-0.003	0.4	2.485	-0.002	1.0	0.994	
		40%	β_1	0.001	0.2	0.001	0.1	2.440	0.009	0.3	0.813	
			β_2	0.005	1.0	-0.001	0.4	2.490	0.005	1.0	0.996	
DC	$N(0, 1)$	20%	β_1	0.002	0.3	-0.003	0.3	0.966	0.028	0.4	0.730	
			β_2	0.008	1.1	0.001	1.1	1.002	0.016	1.3	0.828	
		40%	β_1	0.003	0.4	-0.005	0.4	0.993	0.016	0.5	0.887	
			β_2	0.011	1.4	-0.007	1.4	1.000	-0.019	1.6	0.865	
		EV	20%	β_1	0.004	0.5	-0.001	0.4	1.257	0.049	0.9	0.547
				β_2	-0.004	1.9	-0.006	1.6	1.199	0.012	2.6	0.720
	Exp(1)	40%	β_1	0.006	0.6	-0.005	0.5	1.227	0.032	0.9	0.698	
			β_2	-0.003	2.3	-0.018	1.9	1.173	-0.040	3.1	0.728	
		20%	β_1	0.001	0.3	-0.001	0.1	2.428	0.033	0.4	0.704	
			β_2	0.004	1.2	0.000	0.5	2.434	0.016	1.3	0.926	
		40%	β_1	0.002	0.4	-0.004	0.2	2.339	0.019	0.4	0.982	
			β_2	0.011	1.4	-0.002	0.6	2.211	-0.014	1.4	0.982	

$T \geq W_m$, we treat T as right-censored at W_m .

We summarize the simulation results comparing the performance of the Gehan and log-rank estimators based on average bias (Bias), empirical standard error (ESE), asymptotic standard error (ASE), and coverage probability of 95% confidence intervals (CP). Table 1 presents the results for sample sizes of $n = 200$ and 400 under two scenarios: 30% censoring (6% right-censoring and 24% interval-censoring) and 60% censoring (1%

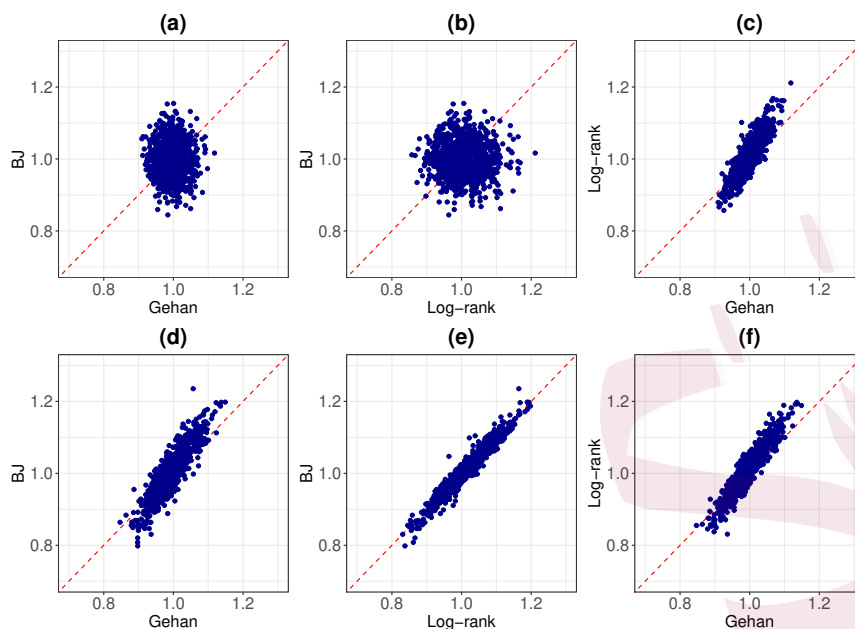


Figure 1: Scatter plots of point estimates of β_1 obtained from the Gehan, log-rank and Buckley-James (BJ) methods, under partly interval-censoring (panels (a)–(c)) and double-censoring (panels (d)–(f)).

left-censoring, 7% right-censoring, and 52% interval-censoring) rates. Overall, the proposed estimators are essentially unbiased, and the ASEs closely match their ESEs. Furthermore, the empirical CPs align well with the nominal levels as predicted by normal approximations. As expected, accuracy improves with larger sample sizes, as evidenced by reductions in MSEs when n increases from 200 to 400. The Gehan and log-rank estimators show similar performance under $N(0, 1)$ and extreme value (EV) error distributions; however, the Gehan estimator outperforms the log-rank estimator when $\text{Exp}(1)$ errors are present. The results corresponding to much larger

n , i.e., $n = 1000$ and 2000 , as presented in Table S1 reveal much smaller Bias, smaller ESEs, and adequate CPs.

