MATEMATIKA, MJIAM, Volume 42, Number 2, 249–265

Cox-Integrated Spline Sieve Estimation for Doubly Censored Data

[1,2]Muhammad Mustapha, [1]Zarina Mohd Khalid* and [3]Adina Najwa Kamarudin

[1]Department of Mathematical Science, Faculty of Sciences, Universiti Teknologi Malaysia, 81310, UTM Johor Bahru, Johor, Malaysia.

[2]Department of Statistics, Faculty of Physical Sciences, Science Complex, University of Maiduguri, Maiduguri, Borno State, Nigeria.

[3]Institute of Mathematical Sciences, Universiti Malaya, 50603, Kuala Lumpur, Malaysia.

*Corresponding author: zarinamkhalid@utm.my

Article history
Received: 20 April 2025
Received in revised form: 3 September 2025
Accepted: 18 September 2025
Published online: 20 July 2026

Abstract When considering doubly censored data, the failure time of interest refers to the duration between the initial and a subsequent event, where both events may be censored. Prior studies investigating the association between these events have yielded statistically inconclusive results due to their reliance exclusively on censored data (initial event). To address this limitation, a spline-based sieve maximum likelihood estimator (MLE) was developed by using integrated spline (I-splines) and M-splines (derivative of I-splines), integrated with Cox Proportional Hazards, to investigate the association between HIV infection and AIDS incubation period. Unlike prior studies, the proposed approach incorporates exactly observed event times and censored data to improve dependency estimates. The estimator treats the initial event as a covariate, allowing for a direct investigation of its relationship with the failure time. Spline functions simultaneously estimate parametric/non-parametric components, simplifying model implementation; multiple imputation ensures robustness to missing data. The asymptotic properties of the estimator are rigorously established. In simulation studies with sample sizes $n = 50, 100, 200, 300$, the proposed estimator achieved a reduction up to 93% in mean squared error and a 95% reduction in bias compared with existing methods, with coverage probabilities close to nominal levels. In the AIDS dataset, the proposed method revealed a statistically significant association between HIV infection timing and AIDS incubation ($\log \mathrm{HR} = 0.580$) that earlier analyses failed to detect ($\log \mathrm{HR} = -0.11$), and produced more precise treatment effect estimates. These findings highlight the enhanced sensitivity and statistical power of the proposed Cox-integrated spline sieve approach.

Keywords Doubly censored data; Cox-Integrated spline estimation; I-splines; M-splines; Semiparametric Sieve Estimation.

Mathematics Subject Classification 62N01, 62G07, 65D07.

---

1 Introduction

Doubly censored data commonly arise from survival analysis when the exact event times are not observed; instead, they are known to fall within a certain interval. Doubly censored data refer to the gap or interval between the initial and the subsequent occurrences, where both occurrences may be subject to censoring [1]. In the literature, other types of double censoring refer to events that can be precisely observed only within a certain interval; outside this interval, the events are left- or right-censored [2–6]. Similarly, recurrent event data subject to doubly censoring within restricted observational windows have been investigated by [7]. Although these frameworks address fundamentally different censoring structures than the gap-time model in [8], they highlight the growing methodological interest in efficient estimation under diverse double-censoring scenarios. Doubly censored data often occur in clinical studies when time-to-event outcomes or survival times for patients or subjects undergoing treatment, such as for a new cancer therapy, are recorded. Some patients may have already begun treatment before enrolling in the study, making it impossible to determine the exact start date of their treatment. In these cases, only the enrolment date is documented, resulting in left-censoring because the initial event (start of treatment) occurred before the study observation began. Additionally, some patients who participated in the study may still be alive or show no evidence of disease progression at the study end, leading to right censoring since the event (progression or death) has not yet happened during the observation period. In contrast, when the timing of disease progression is not exactly observed but can be narrowed to a specific interval, interval censoring occurs. For instance, a patient may enter the study with no indications of cancer progression (left-censored) and later display signs of progression during a later clinical assessment, although the precise period of this event remains unrecorded. Instead, the only available information indicates the event occurred within a known interval. Furthermore, in some cases, survival times or other events may be exactly observed, offering precise information to complement the censored data.

Efforts to refine analytical techniques for doubly censored data have been particularly prominent in the studies of the time between Human Immunodeficiency Virus (HIV) infection and development of Acquired Immune Deficiency Syndrome (AIDS). The foundation for this work was laid by [9], who proposed a nonparametric likelihood approximation for a univariate, singly censored proportional hazards model. This approach was later extended by [10], who established an expectation maximization (EM) algorithm using a proportional hazards model, albeit with a requirement for data discretization for cases involving interval-censored initial time and a right-censored occurrence time, [11] applied Cox regression and presented an EM technique to replace the initial event and fit the model on right-censored data. Building on these ideas, [12] formulated an estimation equation approach built from the marginal likelihood concept for the proportional hazards model. This framework was further extended by [13], whose estimating equation idea was shown to perform comparably to multiple imputations in terms of large sample properties. In a more general scenario, where both initial and subsequent events are subject to interval censoring, [14] considered a non-parametric estimation procedure using a two-step technique. Finally, [15] considered the marginal likelihood and a Cox model-based estimating equation strategy for double censored data and examined the critical assumption of independence.

The methodological landscape includes the work of [16], who applied a maximum likelihood (ML) approach to the proportional hazards regression, employing the Gauss-Seidel method for constrained optimization to generate the ML estimator. Similarly, [15] introduced a method based on estimating equations to test the independence between the initial and subsequent occurrences, an approach that assumed the proportional hazards model was suitable for describing the connection linking HIV infection to the latency phase preceding the AIDS incubation period. Recently, [17] developed a stepwise technique combining multiple imputation (MI) and EM algorithms. Utilizing a Poisson variable within semiparametric transformation frameworks to evaluate the impact of two treatments on incubation periods. A common thread in the above literature is the assessment of the timing characteristics of AIDS incubation periods under the assumption of a lack of association between the initial event and the subsequent event. Notably, Prior analyses have not incorporated potential dependencies between the target failure time and the initial event, leaving a gap in understanding the relationship between HIV infection timing and AIDS progression.

