entropy

MDPI

Article

Modified Local Linear Estimators in Partially Linear Additive Models with Right-Censored Data Based on Different Censorship Solution Techniques

Ersin Yılmaz [1,*], Dursun Aydın [1] and S. Ejaz Ahmed [2]

[1] Department of Statistics, Mugla Sıtkı Kocman University, Mugla 48000, Turkey; duaydin@mu.edu.tr
[2] Department of Mathematics and Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada; sahmed5@brocku.ca
* Correspondence: ersinyilmaz@mu.edu.tr

Abstract: This paper introduces a modified local linear estimator (LLR) for partially linear additive models (PLAM) when the response variable is subject to random right-censoring. In the case of modeling right-censored data, PLAM offers a more flexible and realistic approach to the estimation procedure by involving multiple parametric and nonparametric components. This differs from the widely used partially linear models that feature a univariate nonparametric function. The LLR method is employed to estimate unknown smooth functions using a modified backfitting algorithm, delivering a non-iterative solution for the right-censored PLAM. To address the censorship issue, three approaches are employed: synthetic data transformation (ST), Kaplan–Meier weights (KMW), and the kNN imputation technique (kNNI). Asymptotic properties of the modified backfitting estimators are detailed for both ST and KMW solutions. The advantages and disadvantages of these methods are discussed both theoretically and practically. Comprehensive simulation studies and real-world data examples are conducted to assess the performance of the introduced estimators. The results indicate that LLR performs well with both KMW and kNNI in the majority of scenarios, along with a real data example.

Keywords: partially linear additive models; local linear regression; right-censored data; synthetic data; kNN imputation

Citation: Yılmaz, E.; Aydın, D.; Ahmed, S.E. Modified Local Linear Estimators in Partially Linear Additive Models with Right-Censored Data Based on Different Censorship Solution Techniques. *Entropy* **2023**, *25*, 1307. https://doi.org/10.3390/e25091307

Academic Editors: Donald J. Jacobs, Abhijit Mandal and Suneel Babu Chatla

Received: 12 July 2023
Revised: 31 August 2023
Accepted: 6 September 2023
Published: 7 September 2023

1. Introduction

Partially linear models (PLMs) have gained considerable attention in the field of survival analysis, especially for modeling right-censored data. The flexibility and capability of PLMs to capture both parametric and nonparametric components make them a favored choice for analyzing survival data with complex relationships. The classical PLM is expressed as follows for completely observed data with a sample size n:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + f(t_i) + \varepsilon_i,\ 1 \le i \le n \tag{1}$$

where y_i's are the completely observed response values (or lifetimes in survival analysis), $\mathbf{x}_i \in \mathbb{R}^{n \times p}$ are the parametric covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ denotes the $(p \times 1)$ dimensional vector of regression coefficients, and $f(.)$ is the univariate unknown smooth function to be estimated based on the values of the nonparametric covariate t_i's. Finally, ε_i's are the random error terms with $(i)\ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $(ii)\ Cov(\varepsilon_i, \mathbf{x}_i) = 0$, $(iii)\ E[\varepsilon_i | x_i, t_i] = 0$. Without censored data, model (1) has been studied by many researchers, and some of the notable studies include [1,2], among others. Additionally, ref. [3] proposed the local linear regression (LLR) estimation for model (1). In the right-censored case, the response variable, y_i, is incompletely observed and censored from the right by random censoring variable

$\{c_i\}_{i=1}^n$ under the assumption that x_i and t_i are completely observed. Accordingly, the censoring mechanism and some new variables can be obtained as follows:

$$z_i = \min(y_i, c_i) \text{ with } \delta_i = \begin{cases} 0, \ if \ y_i \ is \ censored \ (y_i > c_i) \\ 1, \ if \ y_i \ is \ uncensored \ (y_i \leq c_i) \end{cases} \tag{2}$$

where z_i denotes the incompletely observed response variable with the censoring indicator δ_i. Thus, instead of y_i, data pairs $\{z_i, \delta_i\}$ are used in the modeling procedure. There are several important studies on the estimation of model (1) under right-censored data, as given in (2), such as refs. [4–6], among others.

While model (1) offers reliable performance for both censored and uncensored data due to its ability to incorporate both parametric and nonparametric components, it encompasses only a singular nonparametric component. This constraint necessitates that researchers select a sole nonparametric covariate from the dataset, a premise that might not align with many real-world situations. Furthermore, adhering to this limitation could result in less dependable estimations unless the dataset genuinely contains only one nonparametric covariate. To improve estimation accuracy and provide a more adaptable model that considers the right-censored response variable, z_i, this research delves into the partially linear additive model (PLAM), tailored for q nonparametric functions:

$$z_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^{q} f_j(t_{ij}) + \varepsilon_i, \ 1 \leq i \leq n \tag{3}$$

Here, q represents the number of nonparametric components, a value determined based on the nature of the relationship between t_{ij} and y_i. When this relationship cannot be adequately captured by a linear parametric component, it is treated as a nonparametric covariate, characterized by an unknown smooth function $f_j(t_{ij})$. As a result, the overall nonparametric component of model (3) is formed by the summation of these functions. The use of PLAMs in survival analysis with right-censored data allows for more realistic modeling of the relationship between covariates and survival outcomes by incorporating both multiple parametric and nonparametric components. By introducing nonparametric components, PLAMs provide a more adaptable framework for capturing potential nonparametric relationships between covariates and survival times. It is crucial to acknowledge that model (3) cannot be estimated unless the censorship problem is suitably addressed. Numerous studies in the literature have concentrated on estimating (3) for data that is fully observed and devoid of any censoring. Ref. [7] discussed the combination of smoothing splines with semiparametric additive models, while ref. [8] studied the asymptotic properties of M-estimators for model (3). Additionally, Ref. [9] presented a comprehensive review of partially linear additive models based on various smoothing techniques.

Distinct from the studies previously mentioned, this paper presents modified LLR estimators for PLAM (3) using three distinct censoring solutions: synthetic data transformation (ST), Kaplan–Meier weights (KMW), and kNN imputation (kNNI). Through the examination of these modified estimators and the exploration of various techniques to tackle censorship, valuable insights can be gained, and the accuracy and effectiveness of modeling right-censored data may be improved. This paper also explains the procedure for obtaining these estimators, encompassing the modified backfitting technique and a non-iterative approach, accompanied by comparative numerical studies. To the best of our knowledge, this research fills a gap in the literature on modeling right-censored data.

The remaining part of the paper is organized as follows: In Section 2, the fundamentals of right-censored data are presented, and solution approaches are explained. Section 3 covers the estimation of PLAM using modified LLR estimators based on various censorship solution techniques. In Section 4, the statistical properties of the estimators are provided. Sections 5 and 6 present simulation and real data studies, respectively. Finally, Section 7 includes the conclusions of the paper.

2. Right-Censored Data and Solution Methods

In this section, we provide theoretical insights into modeling right-censored data. Let F and G represent the probability distribution functions of the F observed response variable (y_i) and the censoring variable (c_i), respectively. Thus, for any arbitrary data point "u", these functions can be expressed as follows:

$$F(u) = P(y_i \leq u) \text{ and } G(u) = P(c_i \leq u), \tag{4}$$

It is essential to highlight that the estimation procedure for the model, utilizing the specified distributions (4), critically relies on two "censorship assumptions". These constrain all variables within model (2). These assumptions, as outlined by ref. [10] and elaborated by ref. [11] in the context of right-censored regression models, hold significant significance. In essence, the dataset must meet the subsequent criteria.

A1. *y_i and c_i are independent.*

A2. $P(y_i \leq c_i | y_i, \mathbf{x}_i, t_{ij}) = P(y_i \leq c_i | y_i)$.

The assumption (A1) and (A2) can be explained as follows: (A2) posits that the covariates in the model lack any information about the censorship in y_i. Assumption (A1) is particularly crucial when implementing censorship solutions. For a more in-depth discussion, one can refer to [10]'s writings. Drawing from the aforementioned details, this section provides the three censorship solutions. Additionally, towards the section's close, a figure is showcased to illustrate the practical distinctions between synthetic data transformation and the kNN imputation methods.