In the DC setup, the left-censoring and right-censoring variables (L, R) are simulated as follows: $\log L \sim (1 - 0.25X_1) \times \text{Uniform}(-6, c_L)$ and $\log R \sim \log L + (1 - 0.5X_2) \times \text{Uniform}(6, c_R)$, respectively. The constants (c_L, c_R) are chosen to achieve the desired left- and right-censoring proportions: $(\pi_L, \pi_R) = (15\%, 15\%)$ and $(30\%, 30\%)$. Table 2 summarizes the simulation results for the DC data. Similar to the PIC setup, the parameter estimates exhibit small biases that tend to decrease as the sample size n increases. The standard error estimates accurately reflect the true variability, and the confidence intervals demonstrate appropriate coverage probabilities.

We also compare the statistical efficiency of our rank estimators with those obtained from the Buckley-James (BJ) method under both the PIC (Gao et al., 2017) and DC (Choi et al., 2021) settings. Data for the PIC and DC scenarios are generated as described above, with censoring rates of 20% and 40%, and a sample size of $n = 400$. Table 3 presents the mean bias and mean squared error (MSE) of these estimators, along with the relative efficiency (RE) of our rank estimators (to the BJ estimator), defined as the ratio of the MSE from the BJ method to the two proposed

estimators. Recall that the BJ estimator, which is essentially a least-square estimator for censored data (Lai and Ying, 1991), yields the most efficient maximum likelihood estimate when the residuals are Gaussian. Even under these conditions, our Gehan estimator performs comparably to the BJ estimator. However, for non-normal residual distributions, the Gehan estimator demonstrates superior performance compared to the other methods. Notably, when the error distribution is highly skewed, as in the Exp(1) case, the Gehan estimator achieves MSEs that are nearly 2.5 times smaller than those of the BJ estimator. Additionally, the Gehan estimator yields consistent MSEs across various error distributions, whereas the MSEs of the other methods are more variable, possibly due to their iterative computation processes. This is further illustrated in Figure 1, which shows scatter plots of 1,000 regression estimates of β_1 , comparing the BJ, Gehan, and log-rank methods under the PIC (upper panel; panels (a)–(c)) and DC (lower panel; panels (d)–(f)) settings. In this scenario, the estimates from the BJ and log-rank methods are quite similar, but the Gehan estimator appears to outperform both. Figure S1 in the Supplementary Material compares the computing times of the three methods.

4.2 Scenario 2: Multivariate data

We further evaluate the proposed methods for clustered PIC and DC settings, where we generate data under the same AFT model (as in Scenario 1) adjusted for the clustered design, specified as $\log T_{ik} = 2 + \beta_1 X_{1ik} + \beta_2 X_{2ik} + \nu_i \varepsilon_{ik}$, $i = 1, \dots, n; k = 1, \dots, m_i$. Here, $X_{1ik} \sim N(0, 1)$, $X_{2ik} \sim \text{Bernoulli}(0.5)$, $(\beta_1, \beta_2) = (1, 1)$, and ε_{ik} follows standard normal distribution. The random effect ν_i is generated from a gamma distribution with shape parameter $1/\theta$ and scale parameter θ , with $\theta = \{0.5, 1\}$. Considering $n = 150$, the cluster size is set to $m_i = (r/10) + 2$ (to initiate the ICS scenario), where r is the r th percentile of the distribution of ν_i satisfying $q_r \leq \nu_i < q_{r+10}$, for $r = 0, 10, \dots, 90$. This configuration yields cluster sizes varying from 2 to 11. The generation of the PIC and DC endpoints follow Section 4.1.

Table 4 summarizes the performance of the Gehan and log-rank estimators via Bias, ESE, ASE and 95% CP, adjusted for ICS ($\varphi_i = 1/m_i$), or ignoring it ($\varphi_i = 1$). The relative efficiencies (RE) of cluster-adjusted method (over unadjusted) are also provided. We observe that although both methods work well, the ICS adjusted estimators can achieve higher statistical efficiencies with much lower empirical standard errors, compared to the unadjusted estimators. This indicates that the inverse-size weighting

helps reduce standard errors, as well as correct potential ICS issues.