Current literature includes a few instances where dependency is rigorously incorporated, notably in the methodologies developed by [18, 19]. Specifically, [18] employed an approximate likelihood technique developed by [20] to investigate the relationship linking the event of interest and initial event time using a frailty model. More recently, [19] examined the dependency assumption and provided theoretical evidence via a flexible likelihood framework incorporating a spline to evaluate the association between the two events under single-interval censorship. Additionally, they investigated the association between HIV infection and AIDS incubation periods, considering that both events are interval censored. However, their results yielded inconclusive evidence regarding the association between HIV infection and the AIDS incubation period, which highlights potential limitations in statistical power or model flexibility due to their reliance on censored data alone. A recent study by [21] extended Kendall's τ to quantify associations between consecutive gap times in doubly interval-censored data, addressing the bias introduced by induced dependent censoring through inverse probability censoring weighting (IPCW) and multiple imputation for missing failure times. However, existing methods for doubly censored data often rely exclusively on censored observations, assume independence between the initial and subsequent events, leading to inconclusive or biased association estimates. While recent advances in spline-based and sieve estimation methods [22–25] are notable, they have not fully addressed scenarios combining exact, interval, and right-censored data with potential event dependencies. Thus, a significant shortfall remains, as existing methods fail to provide precise association estimates when both exact and censored event times are present. This gap motivates the development of the proposed spline-sieve approach, which aims to incorporate exactly observed event times and enforce natural constraints via I-splines and M-splines to achieve more precise and unbiased estimates of the association.

To address these limitations, a semiparametric Cox-integrated spline sieve estimator has been developed. Unlike prior approaches, the proposed methods incorporate exactly observed event times alongside interval and right-censored data, enhancing precision by reducing uncertainty in parameter estimation. The proposed approach used monotonic I-spline functions to describe the cumulative hazard behaviour and M-spline to model the hazard function to ensure non-negativity and adherence to survival model constraints. Unlike the conventional spline approaches, which may not preserve survival analysis constraints, the inclusion of the exactly observed data would further stabilize estimation by providing anchor points for the splines and reducing variability in parameter inference. This novel framework allows a more conclusive test of the association between HIV infection and AIDS incubation.

In this study, both events are to be exactly observed, interval censored, or right censored as considered by [17], recognizing that exact observation provides additional details relative to the interval-censored data approach. To perform inference, we adopt sieve-based techniques due to their demonstrated efficacy in estimating semiparametric survival models under the smoothness assumption.

According to [26], the commonly used M-spline and I-spline functions are connected to the B-spline function, with M-spline being the derivative of the I-spline function. A marginal likelihood technique was formulated together with both M-spline and I-spline for the implementation; the multiple imputation techniques would be employed to make regularly censored data for easy implementation of the proposed methods. The contributions of this work lie in combining these techniques into a unified, semiparametric Cox-integrated spline sieve estimator that accommodates mixed censoring types, incorporates exactly observed data, and enables a more accurate and interpretable analysis of AIDS incubation periods. Establishing a significant association between the timing of HIV infection and AIDS progression would have enhanced and highlighted the need for further investigation into the biological, clinical, and public health implications.

The organization of the research work is constructed in the following manner: Section 2 presents the notation, outlines the model, along with the semiparametric sieve MLE. Section 3 provides the asymptotic properties and asymptotic normality. Section 4 shows the multiple imputation approach to impute the uncensored observation. Section 5 conducts a simulation study to show the performance of the proposed technique. Section 6 shows the application of the proposed model to the doubly-censored observation related to the AIDS incubation period. Section 7 concludes the discussion of the results.

2 Notations, Model and Sieve MLE

Consider doubly censored data with an initial event S_i and a subsequent event T_i for each individual i, and the failure time of interest $T_i^* = T_i - S_i$, in which both T_i and S_i might be censored. In an attempt to investigate the association between the q-dimensional covariates and the failure time of interest, the cumulative hazard function is as follows:

$$\Lambda(t \mid Z, S) = \Lambda(t) \exp(\beta^T Z + \alpha S) \tag{1}$$

Here, $\Lambda(t \mid Z, S)$ represents the cumulative hazard function of T_i^* conditioned on the covariates Z_i and S_i. $\Lambda(t)$ refers to the underlying cumulative risk function, where β and α serve as the associated regression parameters. In this model, the initial occurrence S_i is incorporated as a covariate to investigate the association involving the initial event S_i and the failure time of interest T_i^*, which was also considered by [19] with inconclusive results. Given that both S_i and T_i are subject to censoring, both events are assumed to be observed within the intervals (S_{L_i}, S_{R_i}) and (T_{L_i}, T_{R_i}). It is apparent that $S_{L_i} = S_{R_i}$ can be defined if S_i can be observed exactly, and $T_{L_i} = T_{R_i}$ indicates that T_i is also observable. Additionally, for right-censored observations, set $T_{R_i} = \infty$. To prevent the occurrences of both S_i and T_i from being censored at the same time, we also assume that $S_{R_i} < T_{L_i}$. As justified by [19] in circumstances where both events are censored within their respective intervals, the data will not reveal any differences in the failure time of interest; therefore, such observations need to be deleted. For determining the unobserved variables in the model (1), we consider a situation where the initial occurrence S_i is known. Then, the failure time of interest T_i^* could be determined by setting the interval as $(T_{L_i} - S_i, T_{R_i} - S_i)$ and $T_i^* = T_i - S_i$ for a subject whose subsequent event is exactly observed or interval censored data.

Under this setup, the likelihood function is as follows.

$$\begin{aligned} l_n(\theta, \Lambda \mid Z, S) = &\sum_{i=1}^{n} \delta_{1i} \log \Big[\lambda(T_i - S_i) \exp(\beta^T Z_i + \alpha S_i) \\ &\quad \times \exp\Big(- \Lambda(T_i - S_i) \exp(\beta^T Z_i + \alpha S_i)\Big)\Big] \\ &+ \sum_{i=1}^{n} \delta_{2i} \log \Big[\exp\Big(- \Lambda(T_{L_i} - S_i) \exp(\beta^T Z_i + \alpha S_i)\Big) \\ &\quad - \exp\Big(- \Lambda(T_{R_i} - S_i) \exp(\beta^T Z_i + \alpha S_i)\Big)\Big] \\ &+ \sum_{i=1}^{n} \delta_{3i} \log \Big[\exp\Big(- \Lambda(T_{R_i} - S_i) \exp(\beta^T Z_i + \alpha S_i)\Big)\Big]. \end{aligned} \tag{2}$$