Synthetic data transformation: To incorporate the impact of censorship into the modeling procedure, synthetic data transformation is a commonly employed solution method. Consequently, the incomplete response pairs $\{(z_i, \delta_i),\ i = 1, \ldots, n\}$ must be substituted for a synthetic response variable, as proposed by ref. [12]. Assuming that G is a continuous and known function, it becomes possible to modify the observed lifetimes z_i in a manner that ensures an unbiased estimation:

$$z_{iG} = \frac{\delta_i z_i}{1 - G(z_i)},\ i = 1, 2, \ldots, n \tag{5}$$

where z_{iG} represents the synthetic response variable with $E\left[z_{iG} | \mathbf{x}_i, t_{ij}\right] = E\left[z_i | \mathbf{x}_i, t_{ij}\right] = \mathbf{x}_i \boldsymbol{\beta} + \sum_{j=1}^{q} f_j(t_{ij})$. Nevertheless, the true distribution of the censoring variable G remains unknown. To address this challenge, ref. [12] suggested replacing G with its estimated version, known as the Product-Limit estimator (Kaplan–Meier estimator). This estimator calculates the survival probabilities at the arbitrary positive data point "u" as follows:

$$1 - \hat{G}(u) = \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1} \right)^{\mathrm{I}[z_{(i)} \leq u,\ \delta_{(i)} = 0]},\ u \geq 0 \tag{6}$$

where $z_{(1)} \leq, \ldots, \leq z_{(n)}$ are the sorted values of the right-censored response variable $z_{(i)}$ and $\delta_{(i)}$ are the corresponding censoring indicators associated to $z_{(i)}$. Hence, instead of $G(z_i)$ in (5), $\hat{G}(z_i)$ is used and $\mathbf{z}_{\hat{G}} = \left(z_{1\hat{G}}, \ldots, z_{n\hat{G}}\right)^T$ can be obtained to fit the PLAM.

Kaplan–Meier weights: Kaplan–Meier weights (KMW), as proposed by ref. [13], are a technique used in survival analysis to address the issue of right-censored data. The Kaplan–Meier estimator is a nonparametric method prevalent nonparametric approach used for estimating survival probabilities amidst censoring. Nonetheless, using standard regression techniques on censored data can lead to biased outcomes. Stute (1993) addressed this by presenting Kaplan–Meier weights, derived from the Kaplan–Meier survival probabilities for each data point. These weights are used to adjust the contribution of each observation in the regression analysis, effectively accounting for the censoring mechanism. By incorporating

the Kaplan–Meier weights into the regression model, unbiased estimates of the regression coefficients can be obtained.

Before computing the KMW, let us assume that $z_{(i)}$ denotes the ordered values of the incomplete response values and $\mathbf{x}_{(i)}^T$, $\delta_{(i)}$ and $\mathbf{t}_{(i)} = \left(t_{(i)1}, \ldots, t_{(i)q}\right)$ are the correspondingly ordered values. Then, Kaplan–Meier weight $w_{(i)}$, associating with the $z_{(i)}$, is computed based on the Kaplan–Meier estimator $\hat{F}\left(z_{(i)}\right)$ given in (5) as follows:

$$w_{(i)} = \hat{F}\left(z_{(i)}\right) - \hat{F}\left(z_{(i-1)}\right) = \frac{\delta_{(i)}}{n-i+1} \prod_{r=1}^{i-1} \left(\frac{n-r}{n-r+1}\right)^{\delta_{(r)}} \tag{7}$$

And KMW is obtained for all possible values of z_i as a diagonal matrix $\mathbf{W} = \text{diag}\left(w_{(1)}, \ldots, w_{(n)}\right)$. To reach further information about (7) and implanting these weights into the regression models, see refs. [5,6].

kNN imputation method: kNN imputation is a prevalent technique for addressing missing data across various domains, as discussed by researchers including [14]. Additionally, some studies, such as ref. [15], have adapted the kNN imputation method to manage right-censored data. This method allows for the practical estimation of right-censored data points without the constraints of theoretical limitations. In this context, we provide a succinct overview of the kNN imputation technique and an algorithm tailored for the PLAM dataset. Essentially, the kNN method is a machine learning technique that hinges on the similarity between data points, utilizing distance metrics for predictions. The choice of a suitable similarity measure can greatly impact the results. The Euclidean norm is commonly employed as a measure of distance in numerous studies. The Euclidean norm is a well-known distance and can be computed for the context of censored data points as $d_E\left(x_j, x_i\right) = \sqrt{\sum_{i=1}^{n_c}\left(x_j^c - x_i^c\right)^2}$ where n_c is the number of censored data points and x_j^c and x_i^c denote the j^{th} and i^{th} associated values of a regressor which has a strong correlation between response variable z_i. Details are provided in Algorithm 1. For imputation, the algorithm introduced by ref. [15] can be employed. The choice of the appropriate number of neighbors, "k", is pivotal, especially given the possibility of some neighbors being right-censored. While ref. [16] suggests a smaller value for "k", such as 1 or 2, an optimal "k" ranging between 2 to 10 is chosen in this context to minimize the mean squared error (MSE). This approach ensures precision in imputation, taking into account the distinct attributes of the data.

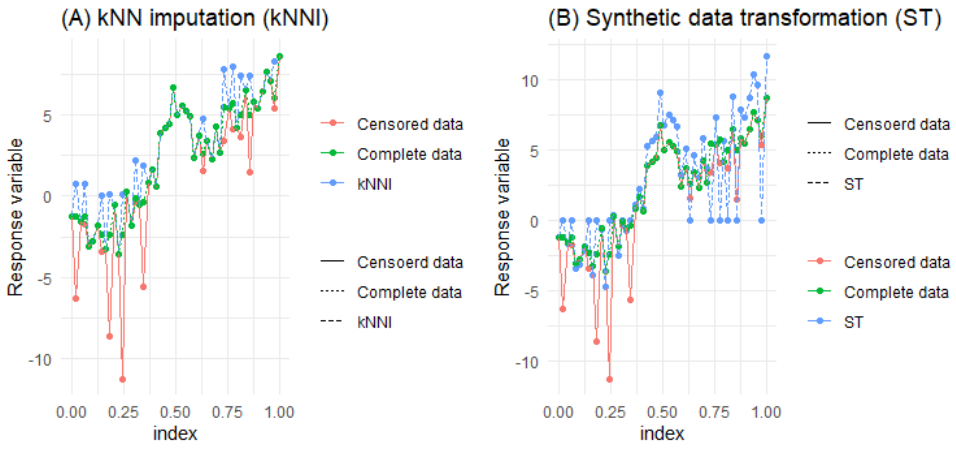

Figure 1. Working procedures of ST in panel (**A**) and KNNI in panel (**B**) for generated data.

Algorithm 1 Algorithm for k NN imputation for the right-censored data

Inputs
I1 : Right − censored dataset z_i
I2 : Censoring indicator δ_i
I3 : Number of nearest neigbours k
I4 : Values of predictor variable x_i (high − correlated one with z_i)
Output : Imputed dataset $\mathbf{z}^{knn} = \left(\mathbf{z}_1^{knn}, \ldots, \mathbf{z}_n^{knn}\right)^T$
1: **begin**
2: **for** $(i = 1$ to $n)$**do**
3: **if**$(\delta_i = 0)$ **do** (if data point is censored)
4: **for**$(j = 1$ to $n)$ **do**
5: Find the distances between x_j *and* x_i for each censored data point
6: Sort the distances from small to large
7: **for** $(j = 1$ to $k)$ **do**
8: Take the first *uncensored* k values of z_i associated to sorted distances
9: Calculate the $i$thimputed value $\left(z_i^{knn}\right)$ with average of nearest $k-$ records of z_i
10: Replace the imputed values $\left(z_i^{knn}\right)$ with censored data points $(z_i, \delta_i = 0)$ in censored data set $\mathbf{z} = (z_1, \ldots, z_n)$
11: Return $\mathbf{z}^{knn} = \left(z_1^{knn}, \ldots, z_n^{knn}\right)^T$
12: **end**

As previously mentioned, Figure 1 has been created to illustrate the practical distinctions between the manipulative solution techniques, namely ST and kNNI. This visualization provides insights into how these methods impact the response variable and the changes they bring about. It should be noted that the effect of KMW is not demonstrated in the figure since it is incorporated into the objective function of the right-censored PLAM as weights. However, further explanation regarding KMW will be provided in the next section when obtaining the modified LLR estimators.