Table 4: Simulation results for multivariate partly interval-censored (PIC) and doubly-censored (DC) data. Table entries are the average bias (Bias), empirical standard error (ESE), asymptotic standard error (ASE), and coverage probability (CP) of the 95% Wald-type confidence intervals of the parameter estimates, obtained from the Gehan and log-rank methods for the unadjusted and adjusted analysis, for $\theta = (0.5, 1)$. Also reported are the relative efficiency (RE) of the parameter estimates between the cluster-adjusted and the unadjusted method.

Type	θ	Method	Par	Unadjusted				Adjusted				RE
				Bias	ESE	ASE	CP	Bias	ESE	ASE	CP	
PIC	0.5	Gehan	β_1	0.001	0.051	0.051	0.954	0.001	0.042	0.041	0.942	1.474
			β_2	-0.001	0.098	0.100	0.964	-0.002	0.083	0.081	0.949	1.393
		Log-rank	β_1	0.014	0.064	0.064	0.946	0.013	0.051	0.057	0.969	1.549
			β_2	0.004	0.125	0.126	0.959	0.003	0.104	0.112	0.963	1.445
	1	Gehan	β_1	0.000	0.044	0.046	0.967	0.000	0.031	0.033	0.973	2.015
			β_2	-0.005	0.090	0.089	0.950	-0.004	0.064	0.063	0.957	1.976
		Log-rank	β_1	0.010	0.064	0.061	0.948	0.009	0.043	0.049	0.983	2.174
			β_2	-0.003	0.126	0.120	0.944	0.001	0.085	0.094	0.968	2.198
DC	0.5	Gehan	β_1	-0.004	0.058	0.059	0.948	-0.003	0.048	0.050	0.955	1.461
			β_2	-0.018	0.117	0.114	0.940	-0.013	0.093	0.096	0.957	1.589
		Log-rank	β_1	0.037	0.069	0.069	0.920	0.021	0.055	0.064	0.965	1.769
			β_2	-0.014	0.140	0.135	0.943	-0.011	0.105	0.123	0.976	1.776
	1	Gehan	β_1	-0.005	0.056	0.055	0.951	-0.003	0.039	0.042	0.968	2.066
			β_2	-0.014	0.103	0.105	0.953	-0.010	0.071	0.078	0.972	2.102
		Log-rank	β_1	0.039	0.069	0.068	0.919	0.018	0.044	0.056	0.986	2.780
			β_2	-0.011	0.130	0.131	0.959	-0.009	0.083	0.106	0.987	2.442

5. Application: Metastatic Colorectal Cancer Data

We applied the proposed method to a dataset from a multi-center, randomized, phase III clinical trial on metastatic colorectal cancer (Peeters et al., 2010). This trial aimed to evaluate the efficacy and safety of panitumumab combined with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared to FOLFIRI alone, following failure of initial treatment for metastatic colorectal cancer. The covariates considered in our analysis include treatment arm (0 = FOLFIRI alone, 1 = Panitumumab + FOLFIRI), patient

tumor KRAS mutation status (0 = wild-type, 1 = mutant), scaled patient age at screening, and gender (0 = male, 1 = female). The primary endpoint is progression-free survival (PFS).

Table 5: Analysis of colorectal cancer data: Table entries are the estimates, standard errors (SE) and p -values of the model parameters obtained from fitting both adjusted (for clustering) and unadjusted AFT models, using the Gehan, log-rank methods and the Buckley-James method (Gao et al., 2017). The estimates corresponding to the unadjusted proportional hazards (PH) and proportional odds (PO) models are also included.

Method	Treatment			Mutation status			Age			Gender		
	Est	SE	p -value	Est	SE	p -value	Est	SE	p -value	Est	SE	p -value
Adjusted AFT Model												
Gehan	0.336	0.102	0.000	-0.011	0.092	0.452	0.049	0.058	0.196	-0.262	0.111	0.009
Log-rank	0.256	0.108	0.009	0.032	0.120	0.395	0.070	0.054	0.098	-0.324	0.104	0.001
Unadjusted AFT Model												
Gehan	0.231	0.082	0.002	-0.138	0.070	0.025	0.057	0.037	0.061	-0.029	0.079	0.358
Log-rank	0.232	0.076	0.001	-0.184	0.083	0.013	0.096	0.046	0.019	-0.009	0.080	0.454
BJ	0.295	0.090	0.001	-0.190	0.097	0.025	0.065	0.047	0.086	-0.027	0.094	0.388
PH (Unadj.)	-0.206	0.088	0.020	0.173	0.084	0.040	-0.086	0.046	0.058	0.006	0.062	0.950
PO (Unadj.)	-0.351	0.117	0.003	0.233	0.117	0.047	-0.101	0.067	0.136	0.031	0.130	0.810