Here, the indicator functions are defined as follows.

$$\delta_{1i} = I(T_{L_i} = T_{R_i}), \quad \delta_{2i} = I(T_{L_i} < T_{R_i} < \infty), \quad \delta_{3i} = I(T_{R_i} = \infty),$$

and the cumulative hazard function is given by:

$$\Lambda(t) = \int_0^t \lambda(u)\, du. \tag{3}$$

In the concept of doubly censored data, if S_i is not directly known, we assume that the true distribution function F_0 of S_i within the interval denoted (S_{L_i}, S_{R_i}) is known. To estimate the unknown parameter θ, the approach proposed by [12] was employed and the following equation will be used.

$$l_n(\theta, \Lambda \mid Z, S) = \int_{S_{L_i}}^{S_{R_i}} \cdots \int_{S_{L_i}}^{S_{R_i}} l_n(\theta, \Lambda \mid Z, s) \prod_{i=1}^{n} dF_0(s_i), \tag{4}$$

where the regression parameters are defined as $\theta = (\beta^T, \alpha)^T$ and $\theta \in \Theta$ defines the parameter space. It is quite challenging to maximize the likelihood function considering the high infinite dimension of the cumulative baseline hazard $\Lambda(t)$ and its instantaneous hazard counterpart $\lambda(t)$ with the intricate nature of the likelihood function. To resolve this difficulty, the baseline hazard function $\lambda(t)$ is proposed to be modeled with the use of the M-spline function, which is also used for a positive function. Subsequently, the cumulative hazard function $\Lambda(t)$ is proposed to use the I-spline function, which is an integration of the M-spline function that fits the relationship between the $\Lambda(t)$ and the $\lambda(t)$, as has been used by [27–29].

Sieve estimation approximates complex, infinite-dimensional functions by a sequence of increasingly flexible models, such as spline bases. As the sample size increases, the number of basis functions expands, enabling the estimator to balance flexibility and stability. This approach allows for consistent and efficient estimation of hazard and cumulative hazard functions under complex censoring. In this study, monotone I-splines were used to estimate cumulative hazards, and their derivatives, M-splines, were employed to model baseline hazard functions, ensuring non-negativity and adherence to the constraints of the survival model. The sieve estimation procedure, originally proposed by [30] reduces the dimensionality of the maximization problem, making likelihood estimation more computationally feasible. To implement the sieve-based likelihood estimation approach, we utilize the B-spline basis to construct both the M-spline as well as the I-spline function to estimate $\Lambda(t)$ and $\lambda(t)$ The B-spline sieve approach for semiparametric models has been employed by [19] for doubly censored data and by [31] for interval-censored data. For approximating the non-negative functional form $\lambda_n(\cdot)$ via B-spline function over the interval $[a, b]$, the interval is partitioned into $K_n + 1$ identical subintervals with knots $d_0 < d_1 < \cdots < d_{K_n+1}$, where the positive integer $K_n = O(n^v)$. The spline function can be defined as

$$\lambda_n(t) = \sum_{j=1}^{q_n} \alpha_j \beta_j(t), \quad \alpha \in A_{q_n}, \tag{5}$$

where $A_{q_n} = (\alpha_1, \alpha_2, \ldots, \alpha_{q_n})$ subject to the constraints $\alpha_{ij} \geq 0, \theta_n \leq 1$. Thus, we can define $\Phi_n = [\lambda_n(t)]$.

The log-likelihood function when λ_n is substituted with λ is

$$\begin{aligned} l_n(\theta, \Lambda|Z, S) = & \sum_{i=1}^{n} \delta_{1i} \log \Bigg[\sum_{j=1}^{q_n} \alpha_j \beta_j(T_i - S_i) \exp(\beta^T Z_i + \alpha S_i) \\ & \times \exp\Bigg(- \int_{s_i}^{t_i} \sum_{j=1}^{q_n} \alpha_j \beta_j(T_i - S_i) \exp(\beta^T Z_i + \alpha S_i) \Bigg) \Bigg] \\ & + \sum_{i=1}^{n} \delta_{2i} \log \Bigg[\exp\Bigg(- \int_{s_i}^{t_{li}} \sum_{j=1}^{q_n} \alpha_j \beta_j(T_{L_i} - S_i) \exp(\beta^T Z_i + \alpha S_i) \Bigg) \\ & - \exp\Bigg(- \int_{s_i}^{t_{ri}} \sum_{j=1}^{q_n} \alpha_j \beta_j(T_{R_i} - S_i) \exp(\beta^T Z_i + \alpha S_i) \Bigg) \Bigg] \\ & + \sum_{i=1}^{n} \delta_{3i} \log \Bigg[\exp\Bigg(- \int_{s_i}^{t_{ri}} \sum_{j=1}^{q_n} \alpha_j \beta_j(T_{R_i} - S_i) \exp(\beta^T Z_i + \alpha S_i) \Bigg) \Bigg]. \end{aligned} \tag{6}$$

The sieve-based MLE framework is formulated through optimization of the log-likelihood (6). Derived from this likelihood, integrating B-spline basis functions introduces computational challenges due to their inherent complexity. The monotone I-spline approach introduced by [26] as a technique for sieve estimation provides an alternative to B-spline sieve estimation. To approximate the baseline hazard function, the monotone I-spline was chosen to eliminate the need for integration. Notably, [29, 32] applied this procedure to estimate the underlying hazard and its cumulative counterpart within time-to-event data.

Following the definition of [26], the I-spline and M-spline functions are denoted as I_j^l and M_j^l, where the M-spline is defined as the derivative of the I-spline, i.e., $M_j^l(t) = \dfrac{dI_j^l(t)}{dt}$. The B-splines and I-splines developed by [26] are closely related, with the I-spline component derived from the cumulative aggregation of the associated B-spline functions. Furthermore, [28] demostrated that the B-spline function of degree l, denoted as $I_j^l(t)$, can be expressed as

$$I_j^l(t) = \sum_{k=j+1}^{q_n+1} B_k^{(l+1)}. \tag{7}$$

Additionally, the derivative of the B-spline function is represented as $M_j^l(t)$, which can be equally expressed as

$$M_j^l(t) = \frac{1}{d_{j+l} - d_j} B_j^l(t), \tag{8}$$

where d_{j+l}, d_j refers to specific knots from the sequences associated with the B-spline basis functions. Notably, I_j^l has a polynomial degree of l, while M_j^l, being a derivative, has a polynomial degree of $l-1$.