3. Modified Estimator for PLAM

3.1. Fundamentals of PLAM

Before explaining the modified LLR estimators, this section provides a concise overview of the fundamental concepts of PLAM and summarizes the steps involved in utilizing the backfitting algorithm. Additionally, we express right-censored PLAM (3) in vector and matrix form as follows:

$$\mathbf{Z} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^{q} \mathbf{f}_j + \boldsymbol{\varepsilon} \tag{8}$$

Below, we present the explicit expressions for the vector and matrices in (8) as follows:

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}, \ \mathbf{f}_j = \begin{bmatrix} f_j(t_{j1}) \\ \vdots \\ f_k(t_{jn}) \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{9}$$

The literature offers only a handful of studies specifically addressing the right-censored partially linear additive model (PLAM). In terms of estimating model (8), ref. [17] presented the primary optimization problem for the nonparametric additive model, which mean $\mathbf{X}\boldsymbol{\beta} = 0$ in model (8), and ref. [18] formulated a similar problem for (8) as follows:

$$\min_{\beta, f} E\left[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \beta_0 - \sum_{j=1}^{q} \mathbf{f}_j\right]^2 \tag{10}$$

Accordingly, the solution expression for the j^{th} function $f_j(\mathbf{z}_j)$ in the objective (10) can be written as $f_j(\mathbf{t}_j) = E\left[\left\{\mathbf{Y} - \sum_{k\neq j} f_k(\mathbf{z}_k)\right\} | \ \mathbf{z}_j\right]$ and, based on this statement, the following equation system can be used for the general solution of the model. Accordingly, let $(\mathbf{S}_1, \ldots, \mathbf{S}_q)$ be smoothing matrices obtained from the LLR procedure. Then, the equation system for the estimation of model (8) can be obtained as follows:

$$\begin{bmatrix} \mathbf{I} & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_q & \mathbf{S}_q & \cdots & \mathbf{I} \end{bmatrix}_{(nq)\times(nq)} \begin{bmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_q \end{bmatrix}_{(nq\times 1)} = \begin{bmatrix} \mathbf{S}_1(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \mathbf{S}_2(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \vdots \\ \mathbf{S}_q(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{bmatrix}_{(nq\times 1)} \tag{11}$$

where $\hat{\boldsymbol{\beta}}$ denotes estimated coefficients by LLR, which is shown in Section 3.2. For further details on (11), refer to [9]. The solution to system (11) effectively yields the estimates of the functions $\left\{f_j(\mathbf{z}_j)\right\}_{j=1}^{q}$. However, it is evident that inverting the matrix on the left-hand side of (11), which comprises the smoothing matrices, becomes infeasible if the dimension of $(nq \times nq)$ is sufficiently large. As the dimension grows, solving the system in (11) becomes progressively more challenging, potentially reaching a point where it is unmanageable and cannot be directly addressed (refer to [18]).

Hence, in practical applications, the system (11) is typically solved using the backfitting method, incorporating initial-valued components notated as $\left\{\hat{\mathbf{f}}_j^0\right\}_{j=1}^{q}$. Consequently, the LLR estimators are derived by the modified backfitting algorithm, which is given at the end of Section 3.

3.2. Local Linear Regression

Local linear regression (LLR) is a widely employed smoothing technique for nonparametric, semiparametric, and additive models. Its effectiveness has been demonstrated across diverse domains, such as medical research, engineering, and the analysis of time-to-event (or survival) data in time-series studies. In this section, we present three LLR estimators for the partially linear additive model (PLAM) described in (8), employing the introduced censorship solution methods. These estimators are derived using a modified backfitting algorithm. Local linear regression (LLR) is a kernel-based method that differs from kernel regression in that it performs a local estimation of a line rather than a constant. To illustrate the working procedure of LLR, let us consider a partially linear model with a univariate function when $q = 1$, as given in (1), involving an unknown smooth function $f(.)$. The key concept of LLR is to estimate model (1) linearly within small input intervals. To estimate the parameters of (1), the backfitting algorithm introduced by ref. [19] is used. Accordingly, the backfitting estimators $\left(\hat{\boldsymbol{\beta}}\ , \hat{\mathbf{f}}\ \right)$ for model (1) where $\hat{\mathbf{f}}_1 = (f_1(t_1), \ldots, f_1(t_n))^T$ by replacing the corresponding matrices that are $\mathbf{S}_{h_1}$ and $\mathbf{H}_1$ in the algorithm given in Algorithm 2 can be obtained where $\mathbf{H}_1 = \mathbf{S}_{h_1} + \tilde{\mathbf{X}}(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}'\left(\mathbf{I} - \mathbf{S}_{h_1}\right)$ for $\tilde{\mathbf{X}} = \left(\mathbf{I} - \mathbf{S}_{h_1}\right)\mathbf{X}$. Here, $\mathbf{S}_{h_1}$ is computed based on the bandwidth parameter $h_1 > 0$ for LLR, which is formed by using nonparametric variables t_{1i}'s.

In order to adapt the LLR method for estimating the parameters of the right-censored PLAM, a closer examination of the elements of the smoother matrix $\mathbf{S}_{h_j}$ is required. Let $\left\{\mathbf{S}_{h_j}\right\}_j^q$ be written with open form as $\mathbf{S}_{h_j} = \left(\mathbf{s}_{j1}, \ldots, \mathbf{s}_{jn}\right)^T$, where $\left(\mathbf{s}_{j1}, \ldots, \mathbf{s}_{jn}\right)$ show the row vectors of $\mathbf{S}_{h_j}$ obtained from values of h^{th} nonparametric covariate $\mathbf{t}_j = \left(t_{j1}, \ldots, z_{jn}\right)^T$. From the theory of LLR, $\mathbf{s}_{jr}^T$ for any $t_{j1} \leq m \ \leq t_{jn}$ can be obtained as follows:

$$\mathbf{s}_{jm}^T = \mathbf{d}_1^T\left(\mathbf{t}_{jm}^T\mathbb{W}_{jm}\mathbf{t}_{jm}\right)^{-1}\mathbf{t}_{jm}^T\mathbb{W}_{jm}$$

where $\mathbf{t}_{jm}$, $\mathbf{d}_1$, and $\mathbb{W}_{jm}$ can be expressed as follows:

$$\mathbf{t}_{jm} = \begin{bmatrix} 1 & (t_{j1} - m) \\ \vdots & \vdots \\ 1 & (t_{jn} - m) \end{bmatrix}, \ \mathbf{d}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\mathbb{W}_{jm} = \text{diag}\left[h^{-1} K\left(\frac{t_{j1} - m}{h} \right), \cdots, h^{-1} K\left(\frac{t_{jn} - m}{h} \right) \right] \tag{12}$$

Based on the provided information, it can be inferred that the extension of LLR estimators to PLAM requires further adjustments. Moreover, it is crucial to satisfy the standard assumptions of LLR, such as where $K(.)$ is the kernel function, which is continuous, and its moment is written as $\mu_i(K) \equiv \int u^i K(u) du = 0$ when $\mu_2(K) \neq 0$ for odd values of j. The density of t_{ji} can be given as $g_t(m) > 0$, for all $m \in sup(g_t)$, and also, as a common assumption, since $n \to \infty$, $h \to 0$, and $nh \to \infty$. Finally, a second derivative of the nonparametric smooth function $f(.)$ exists and is continuous. Details about the assumptions are discussed in detail in ref. [20].

In the backfitting estimation procedure, to make simple the definition of the model (8), some restrictions on $\{f_j(t_{ij})\}_{j=1}^{q}$ are needed. At first, $E[f_j(t_{ij})] = 0$ is assumed. Secondly, the parametric covariates $\mathbf{x}_i^T$'s and right-censored response values z_i's are assumed to be scaled around zero. In order to construct the centered smoother matrix $\mathbf{S}_{h_j}$ used in the LLR estimation, these constraints are necessary. Thus, the conditional expectation of model (8) can be expressed as follows:

$$E(z_i | \mathbf{x}_i, t_i) = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j=1}^{q} f_j(t_{ij}), \ i = 1, \ldots, n \tag{13}$$

By using the modified backfitting algorithm given in Algorithm 2, solutions can be obtained based on $\mathbf{S}_{h_j}$ for PLAM parameters $\boldsymbol{\beta}$ and $\{\mathbf{f}_j\}_{j=1}^{q}$. Thus, without any censoring adjustment, PLAM estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}}\right)$ based on the LLR are obtained.

Furthermore, it should be noted that ref. [20] presented a non-iterative formulation equivalent to the backfitting algorithm based on an additive smoother matrix $\mathbf{S}^A = \sum_{j=1}^{q} \mathbf{S}_j^*$ to demonstrate the LLR estimation process in the absence of censorship issues, which reveals the relationship between $\mathbf{Z}$ and $\hat{\mathbf{f}}^A = \sum_{j=1}^{q} \hat{\mathbf{f}}_j$. Here, $\mathbf{S}_j^*$ is computed from the equation system (11) based on the $\mathbf{S}_{h_j}$ (see ref. [9]). Additionally, this information elucidates the connection between a unique solution and the iterative backfitting process.