Given that tumor response was evaluated every eight weeks during clinic visits, the outcome of interest may be subject to interval-censoring. Setting the baseline assessment at day 0, a patient exhibiting disease progression at the first post-baseline assessment is left-censored, while those showing progression at later assessments are interval-censored. Patients alive without disease progression at the last assessment are right-censored, and exact PFS is observed for patients who died on-study. Among the 855 patients in the dataset, 52 (6.1%) died on-study, 168 (19.6%) were left-censored, 329 (38.5%) were interval-censored, and 306 (35.8%) were right-censored. Additionally, these patients were distributed across 148 clinic centers, with

the number of patients per center ranging from 1 to 23. Previous analyses (Pan et al., 2020) did not account for the potential influence of clinic centers when drawing inferences. Given that the PFS of patients enrolled within the same clinic may be correlated, we propose to account for the effect of the clinic center in our marginal analysis of this data.

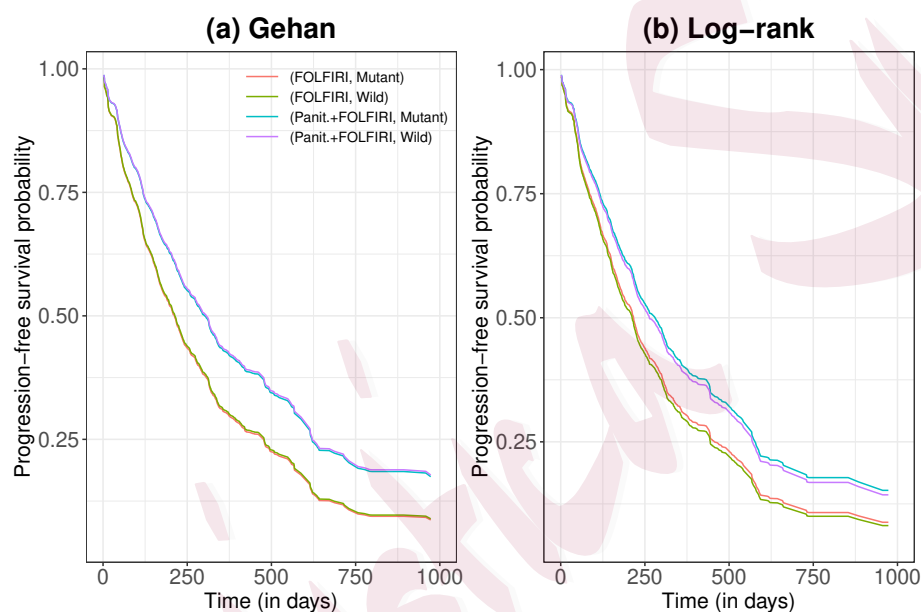


Figure 2: Nonparametric progression-free survival curves corresponding to the 4 groups (combinations of two treatment arms with KRAS mutation status), obtained from the Gehan (Panel a) and log-rank (Panel b) estimators.

Table 5 presents the mean estimates, standard errors, and associated p -values of parameters obtained from fitting both unadjusted and adjusted multivariate AFT models (3.9) using the Gehan and log-rank methods. For comparison, the regression results from the Buckley-James method for the

AFT model (Gao et al., 2017), proportional hazards (PH) and the proportional odds (PO) models (Anderson-Bergman, 2017) are also included. In all these analyses, the treatment effect is statistically significant, indicating that the panitumumab-combined therapy is more effective in prolonging patients' PFS period. For both the Gehan and log-rank methods, the estimated treatment effect in the adjusted AFT model is higher than in the unadjusted model. On the other hand, KRAS mutation status, which is significant in the unadjusted model, loses its significance in the adjusted model. This phenomenon is also confirmed in Figure 2, which shows that the PFS can increase by approximately 11% at 1000 days after initial treatment with panitumumab, with no apparent difference between the two mutation types. Here, we use the self-consistent approach (Gao et al., 2017; Choi et al., 2021) for plotting the nonparametric curves with two cluster-size-adjusted rank estimators. The covariate Age remains mostly non-significant, except from the log-rank test in the unadjusted model. On the other hand, Gender, which remains non-significant in the unadjusted model, becomes significant in the adjusted model, implying that the PFS of females is significantly shorter than that of males.