As shown by [29], it can be demonstrate that the set

$$\Theta_{(N,n)} = \left\{ \int \lambda_n : \lambda_n \in \Phi_n^I \right\} \tag{9}$$

is equivalent to the spline function space, which is defined as

$$\Phi_{(I,n)} = \left\{ \Lambda_n = \sum_{j=1}^{q_n} \eta_j I_j^l : \sum_{j=1}^{q_n} \eta_j \leq b_0, a_0 \leq \frac{l(\eta_j)}{d_{j+1} - d_j} \leq b_0, j = 1, \ldots, q_n \right\}. \tag{10}$$

This means that the estimation problem based on B-splines can be transformed into an equivalent estimation problem using I-splines. Furthermore, it can be simplified to

$$\Theta_{(I,n)} = \left\{ \Lambda_n = \sum_{j=1}^{q_n} \eta_j I_j^l : \sum_{j=1}^{q_n} \eta_j \leq b_0, j = 1, \ldots, q_n \right\}. \tag{11}$$

Based on this discussion, the likelihood function can be rewritten as follows.

$$\begin{aligned} l_n(\theta, \lambda_n, Z, S) = \prod_{i=1}^{n} & \left[\sum_{j=1}^{q_n} \eta_j M_j(T_i - S_i) \exp(\beta^T Z + \alpha S) \exp\left(-\sum_{j=1}^{q_n} \eta_j I_j(T_i - S_i) \exp(\beta^T Z + \alpha S) \right) \right]^{\delta_{1i}} \\ & \times \left[\exp\left(-\sum_{j=1}^{q_n} \eta_j I_j(T_{Li} - S_i) \exp(\beta^T Z + \alpha S) \right) \right. \\ & \quad \left. - \exp\left(-\sum_{j=1}^{q_n} \eta_j I_j(T_{Ri} - S_i) \exp(\beta^T Z + \alpha S) \right) \right]^{\delta_{2i}} \\ & \times \left[\exp\left(-\sum_{j=1}^{q_n} \eta_j I_j(T_{Ri} - S_i) \exp(\beta^T Z + \alpha S) \right) \right]^{\delta_{3i}}. \end{aligned} \tag{12}$$

3 Asymptotic properties

The asymptotic properties were established for the proposed estimators of the parameters under the following regularity conditions.

(C1) The covariates vector $E(Z,S)$ and $E(Z,S)^T$ are invertible matrices; a scalar z_0 and s_0 exist such that $P(\|Z\| < z_0) = 1$ and $P(|S| < s_0)$ implies both Z and S are bounded. Set $|\cdot|$ as the Euclidean norm. The actual regression coefficient β_0 is included in the interior of a set B, which is a compact subset in $\mathbb{R}^q$.

(C2) The actual cumulative hazard function $\Lambda_0(t)$ satisfy the baseline hazard rate $\lambda_0(t)$, such that

$$\Lambda_0(t) = \int_a^b \lambda_0(u)du, \tag{13}$$

while the function $\lambda_0(t)$ is positively bounded and its derivative is bounded on the interval $[a,b]$.

(C3) There exists $\eta > 0$ such that $P(T_L - S_R \geq \eta) = 1$. Subsequently, $P(T_R - T_L \geq \eta) = 0$ and $P(S_R - S_L \geq \eta) = 0$. Both endpoints S_R, S_L, and T_R, T_L are within the intervals $[a,b]$.

(C4) For a certain $\eta \in (0,1)$,

$$\mu^T \text{var}((Z^T,S)^T|L)\mu \geq \eta\mu^T E((Z^T,S)^T(Z^T,S)|L)\mu, \tag{14}$$

and

$$\mu^T \text{var}((Z^T,S)^T|R)\mu \geq \eta\mu^T E((Z^T,S)^T(Z^T,S)|R)\mu. \tag{15}$$

(C5) The joint distribution of $(S_R, S_L, T_R, T_L, Z, S)$ is independent of the parameters (θ, Λ). Additionally, S_R, S_L are independent of S, and similarly, T_R, T_L are independent of T, given covariates (Z,S).

Some of the conditions were mentioned by [19, 29]. They have been used in many studies, and are adjusted here to suit the problems under consideration. C2 implies that the underlying hazard rate $\lambda(t)$ and its first derivative have been bounded within an interval $[a,b]$. Therefore, the survival rate at time t is not zero. C3 also implies that both the initial and subsequent events would not occur in the same interval, such that the failure time of interest has information on both events; it also shows that both events would be exactly observed and non-ignorable. These results are summarized in the next two theorems.

Theorem 1 *Assuming $C1-C5$, it follows that*

$$d((\hat{\theta}_n, \hat{\lambda}_n), (\theta_0, \lambda_0)) = O_p(r^{-\min\{pv,(1-v)/2\}}).$$

The asymptotic normality of the developed estimator is demonstrated here, considering the spline constraints and restrictions on the parameters. This involves both parametric and smooth functions of the non-parametric parts. Assume that a parametric smooth sub-model with parameter $\theta = (\beta, \alpha, \lambda_{c,t})$, where $\lambda_{ct} = \lambda + ct$. Setting $\lambda_0 = \lambda$, then the derivative of $\lambda_{(c,t)}$ concerning c, where $c = 0$ is

$$\left(\frac{d\lambda_{(c,t)}}{dc}\right)_{|c=0} = t. \tag{16}$$

Subsequently, the derivative of $\phi(\lambda_{(s,h)})$ *for* c *where* $c = 0$ *is*

$$\left(\frac{d\phi(\lambda_{(c,t)}}{dc}\right)_{|c=0} = \phi_\lambda(\lambda)[t], \tag{17}$$

where the function of λ *is* ϕ_λ.