Accordingly, LLR estimators for PLAM can be found as for both ST and kNNI by replacing $\mathbf{Z}$ by $\mathbf{Z}^{ST}$ and $\mathbf{Z}^{kNNI}$:

$$\hat{\boldsymbol{\beta}}^A = \left(\mathbf{X}^T \tilde{\mathbf{X}} \right)^{-1} \mathbf{X}^T \tilde{\mathbf{Z}} \tag{14}$$

$$\hat{\mathbf{f}}^A = \mathbf{S}^A \left(\mathbf{Z} - \alpha_0 - \mathbf{X} \hat{\boldsymbol{\beta}}^A \right) \tag{15}$$

Algorithm 2 Modified Backfitting Algorithm for Right-Censored PLAM

Inputs: $\beta_0 = E(Z_i) = \overline{Z}$; $\mathbf{X}$: $(n \times p)$-dimensional covariates of parametric component
$\mathbf{Z}$: $(n \times q)$-dimensional scaled nonparametric covariates; $\left\{f_k^{(0)}\right\}_{k=1}^{q}$: Initial smooth functions
$\boldsymbol{\beta}^{(0)}$: Initial regression coefficients; $\mathbf{Z}$: $(n \times 1)$-dim. vector of right-censored response values
Tolerance value, $tol = 0.05$ and max. iteration = 100.

Outputs: Modified PLAM estimators:
O1:kNNI basis LLR estimators $\hat{\boldsymbol{\beta}}^{imp}$ and $\left(\hat{\mathbf{f}}_1^{imp}, \ldots, \hat{\mathbf{f}}_q^{imp}\right)$
O2:ST basis estimators $\hat{\boldsymbol{\beta}}^{ST}$ and $\left(\hat{\mathbf{f}}_1^{ST}, \ldots, \hat{\mathbf{f}}_q^{ST}\right)$
O3: KMW basis estimators $\hat{\boldsymbol{\beta}}^{KMW}$ and $\left(\hat{\mathbf{f}}_1^{KMW}, \ldots, \hat{\mathbf{f}}_q^{KMW}\right)$

Begin

Initialize $\boldsymbol{\beta}$ and $(\mathbf{f}_1, \ldots, \mathbf{f}_q)$ as $\boldsymbol{\beta}^{(0)}$ and $\left\{f_j^{(0)}\right\}_{j=1}^{q}$ by covariates $\mathbf{X}$ and $\mathbf{t}_1, \ldots, \mathbf{t}_q$.

while $(tol \geq 0.05)$ and $(i < max.iteration)$

Selection of optimal bandwidth parameter h_j by GCV between steps: 3–8

Create a sequence of tunning parameter $h_{seq} = [0.01,\ 1.5]$ for determined length

  for $(l\ in\ 1 : length)$ **do**

    Compute the smoothing matrix $\mathbf{S}_{h_{seq}}^{(l)}$.

      if censorship solution is **KMW**

        Compute $\tilde{\mathbf{X}}$ and $\mathbf{H}_j^{(l)} = \mathbf{S}_{h_{seq}}^{(l)} + \tilde{\mathbf{X}}(\left(\tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{I} - \mathbf{S}_{h_{seq}}^{(l)}\right)$ where $\tilde{\mathbf{X}} = \left(\mathbf{I} - \mathbf{S}_{h_{seq}}^{(l)}\right)\mathbf{X}$

      Else

        Compute $\tilde{\mathbf{X}}$ and $\mathbf{H}_j^{(l)} = \mathbf{S}_{h_{seq}}^{(l)} + \tilde{\mathbf{X}}(\left(\tilde{\mathbf{X}}'\mathbf{W}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{I} - \mathbf{S}_{h_{seq}}^{(l)}\right)$ where $\tilde{\mathbf{X}} = \left(\mathbf{I} - \mathbf{S}_{h_{seq}}^{(l)}\right)\mathbf{X}$

        Calculate GCV$\left(h_{seq}^{(l)}\right)$ as given in Equation (24)

    end

    Select optimal $\hat{h}_j$ which minimizes $GCV\left(h_j\right)$ for j^{th} function $\mathbf{f}_j$.

    Compute $\mathbf{S}_{\hat{h}_j}$ for each criterion (and method).

Solution of censorship problem between steps: 14–25

    if the censorship solution is **kNNI**

      Replace $\mathbf{Z}$ with $\mathbf{Z}^{imp}$ using algorithm in Algorithm 1.

    if the censorship solution is **ST**

      Replace $\mathbf{Z}$ with $\mathbf{Z}^{ST}$ as shown in Equation (5)

  for $(j\ in\ 1 : q)$ **do**

      if the censorship solution is **KMW**

$$\hat{\boldsymbol{\beta}}_j^{(i)} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\left(\mathbf{Z} - \beta_0 - \sum_{m<j}^{q}\hat{\mathbf{f}}_m^{(i)} - \sum_{m>j}^{q}\hat{\mathbf{f}}_m^{(i-1)}\right)$$

$$\hat{\mathbf{f}}_j^{(i)} = \mathbf{S}_{\hat{h}_j}\left(\mathbf{Z} - \beta_0 - \mathbf{X}\hat{\boldsymbol{\beta}}_j^{(i)} - \sum_{m<j}^{q}\hat{\mathbf{f}}_m^{(i)} - \sum_{m>j}^{q}\hat{\mathbf{f}}_m^{(i-1)}\right)$$

      Else

$$\hat{\boldsymbol{\beta}}_j^{(i)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Z} - \beta_0 - \sum_{m<j}^{q}\hat{\mathbf{f}}_m^{(i)} - \sum_{m>k}^{q}\hat{\mathbf{f}}_m^{(i-1)}\right)$$

$$\hat{\mathbf{f}}_j^{(i)} = \mathbf{S}_{\hat{\lambda}_k}\left(\mathbf{Y} - \alpha_0 - \mathbf{X}\hat{\boldsymbol{\beta}}_k^{(i)} - \sum_{m<k}^{q}\hat{\mathbf{f}}_m^{(i)} - \sum_{m>k}^{q}\hat{\mathbf{f}}_m^{(i-1)}\right)$$

      end

$$i = i + 1$$

$$tol = (nq)^{-1}\left|\left(\mathbf{f}_k^{(i)} - \mathbf{f}_k^{(i-1)}\right)^T\mathbf{1}\right| \text{ where } \mathbf{1} = (1, \ldots, 1)^T.$$

end

Return $\hat{\boldsymbol{\beta}}$ and $\left(\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_q\right)$

end

And for KMW solution, non-iterative estimators are obtained as follows:

$$\hat{\boldsymbol{\beta}}_{KMW}^{A} = \left(\mathbf{X}^T\mathbf{W}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^T\mathbf{W}\tilde{\mathbf{Z}} \tag{16}$$

$$\hat{\mathbf{f}}_{KMW}^{A} = \mathbf{S}^{A}\left(\mathbf{Z} - \alpha_0 - \mathbf{X}\hat{\boldsymbol{\beta}}^{A}\right) \tag{17}$$

where $\tilde{\mathbf{X}} = \left(\mathbf{I} - \mathbf{S}^{A}\right)\mathbf{X}$, $\tilde{\mathbf{Z}} = \left(\mathbf{I} - \mathbf{S}^{A}\right)\mathbf{Z}$. It should be noted that the validity of Equations (14)–(17) depends on the existence of a unique solution. Furthermore, the vector of fitted values for LLR can be expressed as follows:

$$\hat{\boldsymbol{\mu}} = E[\mathbf{Z}|\mathbf{X}, \mathbf{Z}] = \hat{\mathbf{Z}} = \mathbf{H}^{A}\mathbf{Z} \tag{18}$$

where $\mathbf{H}^{A} = \mathbf{S}^{A} + \tilde{\mathbf{X}}\left[\tilde{\mathbf{X}}^{T}\tilde{\mathbf{X}}\right]^{-1}\mathbf{X}^{T}\left(\mathbf{I} - \mathbf{S}^{A}\right)$ and for the KMW solution $\mathbf{H}_{KMW}^{A} = \mathbf{S}^{A} + \tilde{\mathbf{X}}\left[\tilde{\mathbf{X}}^{T}\mathbf{W}\tilde{\mathbf{X}}\right]^{-1}\mathbf{X}^{T}\mathbf{W}\left(\mathbf{I} - \mathbf{S}^{A}\right)$. Note that under completely observed data, $\mathbf{H}^{A}$ is derived by [21] for the LLR estimator of PLAM.