6. Conclusions

In this article, we present a unified rank-based estimation procedure for data with PIC and DC endpoints, extending it to multivariate cases. Most existing work in this area relies on Cox-type models, which often require the simultaneous and complex estimation of both regression and nuisance function parameters. In contrast, our proposed rank-based method directly estimates the regression coefficients without needing to address the residual distribution function, thereby significantly reducing computational complexity. Our simulation studies demonstrate that the proposed rank estimator nearly matches the statistical efficiency of the nonparametric maximum likelihood estimator.

As pointed out by a reviewer, the two rank estimators (Gehan and log-rank) are asymptotically similar, and, in general, we do not claim one to be better than the other. The Gehan estimator is easy to compute, while the log-rank estimator has a very similar form to the Cox estimator. If one wants to give variable weights to early events, the Gehan estimator should be used. On the other hand, if one believes that all data points contribute equally, the log-rank estimator is expected to be better.

An associate editor confirms that the proposed Gehan and log-rank estimators are suboptimal with respect to statistical efficiency, although

they are computationally efficient. To obtain full efficiency, we should utilize the nonparametric function estimator F . The optimal weight function should now have the form $\phi^{\text{opt}}(t) = \lambda'(t)/\lambda(t)$, with the optimal estimating equation (Lin and Chen, 2013) given as

$$S_{\phi}^*(\beta) = n^{-1} \sum_{i=1}^n \phi^{\text{opt}}(v_i(\beta)) \eta_{2i} \left\{ X_i - \frac{\sum_{j=1}^n \eta_{1j} X_j I\{v_i(\beta) \leq u_j(\beta)\}}{\sum_{j=1}^n \eta_{1j} I\{v_i(\beta) \leq u_j(\beta)\}} \right\}.$$

However, this approach would require completely different modeling strategies, and will be pursued in a separate work.

PIC and DC data typically contain a substantial amount of exact failure time observations. In the absence of exact observations, these data types reduce to case-2 and case-1 (or current status) interval-censored data, respectively, and can further generalize to panel count data if some intervals contain more than one count (Sun and Zhao, 2013). Rank regression analysis for such data is more complex due to the need for intricate theoretical considerations, which are currently being explored by the authors. Additionally, our rank-based estimation procedure does not identify the intercept term in the linear model, necessitating an extra ad-hoc approach to complete the analysis. Consistent and efficient estimation of the intercept term is crucial for survival prediction using the proposed linear model, highlighting a valuable area for future research.

Supplementary Materials

Proofs of Theorems 1 and 2, details on estimation under the DC setup, and additional simulation studies are relegated to the Supplementary Materials.

Acknowledgements

The colorectal cancer data was derived based on raw data sets obtained from <https://www.projectdatasphere.org/>, which is maintained by Project Data Sphere, LLC. The authors acknowledge support from grant (RS-2024-00340298) from the National Research Foundation of S. Korea (T. Choi), grants (2022R1A2C1008514, 2022M3J6A1063595) from the National Research Foundation of S. Korea and grant (K2305261) from Korea University (S. Choi), and grants (P20CA252717, P20CA264067, and R21DE031879) from the United States National Institutes of Health (D. Bandyopadhyay).

References

- Anderson-Bergman, C. (2017). icenReg: Regression models for interval censored data in R. *Journal of Statistical Software* 81(12), 1–23.
- Bogaerts, K., A. Komárek, and E. Lesaffre (2017). *Survival Analysis with Interval Censored Data: A Practical Approach with Examples in R, SAS, and BUGS*. London: Chapman & Hall/CRC.