Let μ *denote the class of functions* t*; the limitations on the spline parameters* η_M, η_I *would be taken into consideration by the class* μ*. To verify that the values of these parameters remain within acceptable ranges, these constraints would be incorporated as a requirement to achieve the smoothness of the spline terms, specifically, the constraints* $\eta_1 \leq \eta_2 \leq \cdots \leq \eta_{qn}$*. For the score operator* λ*, the directional derivative of the log-likelihood with respect to* t *can be expressed as*

$$l_\lambda(\theta : x)[t] = \frac{d}{dc} l(\theta, \lambda_{(c,t)}; x)\bigg|_{c=0} \equiv f_t(\theta, \lambda, x). \tag{18}$$

If the distribution function F_s *contributes to the directional derivative of the log-likelihood since* $l(\theta, \lambda; x)$ *includes* $\log f_s(S_i)$*, hence the score operator*

$$f_t(\theta, \lambda, x) = d/dc[\log F_{(T/Z,S)}(T/Z, S; \lambda) + \log f_s(S)]. \tag{19}$$

The second-order directional derivatives with respect to t_1 *and* t_2 *is also obtained as*

$$l_{(\lambda,\lambda)}(\theta : x)[t_1][t_2] = d/dc f_{(t_1)}(\theta, \lambda_{(c,t_2)}; x)|_{c=0}. \tag{20}$$

Furthermore, setting $t = (t_1, \ldots, t_d)^T$ *with each* $t_c \in \mu$ *for* $c = 1, \ldots, d$*; let* $l_\lambda(\theta, \lambda)[t]$ *represent d-dimensional that has* $l_\lambda(\theta, \lambda)[t_c]$*. In this situation, the smoothness and ordering conditions* $\eta_1 \leq \eta_2 \leq \cdots \leq \eta_{q_n}$ *will be respected by the restrictions on the spline parameters. For* $t_1 = (t_{1,1}, \ldots, t_{1,d})^T$ *and* $t_2 = (t_{2,1}, \ldots, t_{2,d})^T$*, let* $l_{(\lambda,\lambda)}(\theta : x)[t_1][t_2]$ *denotes a* $d \times d$ *matrix with i-th row and j-th column entry* $l_{(\lambda,\lambda)}(\theta : x)[t_{(1,i)}][t_{(2,j)}]$*. For the d-dimensional vector* $\beta = (\beta_1, \ldots, \beta_d)^T$*, define* $l_\beta(\theta : x) = (l_{\beta_1}(\theta : x), \ldots, l_{\beta_d}(\theta : x))^T$*, where* $l_{\beta_c}(\theta : x)$ *correspond to the derivative of the log-likelihood* $l(\theta, x)$ *relative to* β_c*,* $c = 1, \ldots, d$*.*

Now, define the function $\phi_c(\theta, t)$ *as*

$$\phi_c(\theta, t) = \left\{l_{\beta c}(\theta, x) - l_\lambda(\theta, x)[t]\right\}^2 \quad \textit{for } c = 1, \ldots, d. \tag{21}$$

Let $t_c^* = \arg\min_{t\in\mu} \phi_c(\theta_0, t)$ *be the value of* t *that minimizes the function* ϕ_c *at the true parameter* θ_0*, subject to constraints* η_M, η_I*. Then, the coefficient score of* β_0 *is given by* $l_\beta(\theta : x) - l_\lambda(\theta : x)[t^*]$ *with* $t^* = (t_1^*, \ldots, t_d^*)$*. [33].*

Lastly define the function $\phi(\theta, h)$ *as*

$$\phi(\theta, h) = \{l_\beta(\theta : x) - l_\lambda(\theta : x)\}^{\otimes 2} \tag{22}$$

where $v^{\otimes 2} = vv^T$ *represents the outer product column vector* v*. Then the information matrix for* β_0 *is*

$$I(\beta_0) = P\phi(\theta_0, h^*). \tag{23}$$

Theorem 2 *Assuming $C1 - C5$, it follows that*

$$\sqrt{n}(\hat{\theta}_0 - \theta_n) \xrightarrow{d} N\left(0, I^{-1}(\theta_0, \lambda_0)\right).$$

To establish the asymptotic normality of $\hat{\theta}_n = (\hat{\beta}_n, \lambda_n)$, defining the parameter $\theta = (\beta, \alpha, \lambda)$, the score function which gives the first derivatives and is used to determine the maximum likelihood estimator of $\hat{\theta}_n$ is as follows

$$\omega_n(\theta) = \left(\frac{\partial l_n(\theta, \lambda \mid Z, S; x)}{\partial \theta}\right), \quad \omega_n(\lambda) = \left(\frac{\partial l_n(\theta, \lambda \mid Z, S; x)}{\partial \lambda}\right). \tag{24}$$

The Fisher information matrix is set as

$$I_n(\theta, \lambda) = \omega_n(\theta) \cdot \omega_n(\lambda). \tag{25}$$

By the Central Limit Theorem and regularity conditions, it can be shown that

$$\sqrt{n}(\hat{\theta}_0 - \theta_n) \xrightarrow{d} N\left(0, I^{-1}(\theta_0, \lambda_0)\right). \tag{26}$$

4 Multiple Imputation Procedure

Estimating the parameters of the maximum likelihood estimator by direct maximization is very difficult due to the integral in the likelihood function and the unknown exact values of S_i. However, S_i is known to lie within an interval (S_{L_i}, S_{R_i}). For simplicity and computational convenience, the multiple imputation procedure would be employed to estimate the parameters $(\hat{\theta}_n, \hat{\lambda}_n)$. This technique imputes S_i from the distribution function F_S within the interval (S_{L_i}, S_{R_i}), where for each individual i, S_i^m is drawn from

$$S_i^m \sim F(\cdot \mid S_{L_i} \leq S_i \leq S_{R_i}),$$

with F is the distribution function derived from the spline model. The exact values of S_i are imputed M times, where M is the pre-specified number of imputations. Each imputed dataset is used to estimate $(\hat{\theta}_n^{(m)}, \hat{\lambda}_n^{(m)})$. The imputed S_i^m adhere to the observed censoring intervals (S_{L_i}, S_{R_i}). This idea has been successfully applied by [13, 17, 19].