To effectively demonstrate and interpret each nonparametric component individually, the introduced modified backfitting algorithm is more suitable than Equations (16)–(18), which yield an additive outcome for the nonparametric component. Additionally, computing $\mathbf{S}_{LL}^{A}$ becomes significantly challenging as the dimension of the additive component increases. In this paper, the modified backfitting estimators $\left(\hat{\boldsymbol{\beta}}^{A}, \hat{\mathbf{f}}^{A}\right)$ of LLR, obtained through an algorithm given in Algorithm 2, are employed. This approach aims to showcase the performance of the estimated functions $\hat{\mathbf{f}} = \left\{\hat{\mathbf{f}}_j\right\}_{j=1}^{q}$. In the introduced algorithm given in Algorithm 2, to calculate the selection criterion GCV, the degrees of freedom of (DF) are computed by $DF_j = tr\left[\left(\mathbf{I} - \mathbf{H}_j\right)^{T}\left(\mathbf{I} - \mathbf{H}_j\right)\right] = n - 2tr\left(\mathbf{H}_j\right) + tr\left(\mathbf{H}_j^{T}\mathbf{H}_j\right)$ where $\mathbf{H}_j$ denotes the hat matrix based on the j^{th} nonparametric component. Also, to see details about the algorithm given in Algorithm 2, see ref. [9].

4. Properties of the Estimator

The objective of this section is to assess the bias and variance of the modified LLR estimators introduced in the previous section. When evaluating the performance of the parametric component, the variances and biases of the regression coefficients are calculated using the non-iterative solutions given in Equations (14)–(17), owing to its theoretical simplicity.

Empirical studies can be conducted to calculate the bias and variance properties of the estimators. However, when considering LLR as demonstrated in Equations (14)–(17), non-iterative formulations can be employed to compute finite-sample properties for the other two methods. In this matter, conditional bias $E\left[\left(\hat{\boldsymbol{\beta}}^{A} - \boldsymbol{\beta}\right)\middle|\mathbf{X}, \mathbf{t}\right]$ and variance $Var\left(\hat{\boldsymbol{\beta}}^{A}\right)$ are obtained based on Equations (14)–(17).

Let us rewrite $\hat{\boldsymbol{\beta}}^{A}$ as:

$$\hat{\boldsymbol{\beta}}^{A} = \boldsymbol{\beta} + \left(\mathbf{X}^{T}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^{T}\tilde{\mathbf{f}}^{A} + \left(\mathbf{X}^{T}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^{T}\left(\mathbf{I} - \mathbf{S}^{A}\right)\boldsymbol{\varepsilon}$$

where $\mathbf{S}^{A} = \sum_{j=1}^{q}\mathbf{S}_j^{*}$, and $\tilde{\mathbf{f}}^{A} = \left(\tilde{\mathbf{f}}_1 + \ldots + \tilde{\mathbf{f}}_q\right)$ for $\left\{\tilde{\mathbf{f}}_j = \left(\mathbf{I} - \mathbf{S}_{h_k}\right)\mathbf{f}_j\right\}_{j=1}^{q}$. Then $B\left(\hat{\boldsymbol{\beta}}^{A}\right)$ and $Var\left(\hat{\boldsymbol{\beta}}^{A}\right)$ can be given by:

$$B\left(\hat{\boldsymbol{\beta}}^{A}\right) = E\left[\left(\hat{\boldsymbol{\beta}}^{A} - \boldsymbol{\beta}\right)\middle|\mathbf{X}, \mathbf{t}\right] = \left(\mathbf{X}^{T}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^{T}\tilde{\mathbf{f}}^{A} \tag{19}$$

$$Var\left(\hat{\boldsymbol{\beta}}^{A}\right) = \hat{\sigma}_{\varepsilon}^{2}\left(\mathbf{X}^{T}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^{T}\left(\mathbf{I} - \mathbf{S}^{A}\right)^{2}\mathbf{X}\left(\mathbf{X}^{T}\tilde{\mathbf{X}}\right)^{-1} \tag{20}$$

And for the KMW solution, Equations (19) and (20) are given by:

$$B\left(\hat{\boldsymbol{\beta}}_{KMW}^{A}\right) = E\left[\left(\hat{\boldsymbol{\beta}}^{A} - \boldsymbol{\beta}\right)\middle|\mathbf{X}, \mathbf{t}\right] = \left(\mathbf{X}^{T}\mathbf{W}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^{T}\mathbf{W}\tilde{\mathbf{f}}_{KMW}^{A} \tag{21}$$

$$Var\left(\hat{\boldsymbol{\beta}}_{KMW}^{A}\right) = \hat{\sigma}_{\varepsilon}^{2}\left(\mathbf{X}^{T}\mathbf{W}\tilde{\mathbf{X}}\right)^{-1}\mathbf{X}^{T}\mathbf{W}\left(\mathbf{I} - \mathbf{S}^{A}\right)^{2}\mathbf{X}\left(\mathbf{X}^{T}\mathbf{W}\tilde{\mathbf{X}}\right)^{-1} \tag{22}$$

where $\hat{\sigma}_{\varepsilon}^{2}$ is the model variance estimated based on LLR and it can be computed using the hat matrix $\mathbf{H}^{A}$ or $\mathbf{H}_{KMW}^{A}$ for the KMW solution that are defined after Equation (18). In addition, one can replace $\mathbf{Z}$ by $\mathbf{Z}^{ST}$ or $\mathbf{Z}^{imp}$. Accordingly, $\hat{\sigma}_{\varepsilon}^{2}$ is formulated as follows:

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{\mathbf{Z}^{T}\left(\mathbf{I} - \mathbf{H}^{A}\right)^{\mathbf{T}}\left(\mathbf{I} - \mathbf{H}^{A}\right)\mathbf{Z}}{tr\left[\left(\mathbf{I} - \mathbf{H}_{LL}^{A}\right)^{T}\left(\mathbf{I} - \mathbf{H}_{LL}^{A}\right)\right]} \tag{23}$$

where the degree of freedom (DF), which is given in the denominator of (23), is calculated by $DF_{A} = tr\left[\left(\mathbf{I} - \mathbf{H}^{A}\right)^{T}\left(\mathbf{I} - \mathbf{H}^{A}\right)\right] = n - 2tr\left(\mathbf{H}^{A}\right) + tr\left(\left(\mathbf{H}^{A}\right)^{T}\mathbf{H}^{A}\right)$ and $\mathbf{H}_{KMW}^{A}$ is used for the KMW solution. For the further details of DF_{A}, see ref. [17]. The modified backfitting algorithm provided in Algorithm 2 requires the estimation of the model variance for each individual nonparametric function in order to calculate the GCV score for bandwidth parameter selection. Consequently, if $\mathbf{H}^{A}$ is replaced by $\mathbf{H}_{j}$ or $\mathbf{H}_{KMW_{j}}$ in (23), then the individual variance estimator $\hat{\sigma}_{\varepsilon_{j}}^{2}$ can be easily obtained. The fundamental concept behind computing $\hat{\sigma}_{\varepsilon_{j}}^{2}$ lies in selecting the appropriate smoothing and bandwidth parameters using the GCV criterion, as it relies on the estimated model variance. The GCV criterion can be summarized as follows.

***GCV*criterion**: Generalized cross-validation is used to obtain a minimum score based on the optimal tuning parameter for the regression model. In terms of bandwidth selection in additive models with LLR, ref. [22] presented a detailed work on using GCV and its properties. Accordingly, to choose the optimal h_{j} for j^{th} function $\mathbf{f}_{j}$, $GCV\left(h_{j}\right)$ score can be computed based on $\hat{\boldsymbol{\mu}}$ given in (18):

$$GCV\left(h_{j}\right) = \frac{\left(\mathbf{Z} - \hat{\boldsymbol{\mu}}\right)^{T}\left(\mathbf{Z} - \hat{\boldsymbol{\mu}}\right)}{n\left\{1 - \left(n^{-1}tr\left(\mathbf{H}_{j}\right)\right)\right\}^{2}} \tag{24}$$

where $\mathbf{H}_{j}$ is the hat matrix obtained for $\mathbf{f}_{j}$ which is provided at the end of the Section 3. Notice that calculating the true DF_{j} in PLAM is asymptotically justifiable if parametric and nonparametric covariates $\left(\mathbf{x}_{i}, \mathbf{t}_{j}\right)$ are independent. If there is multicollinearity, then Equation (24) may be regularized properly due to overestimated DF_{j}.