REFERENCES

-
- Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika* 66(3), 429–436.
- Cai, T. and S. Cheng (2004). Semiparametric regression analysis for doubly censored data. *Biometrika* 91(2), 277–290.
- Chen, K., Z. Jin, and Z. Ying (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* 89(3), 659–668.
- Choi, S. and X. Huang (2021). Efficient inferences for linear transformation models with doubly censored data. *Communications in Statistics-Theory and Methods* 50(9), 2188–2200.
- Choi, T. and S. Choi (2021). A fast algorithm for the accelerated failure time model with high-dimensional time-to-event data. *Journal of Statistical Computation and Simulation* 91(16), 3385–3403.
- Choi, T., A. K. Kim, and S. Choi (2021). Semiparametric least-squares regression with doubly-censored data. *Computational Statistics & Data Analysis*. 164, 107306.
- Fygenson, M. and Y. Ritov (1994). Monotone estimating equations for censored data. *The Annals of Statistics* 22(2), 732–746.
- Gao, F., D. Zeng, and D. Y. Lin (2017). Semiparametric estimation of the accelerated failure time model with partly interval-censored data. *Biometrics* 73(4), 1161–1168.
- Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika* 52(3/4), 650–653.

REFERENCES

-
- Groeneboom, P. and K. Hendrickx (2018). Current status linear regression. *The Annals of Statistics* 46(4), 1415–1444.
- Jin, Z., D. Y. Lin, L. J. Wei, and Z. Ying (2003). Rank-based inference for the accelerated failure time model. *Biometrika* 90(2), 341–353.
- Jin, Z., D. Y. Lin, and Z. Ying (2006a). On least-squares regression with censored data. *Biometrika* 93(1), 147–161.
- Jin, Z., D. Y. Lin, and Z. Ying (2006b). Rank regression analysis of multivariate failure time data based on marginal linear models. *Scandinavian Journal of Statistics* 33(1), 1–23.
- Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65(2), 489–502.
- Kim, Y.-J. (2010). Regression analysis of clustered interval-censored data with informative cluster size. *Statistics in Medicine* 29(28), 2956–2962.
- Koenker, R. (2008). Censored quantile regression redux. *Journal of Statistical Software* 27(6), 1–25.
- Lai, T. L. and Z. Ying (1991). Large sample theory of a modified buckley-james estimator for regression analysis with censored data. *The Annals of Statistics* 19(3), 1370–1402.
- Lam, K. F., C. Y. Lee, K. Y. Wong, and D. Bandyopadhyay (2021). Marginal analysis of current status data with informative cluster size using a class of semiparametric transformation cure models. *Statistics in Medicine* 40(10), 2400–2412.

REFERENCES

-
- Li, L. and Z. Pu (2003). Rank estimation of log-linear regression with interval-censored data. *Lifetime Data Analysis* 9, 57–70.
- Lin, Y. and K. Chen (2013). Efficient estimation of the censored linear regression model. *Biometrika* 100(2), 525–530.
- Pan, C., B. Cai, and L. Wang (2020). A Bayesian approach for analyzing partly interval-censored data under the proportional hazards model. *Statistical Methods in Medical Research* 29(11), 3192–3204.
- Peeters, M., T. Price, A. Cervantes, et al. (2010). Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. *Journal of Clinical Oncology* 28(31), 4706–4713.
- Sun, J. and X. Zhao (2013). *Statistical Analysis of Panel Count Data*. New York: Springer.
- Tian, L. and T. Cai (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* 93, 329–342.
- Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* 69(345), 169–173.
- van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.
- Wang, S., B. Nan, J. Zhu, and D. G. Beer (2008). Doubly penalized Buckley–James method for survival data with high-dimensional covariates. *Biometrics* 64(1), 132–140.

REFERENCES

Wellner, J. A. and Y. Zhan (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association* 92(439), 945–959.

Xu, Y., D. Zeng, and D. Lin (2023). Marginal proportional hazards models for multivariate interval-censored data. *Biometrika* 110(3), 815–830.

Zeng, D. and D. Y. Lin (2008). Efficient resampling methods for nonsmooth estimating functions. *Biostatistics* 9(2), 355–363.

Zhang, J. and D. F. Heitjan (2006). A simple local sensitivity analysis tool for nonignorable coarsening: application to dependent censoring. *Biometrics* 62(4), 1260–1268.

Zhao, X., Q. Zhao, J. Sun, and J. S. Kim (2008). Generalized log-rank tests for partly interval-censored failure time data. *Biometrical Journal* 50(3), 375–385.

School of Mathematics, Statistics & Data Science, Sungshin Women’s University, Seoul, South Korea

E-mail: tchoi@sungshin.ac.kr

Department of Statistics, Korea University, Seoul, South Korea

E-mail: choisang@korea.ac.kr

Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia, U.S.A.

E-mail: dbandyop@vcu.edu