For the m-th imputation, the adjusted variables can be defined as

$$T_i^{(*m)} = T_i - S_i^m, \quad L_i^m = T_{L_i} - S_i^m, \quad R_i^m = T_{R_i} - S_i^m. \tag{27}$$

These transformed variables enable efficient computation within the likelihood framework. The final estimator combines results from all imputed datasets

$$\hat{\theta}_n = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_n^{(m)}, \quad \hat{\lambda}_n = \frac{1}{M}\sum_{m=1}^{M} \hat{\lambda}_n^{(m)}. \tag{28}$$

The variance estimate is computed as

$$\hat{\Sigma} = \frac{1}{M}\sum_{m=1}^{M} I_{(m)}^{-1} + \left(1 + \frac{1}{M}\right) \frac{\sum_{m=1}^{M} (\hat{\theta}_n^{(m)} - \hat{\theta}_n)^\top (\hat{\theta}_n^{(m)} - \hat{\theta}_n)}{M-1}, \tag{29}$$

where $I_{(m)}$ denotes the covariance estimate for $\hat{\theta}^{(m)}$. According to [13], a small M (e.g., $M = 5$) is often sufficient, but larger M improves robustness. It is recommended to incrementally increase M to assess result stability.

5 Simulation Research

A simulation experiment was performed to determine the effectiveness of the suggested approaches. For each subject i, the initial occurrence time S_i and the failure time of interest T_i^* were generated using an exponential distribution with means of 2 and $\dfrac{1}{\exp(\beta^T Z + \alpha S)}$. A two-dimensional covariate $Z_i = (Z_1, Z_2)$ was generated using a Bernoulli distribution (0.5) for Z_1 and a normal distribution (0, 1) for Z_2, with $\beta_i = (0.5, -0.5)^T$ and $\alpha = 1$. We then set $T_i = S_i + T_i^*$. To simulate a sequence of visitation instances for each subject i, the initial visitation instances were simulated independently from a uniform distribution over (0, 1). Subsequently, the gaps between consecutive visitation instances were simulated independently using a uniform distribution over (0, 0.5). The study length was set to 4. If the interval length is smaller than 0.3, we utilize the above-simulated S_i and T_i in place of the interval-censored observation for every individual i, thereby generating the exact observed S_i. In this simulation framework, approximately 20% of the initial events are exactly observed, while about 14% of subsequent events are interval censored and 49–50% are right censored. These choices were made to reflect realistic monitoring intervals and censoring patterns commonly seen in AIDS incubation studies. Additional sensitivity analyses with alternative censoring rates, visit intervals, and baseline hazards produced consistent results, confirming the robustness of the findings.

The sieve MLE F_s is computed using a cubic spline with the Generalized Gradient Projection (GGP) Algorithm (see Algorithm in the Appendix), as employed by [19, 28, 29, 31]. Subsequently, the estimate of θ_n and its associated standard error (SE) are calculated using the multiple imputation procedure. In the study, we used $M = 20$ imputations. for the cubic spline representation, the number of interior knots was chosen as $K_n = \lfloor n^{1/3} \rfloor$, where n is the sample size; for $n = 200$, this gives $K_n = 5$. Boundary knots were placed at the observed minimum and maximum event times. Spline functions were then used to estimate F_s and θ_n. A total of 500 simulation runs were carried out to evaluate the biases, observed and estimated standard errors, and coverage probabilities of the 95% confidence intervals (CI) for the regression coefficients. Table 1 demonstrates that the biases are very minimal and decrease notably with increasing sample size. Both the standard error (SE) and the standard deviation (SD) provide reliable estimates consistent with the employed techniques. Furthermore, the coverage probabilities, all of which are around 95%, indicate that a Gaussian approximation suitably characterizes the sampling variability of the proposed estimator.

Table 1: Results obtained through a spline-integrated sieve maximum likelihood approach

Reg. parameters	$\boldsymbol{\beta_1}$				$\boldsymbol{\beta_2}$				$\boldsymbol{\alpha}$			
Sample Size n	50	100	200	300	50	100	200	300	50	100	200	300
Bias	-0.004	-0.027	-0.044	-0.044	-0.027	0.029	0.018	0.019	0.004	0.003	-0.004	-0.004
SE	0.065	0.060	0.054	0.052	0.066	0.057	0.052	0.050	0.069	0.059	0.058	0.052
SD	0.078	0.082	0.065	0.061	0.079	0.068	0.060	0.054	0.082	0.070	0.065	0.055
CP	0.954	0.948	0.972	0.966	0.966	0.972	0.968	0.988	0.942	0.980	0.978	0.974

6 Application of AIDS data

Doubly censored data that include the AIDS incubation period—the time between HIV infection and the onset of the disease. Within this dataset, the initial event (HIV infection) and subsequent event (onset of AIDS-related symptoms) may be precisely observed, interval-censored, or subject to right censoring. Consequently, the exact occurrence time of each event is either known or confined to a defined interval. We present the proposed approach to analyze such doubly censored data, which has been applied in studies by [10, 17, 19] among others.

The analyses concentrate on 188 individuals diagnosed with HIV infection; 23 of whom have interval-censored measurements on the development of AIDS-related symptoms, while 147 of these individuals experienced right censoring. These patients were categorized into two cohorts—lightly and heavily treated patients—based on the quantity of blood factor they received. The main objective of the analysis is to examine the association between the AIDS incubation period and treatment Z as well as the inconclusiveness between the AIDS incubation period T^* and the duration of HIV infection S. With Z and S as covariates, we fit the proportional hazard model to analyze the AIDS incubation period T^*. Let $Z = 1$ indicate that the patient belongs to the heavily treated group, and $Z = 0$ otherwise. This binary variable will be used to implement the proposed approach.

Table 2: Findings on the impact of a covariate on the association in the AIDS study.

Covariates	**Log Hazard Ratio**	**Standard Error**	**p-values**
Z	0.378	0.161	0.019
S	0.580	0.049	<0.001

Table 2 summarizes the outcomes from implementing the developed framework, featuring the estimated log-hazard ratios ($\log \mathrm{HR}$), standard errors (SE), and p-values for the covariate effects of both Z and S from the proportional hazard regression. The treatment group Z shows a significant effect ($\log \mathrm{HR} = 0.378$, $\mathrm{SE} = 0.161$, $p = 0.019$), indicating that patients in the lightly treated group experience significantly longer AIDS incubation periods than those in the heavily treated group, suggesting a reduced risk of AIDS-related signs of infection.