4.1. Evaluation of Performance

4.1.1. Metrics for the Parametric Component

In this section, two metrics are presented to assess the performance of the LLR estimator of the parametric component of the model $\hat{\boldsymbol{\beta}}$ that are scalar versions of the dispersion error (SMDE) and the relative efficiency (RE), which is computed by ratio of the SMDE values. The formulations are given below:

$$SMDE\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) = E\left[\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right] = tr\left[\left(MSE\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right)\right)\right] \tag{25}$$

where $MSE\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right)$ is expressed as a summation of bias square and variance of $\hat{\boldsymbol{\beta}}$, and given by:

$$MSE\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}\right) = E\left[\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right] = Var\left(\hat{\boldsymbol{\beta}}\right) + \left[B\left(\hat{\boldsymbol{\beta}}\right)\right]^{2} \tag{26}$$

Then, using (25), *RE*s of the methods on estimating $\boldsymbol{\beta}$ can be computed. In this paper, methods are considered for use as censorship solution techniques for *RE*s.

Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ represent the estimates of parametric components based on two different censorship solutions. Accordingly, *RE* can be formulated as follows:

$$RE(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = SMDE(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}) / SMDE(\hat{\boldsymbol{\beta}}_2, \boldsymbol{\beta}) \tag{27}$$

where $RE(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) < 1$ indicates that $\hat{\boldsymbol{\beta}}_1$ is more efficient than $\hat{\boldsymbol{\beta}}_2$.

4.1.2. Metrics for the Nonparametric Component

To evaluate the quality of the estimated nonparametric component, two measures are presented. The first measure is the root mean squared error (*RMSE*), which measures the accuracy of each individual estimated function in the model. The second measure is the averaged root mean squared error (*ARMSE*) which is specifically designed to assess the performance of the overall additive component $\hat{\mathbf{f}} = \left(\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_q\right)$. The formulations of *RMSE* and *ARMSE* are written as:

$$RMSE_j\left(f_j, \hat{f}_j\right) = \sqrt{n^{-1} \sum_{i=1}^{n} \left[f_j(z_{ij}) - \hat{f}_j(z_{ij})\right]^2}, \; 1 \le j \le q \tag{28}$$

and

$$ARMSE\left(\mathbf{f}^A, \hat{\mathbf{f}}^A\right) = q^{-1} \sum_{j=1}^{q} RMSE_j\left(\mathbf{f}_j, \hat{\mathbf{f}}_j\right) \tag{29}$$

where $\mathbf{f} = \sum_{j=1}^{q} \mathbf{f}_j$ and $\hat{\mathbf{f}} = \sum_{j=1}^{q} \hat{\mathbf{f}}_j$.

5. Simulation Study

The practical performance of the modified LLR estimators in the context of right-censored PLAM with various censorship solution methods is analyzed in this section. To achieve this, different settings for sample size (n), the number of additive nonparametric components (q), and the level of censoring (CL) are considered. Specifically, three sample sizes (n = 50, 100, and 200) and three levels of censoring (CL = 5%, 20%, and 35%) are chosen. A total of eight scenarios are obtained by combining these configurations. Additionally, a total of 24 cases for analysis are formed by using three censorship solution methods. Moreover, accelerated failure time model estimation results are presented as benchmark performance scores. To achieve that existing function, the survival library in R is used. Note that the function written in R for this paper is provided via link: https://github.com/yilmazersin13/Censored-Partially-linear-additive-models/tree/main, accessed on 9 August 2023. The simulation design and setup used in this study are designed in a manner commonly found in the literature (see ref. [4]). Small, medium, and large sample sizes are chosen, along with three different censoring levels, in accordance with reference articles. Furthermore, the nonparametric component count has been determined in two distinct ways, introducing a novel approach that differs from most similar studies (see ref. [9]).

After establishing the design, the data generation procedure for the right-censored PLAM is outlined here. Firstly, PLAM with completely observed responses is generated as:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{\mathrm{j}=1}^{q} f_j(t_{ji}) + \varepsilon_i, \; 1 \le i \le n \tag{30}$$

where $\mathbf{x}_i^T = (x_{i1}, x_{i2})^T$, is $(n \times 2)$ dimensional parametric covariate matrix with normally distributed and independently $\mathbf{x}_i$'s that are generated as $\mathbf{x}_i \sim N(\mu_x = 0, \sigma_x^2 = 1)$. Also, the vector of regression coefficients is determined as $\boldsymbol{\beta} = (1, -0.5)^T$. Regarding the nonparametric component, smooth functions are generated by $f_1(t_1) = 1 - 48t_1 + 218t_1^2 - 315t_1^3 + 145t_1^4$ with $t_1 = \{(i - 0.5)/n\}_{i=1}^{n}$ and $f_2(t_2) = \sin(2t_2) + 2e^{-16t_2^2}$ with $t_2 = U[-2, 2]$

when $q = 2$. Note that, due to how all the variables are scaled in the simulation study, the constant term α_0 is not used throughout the section. Finally, the random error terms ε_i's are independent and identically distributed with zero mean and constant variance, which can be shown as $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2 = 0.5\right)$.

After generating (30), by applying the censorship procedure given in Algorithm 3, right-censored response variable $\mathbf{Z}$ is generated based on random censoring variable $\mathbf{C} = (c_1, \ldots, c_n)^T$ and censoring indicator $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$.

Algorithm 3 Censoring Procedure

Input: Completely observed y_i
Output: Right-censored dependent variable z_i
1: For given censoring level (CL), produce $\delta_i = I(y_i \leq c_i)$ from the binomial distribution
2: **for** $(i\ in\ 1\ to\ n)$
3: **If** $(\delta_i = 0)$
4: **while** $(y_i \leq c_i)$
5: generate $c_i \sim N\left(\mu_y, \sigma_y^2\right)$
6: **Else**
7: $c_i = z_i$
8: **end** (for loop in Step 2)
9: **for** $(i\ in\ 1\ to\ n)$
10: **If** $(y_i \leq c_i)$
11: $z_i = y_i$
12: **Else**
13: $z_i = c_i$
14: **end** (for loop in Step 9)

Then, right-censored PLAM is obtained with the incomplete response variable $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$. Accordingly, the following figures and tables are provided based on the censorship solution techniques. Algorithms 2 and 3 present the results for the performance of the parametric component estimation, specifically the SMDE and RE values, respectively. In addition, as a benchmark method, the performance of AFT model estimation based on Cox's semiparametric proportional hazards (CPH) estimator is provided in both simulation and real data examples. The estimates are obtained a using "*Survival*" package in R.

Prior to presenting the findings, we offer a visual representation in Figure 2 that elucidates the process of bandwidth selection across diverse scenarios. This illustration sheds light on how the choice of bandwidth is intricately intertwined with the extent of censoring and the specific methods employed for addressing censorship. The discerning eye will note that in the context of f_1, the selection of bandwidth appears to exhibit a lesser degree of sensitivity to variations in the level of censoring and sample size. However, in the case of the f_2 function, it becomes clear that the level of censorship exerts a discernible influence on the chosen bandwidth value. Notably, when confronted with elevated censorship levels across all solution strategies, a preference for smaller bandwidths becomes evident. This outcome is intuitively reasonable since, especially in scenarios involving ST and kNNI, the structural complexity of the data to be fitted takes on a more undulating nature. Therefore, it is evident that we can extrapolate that accounting for the degree of censorship is a pivotal factor when navigating the terrain of bandwidth selection. These findings resonate with prior research in this domain. Ref. [23] demonstrated similar behavior in a related context, highlighting the sensitivity of bandwidth to censorship levels. In line with the in-depth investigations of ref. [24], our observations underscore the need for cautious bandwidth selection in scenarios characterized by substantial censorship, promoting the accurate modeling of intricate data structures.