Additionally, The initial event time S exhibits a strong positive association with the incubation period ($\log \mathrm{HR} = 0.580$, $\mathrm{SE} = 0.049$, $p < 0.001$), indicating that later HIV infection corresponds to a delayed onset of AIDS-related symptoms. These results provide a clear association between the AIDS incubation period, treatment, and HIV infection time. Notably, the estimated SEs are smaller than those reported by [21] and [19]. Moreover, the significant p-value for the association between the AIDS incubation period and the initial event time contrasts with [19], who reported inconclusive results at the 5% level of significance. These results align with the simulation findings, confirming that the proposed estimator provides efficient estimates with smaller SEs and enhanced sensitivity in detecting associations.

7 Discussion and Conclusion

In this study, a semiparametric sieve maximum likelihood approach was used to analyse doubly censored data and investigate the association between AIDS incubation and covariates. The proposed method integrates M-spline and I-spline techniques for parametric and nonparametric estimation, establishing the asymptotic properties of both components. Multiple imputation was utilized to accommodate exactly observed and interval-censored event times, providing a practical framework for assessing associations in AIDS incubation data.

Unlike [19], who replaced exactly observed event times with artificial intervals and treated them as doubly censored, the proposed approach directly incorporates exact observations, yielding smaller variance estimates than those of [21] and [19] by retaining more information for inference. Simulation studies confirmed the estimator's efficiency, with improved precision and coverage compared with existing methods.

Nevertheless, several limitations should be noted. First, the sieve estimation procedure relies on smoothness assumptions and constraints to ensure identifiability and stability. Second, the method is computationally intensive; as sample sizes increase, the number of basis functions grows, which can significantly increase computational complexity for very large datasets. Finally, the proposed model assumes moderate-dimensional covariate structures and may require further adaptation for high-dimensional data or variable selection.

Future work may address these limitations by developing data-driven strategies for sieve dimension selection to balance flexibility and computational efficiency, investigating alternative models such as additive hazard or transformation models, and extending the framework to handle high-dimensional covariates with variable selection. Additionally, alternative estimation procedures, such as self-consistency algorithms or nonparametric maximum likelihood estimators, could be explored to further improve computational performance and robustness.

Acknowledgments

Heartfelt gratitude is extended to the Tertiary Education Trust Fund (TETFUND), as well as the University of Maiduguri, for their generous scholarship and financial support. Appreciation is also expressed to the Universiti Teknologi Malaysia for providing the study opportunity and academic resources. All of these contributions have been instrumental in making this research possible.

References

[1] Sun, J. Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies, *Biometrics*, 51(3) (1995), 1096–1104. DOI: `https://doi.org/10.2307/2533008`.

[2] Cai, T. and Cheng, S. Semiparametric regression analysis of doubly censored data, *Biometrika*, 91(2) (2004), 277–290. DOI: `https://doi.org/10.1093/biomet/91.2.277`.

[3] Li, S., Hu, T., Wang, P. and Sun, J. A class of semiparametric transformation models for doubly censored failure time data, *Scandinavian Journal of Statistics*, 45(3) (2018), 682–698. DOI: `https://doi.org/10.1111/sjos.12319`.

[4] Choi, T., Kim, K. H. A. and Choi, S. Semiparametric least-squares regression with doubly-censored data, *Computational Statistics & Data Analysis*, 164 (2021), 107306. DOI: `https://doi.org/10.1016/j.csda.2021.107306`.

[5] Choi, S. and Huang, X. Efficient inferences for linear transformation models with doubly censored data, *Communications in Statistics – Theory and Methods*, 50(9) (2021), 2188–2200. DOI: `https://doi.org/10.1080/03610926.2019.1662046`.

[6] Shen, P. S. The Cox–Aalen model for doubly censored data, *Communications in Statistics – Theory and Methods*, 51(23) (2022), 8075–8092. DOI: `https://doi.org/10.1080/03610926.2021.1887241`.

[7] Sankara, P., Hari, S. and Sreedevi, E. P. Semiparametric regression analysis of doubly censored recurrent event data, *Japanese Journal of Statistics and Data Science*, 7 (2024), 183–202. DOI: `https://doi.org/10.1007/s42081-023-00234-x`.

[8] Sun, J. *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York, 2006.

[9] De Gruttola, V. and Lagakos, S. W. Analysis of doubly-censored survival data, with application to AIDS, *Biometrics*, 45(1) (1989), 1–11. DOI: `https://doi.org/10.2307/2532030`.

[10] Kim, J., De Gruttola, V. and Lagakos, S. W. Analyzing doubly censored data with covariates, with application to AIDS, *Biometrics*, 49(1) (1993), 13–22. DOI: `https://doi.org/10.2307/2532598`.

[11] Goggins, W. B., Finkelstein, D. M. and Zaslavsky, A. M. Applying the Cox proportional hazards model for analysis of latency data with interval censoring, *Statistics in Medicine*, 18(20) (1999), 2737–2747. DOI: https://doi.org/10.1002/(SICI)1097-0258(19991030)18:20<2737::AID-SIM199>3.0.CO;2-7.

[12] Sun, J., Liao, Q. and Pagano, M. Regression analysis of doubly censored failure time data with applications to AIDS, *Biometrics*, 55(3) (1999), 909–914. DOI: `https://doi.org/10.1111/j.0006-341X.1999.00909.x`.

[13] Pan, W. A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies, *Biometrics*, 57(4) (2001), 1245–1250. DOI: `https://doi.org/10.1111/j.0006-341X.2001.01245.x`.

[14] Fang, H. B. and Sun, J. Consistency of nonparametric maximum likelihood estimation of a distribution function based on doubly interval-censored failure time data, *Statistics & Probability Letters*, 55(3) (2001), 311–318. DOI: `https://doi.org/10.1016/S0167-7152(01)00160-2`.

[15] Sun, J., Lim, H. J. and Zhao, X. An independence test for doubly censored failure time data, *Biometrical Journal*, 46(5) (2004), 503–511. DOI: https://doi.org/10.1002/bimj.200310060.

[16] Kim, Y., Kim, J. and Jang, W. An EM algorithm for the proportional hazards model with doubly censored data, *Computational Statistics & Data Analysis*, 57(1) (2013), 41–51. DOI: https://doi.org/10.1016/j.csda.2012.06.001.