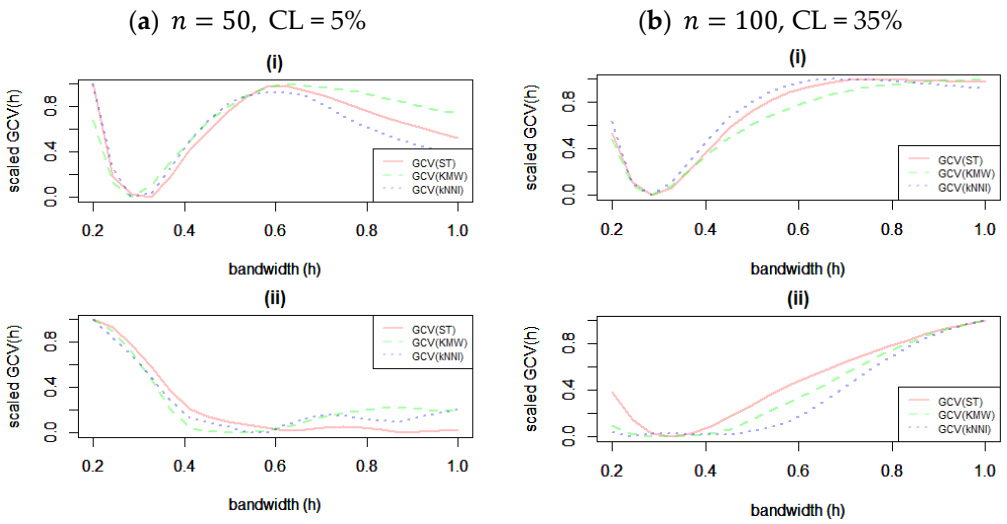

Figure 2. Selection of bandwidth parameter (h) for different scenarios and censorship solution methods when $n = 50$. In each panel, (i) and (ii) involve the selection processes for $f_1(t_1)$ and $f_2(t_2)$, respectively.

The results in Table 1 demonstrate that the estimation quality of the modified LLR estimators for the parametric component $\boldsymbol{\beta}$ improves with lower censoring levels and larger sample sizes across all censorship techniques. These tendencies align with the expected theoretical behavior. Specifically, the LLR-KMW estimator exhibits dominant performance in many simulation combinations, closely followed by the LLR-kNNI estimator with competitive SMDE scores. However, the LLR-ST does not yield good performance. Also, as a benchmark method for the model, SMDE scores of the CPH estimator are presented in the table. It is evident that due to the model involving serious complexity with two different nonparametric functions, there is a significant distance between the LLR-based estimators and the CPH estimator, which is expected.

Table 1. Calculated $SMDE$ values for all simulation combinations.

n	CL	LLR-ST	LLR-KMW	LLR-kNNI	CPH
50	5%	0.561	0.557	**0.545**	0.991
	20%	0.724	0.681	**0.624**	1.029
	35%	1.084	**0.738**	0.744	1.173
100	5%	0.121	**0.103**	0.104	0.702
	20%	0.140	**0.122**	0.135	0.764
	35%	0.168	**0.142**	0.148	0.834
200	5%	0.027	**0.024**	0.026	0.471
	20%	0.031	0.029	**0.028**	0.480
	35%	0.034	**0.031**	0.033	0.497

Bold color denotes the best performance score.

Interestingly, in cases where $n = 50$ and $CL = 5\%$ or $CL = 20\%$, the LLR-kNNI estimator outperforms the LLR-KMW estimator. As the sample size increases, LLR-KMW takes the lead, in accordance with its theoretical behavior. It is worth noting that due to its fully nonparametric nature, LLR-kNNI may yield better results under different configurations, demonstrating relative independence from specific simulation settings. This characteristic is observed in the combination of $n = 200$ and $CL = 20\%$.

Additionally, to assess the impact of censorship on the solution techniques, the increase in SMDE scores between censorship levels is examined. The results indicate that the the LLR-ST estimator is the most affected by censorship, which aligns with the theoretical background of ST presented in Section 2.

In Table 2, the calculation of the RE scores follows a decision where the nominators represent the columns, and the denominators represent the rows. Therefore, an RE value of less than 1 in Table 2 indicates that the method in the column is more effective than the methods in the corresponding row. Please note that, for the sake of saving space, only certain simulation configurations are considered in Table 2. The results in the table confirm that LLR-KMW is more efficient than LLR-ST in all cases. Simultaneously, LLR-KMW and LLR-kNNI exhibit similar outcomes, indicating that they are not distinctly efficient in any simulation configurations for estimating the parametric component of the PLAM.

Table 2. Comparative *RE* scores for the modified LLR estimators.

n	*CL*	Method	LLR-ST	LLR-KMW	LLR-kNNI	CPH
50	5%	LLR-ST	1.000	**0.992**	**0.970**	1.766
		LLR-KMW	1.007	1.000	**0.977**	1.779
		LLR-kNNI	1.030	1.023	1.000	1.818
		AFT	**0.566**	**0.562**	**0.549**	1.000
	35%	LLR-ST	1.000	**0.686**	**0.680**	1.082
		LLR-KMW	1.456	1.000	**0.991**	1.589
		LLR-kNNI	1.468	1.008	1.000	1.576
		AFT	**0.924**	**0.629**	**0.634**	1.000
200	5%	LLR-ST	1.000	**0.974**	**0.918**	6.333
		LLR-KMW	1.025	1.000	**0.942**	7.125
		LLR-kNNI	1.088	1.060	1.000	6.576
		AFT	**0.158**	**0.140**	**0.152**	1.000
	35%	LLR-ST	1.000	**0.963**	**0.920**	5.794
		LLR-KMW	1.038	1.000	**0.956**	6.354
		LLR-kNNI	1.085	1.045	1.000	5.969
		AFT	**0.173**	**0.157**	**0.167**	1.000

Bold color denotes the best performance score.

Furthermore, when the censoring level is very high ($CL = 35\%$), the RE scores deviate from 1, making the performance differences among the LLR estimators based on the solution techniques more apparent. Once again, it is evident that, especially for $n = 50$, ST is the most sensitive technique to censorship compared with the other two methods. Additionally, the results reveal that LLR-kNNI and LLR-KMW display similar RE scores in every combination. In addition, in Table 2, REs of CPH show that there is a clear dominance of LLR-basis estimators for the estimation of right-censored PLAM. This result also proves that the introduced estimator has important potential to be an alternative estimator for the model of interest that is used in survival analysis.

In Figure 3, the averaged values of the RE scores are displayed, confirming the interpretations from Table 2. The figure also shows both the effects of censorship and the sample size. In panel (a), the RE values are very close to each other due to the very low censoring level ($CL = 5\%$). Panels (b) and (c) demonstrate the change in RE scores as the censoring level increases, with the differences between the estimators becoming more distinct, as mentioned earlier. Consequently, the LLR-kNNI and LLR-KMW estimators are more efficient than the LLR-ST estimator. In panel (c), the performances are once again close to each other, reflecting the large sample size ($n = 200$).

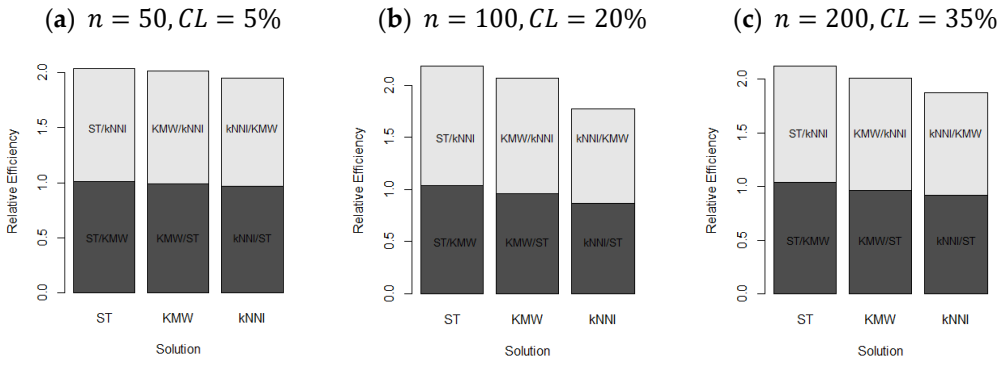

Figure 3. Bar plots of averaged RE scores.

After analyzing the parametric component, the estimation of the additive nonparametric components is presented in Tables 3 and 4. Table 3 displays the RMSE values computed for the individual functions, while Table 4 provides the ARMSE values for all simulation configurations, serving as a measure of the overall performance in estimating the nonparametric component of the right-censored PLAM. Upon initial examination, the LLR-KMW estimator demonstrates a significantly superior performance compared with the other two estimators across all simulation configurations. This dominance is further evidenced by the ARMSE results presented in Table 4, which contrast the outcomes observed in the parametric component estimation.

Table 3. RMSE values of individual nonparametric functions for both functions $f_1(t_1)$ and $f_2(t_2)$.