[17] Li, S., Sun, J., Tian, T. and Cui, X. Semiparametric regression analysis of doubly censored failure time data from cohort studies, *Lifetime Data Analysis*, 26(2) (2020), 315–338. DOI: https://doi.org/10.1007/s10985-019-09477-x.

[18] Kim, Y. J. Regression analysis of doubly censored failure time data with frailty, *Biometrics*, 62(2) (2006), 458–464. DOI: https://doi.org/10.1111/j.1541-0420.2005.00487.x.

[19] Li, Z. and Owzar, K. Fitting Cox models with doubly censored data using spline-based sieve marginal likelihood, *Scandinavian Journal of Statistics*, 43(2) (2016), 476–486. DOI: https://doi.org/10.1111/sjos.12186.

[20] Goetghebeur, E. and Ryan, L. Semiparametric regression analysis of interval-censored data, *Biometrics*, 56(3) (2000), 1139–1144. DOI: https://doi.org/10.1111/j.0006-341X.2000.01139.x.

[21] Kang, S. H. and Kim, Y.-J. Association measure of doubly interval censored data using a Kendall's τ estimator, *Communications for Statistical Applications and Methods*, 28(2) (2021), 151–159. DOI: https://doi.org/10.29220/CSAM.2021.28.2.151.

[22] Wong, Y. K., Zhou, Q. and Hu, T. Semiparametric regression analysis of doubly-censored data with applications to incubation period estimation, *Lifetime Data Analysis*, 29(1) (2023), 76–114. DOI: https://doi.org/10.1007/s10985-022-09567-3.

[23] Yuan, C., Zhao, S., Li, S. and Song, X. Sieve maximum likelihood estimation of partially linear transformation models with interval-censored data, *Statistics in Medicine*, 43(30) (2024), 5765–5776. DOI: https://doi.org/10.1002/sim.10225.

[24] Lu, X., Wang, Y., Bandyopadhyay, D. and Bakoyannis, G. Sieve estimation of a class of partially linear transformation models with interval-censored competing risks data, *Statistica Sinica*, 33(2) (2023), 685–704. Available at: https://www.jstor.org/stable/27249936.

[25] Wu, Y., Zhang, Y. and Zhou, J. A spline-based nonparametric analysis for interval-censored bivariate survival data, *Statistica Sinica*, 32(3) (2022), 1541–1562. DOI: https://doi.org/10.5705/ss.202019.0296.

[26] Ramsay, J. O. Monotone regression splines in action, *Statistical Science*, 3(4) (1988), 425–461. DOI: https://doi.org/10.1214/ss/1177012761.

[27] Lu, M., Zhang, Y. and Huang, J. Estimation of the mean function with panel count data using monotone polynomial splines, *Biometrika*, 94(3) (2007), 705–718. DOI: https://doi.org/10.1093/biomet/asm057.

[28] Wu, Y. and Zhang, Y. Partially monotone tensor spline estimation of the joint distribution function with bivariate current status data, *The Annals of Statistics*, 40(3) (2012), 1609–1636. DOI: https://doi.org/10.1214/12-AOS1016.

[29] Wu, Y., Chambers, D. C. and Xu, R. Semiparametric sieve maximum likelihood estimation under cure model with partly interval-censored and left truncated data for application to spontaneous abortion, *Lifetime Data Analysis*, 25(3) (2019), 507–528. DOI: https://doi.org/10.1007/s10985-018-9445-4.

[30] Geman, S. and Hwang, C. Nonparametric maximum likelihood estimation by the method of sieves, *The Annals of Statistics*, 10(2) (1982), 401–414. DOI: https://doi.org/10.1214/aos/1176345782.

[31] Zhang, Y., Hua, L. and Huang, J. A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data, *Scandinavian Journal of Statistics*, 37(2) (2010), 338–354. DOI: https://doi.org/10.1111/j.1467-9469.2009.00680.x.

[32] Joly, P., Commenges, D. and Letenneur, L. A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia, *Biometrics*, 54(1) (1998), 185–194. DOI: https://doi.org/10.2307/2534006.

[33] Bickel, P. J., Klassen, A. J. C., Ritov, Y. and Wellner, J. A. *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York, 1993.

Appendix: Generalized Gradient Projection Algorithm

Algorithm 1 Generalized Gradient Projection (GGP) for Constrained Likelihood Optimization

Require: Initial parameter vector $\theta^{(0)}$, negative log-likelihood function $\ell(\theta)$, gradient $\nabla\ell(\theta)$, Hessian $H(\theta)$, constraint matrix A, number of imputations D, tolerance `tol`, maximum iterations `max_iter`

Ensure: Final estimate $\hat{\theta}$, variance estimates (within- and between-imputation)

procedure Stage 1: Gradient and Hessian Computation Across Imputations
    for $d = 1, \ldots, D$ **do**
        Compute $g_d \leftarrow \nabla\ell_d(\theta)$
        Compute $H_d \leftarrow H_d(\theta)$
    end for
    Compute averages: $\bar{g} \leftarrow \frac{1}{D}\sum_d g_d, \quad \bar{H} \leftarrow \frac{1}{D}\sum_d H_d$
end procedure

procedure Stage 2: GGP Iterative Optimization
    Initialize $\theta \leftarrow \theta^{(0)}$
    for `iter` $= 1, \ldots,$ `max_iter` **do**
        Regularize Hessian: $H_{\text{reg}} \leftarrow \bar{H} + \lambda I$
        $W \leftarrow H_{\text{reg}}^{-1}$
        Projection matrix: $P \leftarrow I - WA^\top(AWA^\top + \epsilon I)^{-1}A$
        Compute projected Newton direction: $d \leftarrow -PW\bar{g}$
        Apply trust-region scaling to d
        Perform line search to choose step size α
        Update: $\theta \leftarrow \theta + \alpha d$
        Enforce monotonicity and non-negativity constraints
        if $\max|\Delta\theta| <$ `tol` **or** $|\Delta\ell| <$ `tol` **then**
            break
        end if
    end for
end procedure

procedure Stage 3: Variance Estimation
    for $d = 1, \ldots, D$ **do**
        Compute within-imputation variance: $V_d \leftarrow \text{diag}(H_d^{-1})$
    end for
    Compute between-imputation variance B and total variance $T \leftarrow B + \text{diag}(\text{mean}(V_d))$
end procedure
return $\hat{\theta} = \theta, T$