	Functions		$f_1(t_1)$			$f_2(t_2)$	
n	***CL***	**LLR-ST**	**LLR-KMW**	**LLR-kNNI**	**LLR-ST**	**LLR-KMW**	**LLR-kNNI**
50	5%	0.283	**0.256**	0.260	0.491	**0.473**	0.478
	20%	0.353	**0.241**	0.271	0.535	**0.433**	0.483
	35%	0.447	**0.256**	0.273	0.613	**0.406**	0.479
100	5%	0.383	**0.340**	0.364	0.689	**0.637**	0.668
	20%	0.408	**0.319**	0.366	0.704	**0.581**	0.657
	35%	0.466	**0.323**	0.371	0.754	**0.527**	0.655
200	5%	0.516	**0.483**	0.507	0.936	**0.896**	0.931
	20%	0.537	**0.438**	0.514	0.967	**0.800**	0.927
	35%	0.557	**0.452**	0.517	1.010	**0.727**	0.923

Bold color denotes the best performance score.

Table 4. $ARMSE\left(\hat{f}_1, \hat{f}_2\right)$ values for all simulation configurations.

n	***CL***	**LLR-ST**	**LLR-KMW**	**LLR-kNNI**	**CPH**
50	5%	0.281	**0.267**	0.271	0.872
	20%	0.319	**0.247**	0.275	0.967
	35%	0.374	**0.233**	0.276	1.008
100	5%	0.393	**0.362**	0.386	0.778
	20%	0.402	**0.334**	0.377	0.814
	35%	0.442	**0.310**	0.381	0.860
200	5%	0.544	0.519	0.539	0.775
	20%	0.565	**0.463**	0.541	0.784
	35%	0.583	**0.438**	0.538	0.841

Bold color denotes the best performance score.

An interesting distinction in estimating the nonparametric component is that the performances of the introduced estimators deteriorate as the sample size increases. To explain this phenomenon, it is crucial to note that in the estimation of PLAMs, there exists

a balance between the estimation of parametric and nonparametric components, which exhibits an inverse relationship. Furthermore, when data points are scattered widely around the representative smooth curve, the bias of the fitted curve increases. Additionally, the RMSE scores for the three modified LLR estimators are fairly similar to each other, confirming that the modified backfitting algorithm functions effectively with the censorship solution techniques.

Table 4 presents a strong case, confirming the dominant role of the LLR-KMW estimator in estimating nonparametric components within the context of right-censored PLAM. The success of the LLR-KMW estimator lies in its clever use of weighted estimation, which works well for both the parametric and nonparametric aspects of PLAM. Notably, the LLR-KMW estimator does not just improve $\boldsymbol{\beta}$ estimates, it also works well together with the LLR-kNNI estimator, forming a powerful estimation duo. When we carefully analyze Table 4 and take a close look at Figures 4 and 5, a clear pattern emerges. Both the LLR-KMW and LLR-kNNI estimators perform very similarly when it comes to estimating the nonparametric component. What is even more interesting is that both estimators outperform the LLR-ST estimator, as these enlightening visuals below beautifully demonstrate. In terms of estimating nonparametric components, it is naturally expected that the CPH estimator does not show a good performance due to its theoretical structure. However, its behaviors are similar to LLR-basis estimators in sample size and censoring level changes. In summary, the introduced LLR-basis estimators show better performance than the classical CPH estimator.

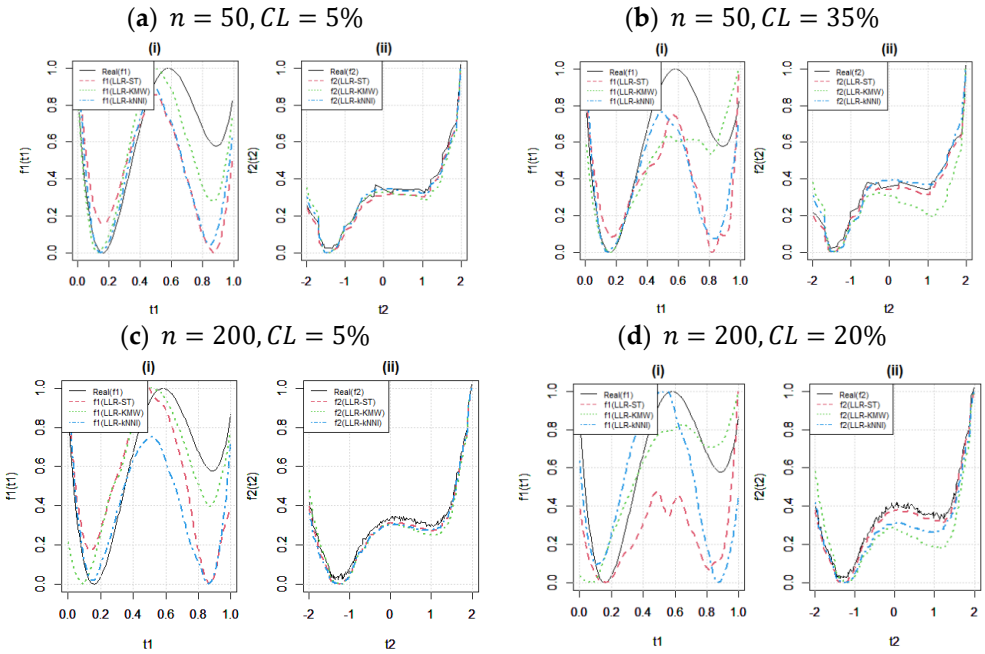

Figure 4. Fitted curves to show the effect of the censoring level (CL). In each panel, (i) and (ii) show fitted curves for $f_1(t_1)$ and $f_3(t_2)$ respectively.

Figure 4 illustrates the behavior of the estimators under different censoring levels with fixed sample sizes. In panels (a)–(b), the effect of the censoring level is investigated when the sample size is small ($n = 50$). It can be observed that while $f_2(t_2)$ is not significantly affected, the estimate of $f_1(t_1)$ is heavily influenced by the censored data points. It is important to note that this inference is also related to the initial values $\left(\boldsymbol{\beta}^{(0)}, \mathbf{f}^{(0)}\right)$ determined in the algorithm and their compatibility with the unknown functions f_1 and f_2, respectively (see [9] for further discussions). Furthermore, the results demonstrate that the weakness of the LLR-ST estimator (red dotted line) is clear in all four panels (a), (b), (c), and (d), for both $n = 50$ and $n = 200$. Additionally, panels (c) and (d) support the findings of Tables 3 and 4, leading to the conclusion that, for larger sample sizes, the fitted curves become more sensitive to the censoring level, resulting in a decrease in their performance.

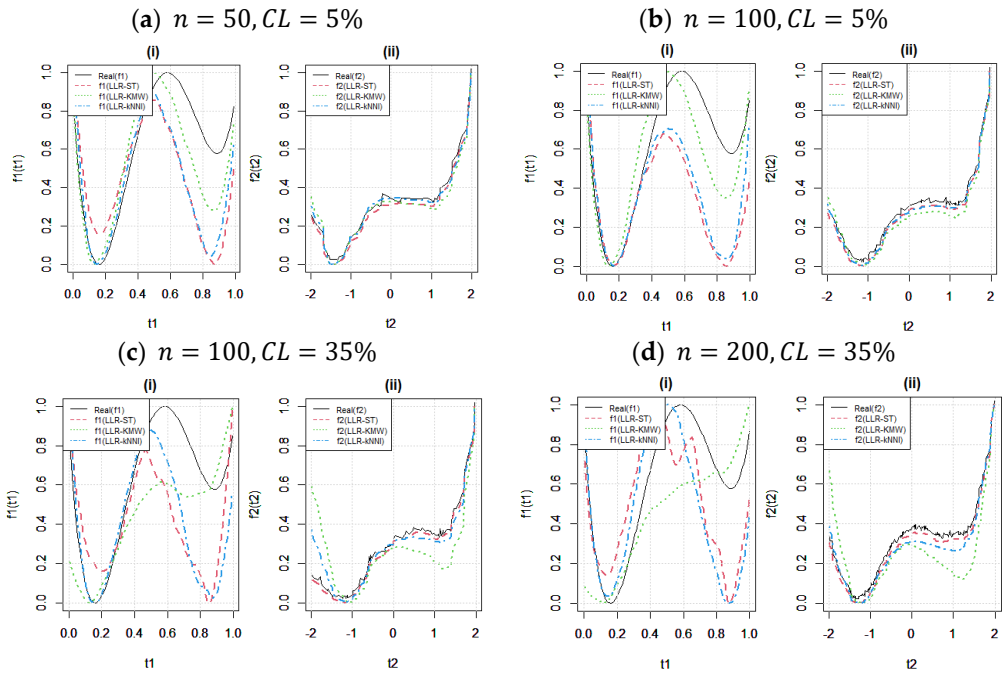

Figure 5. Fitted curves to show the effect of the sample size (n). In each panel, (i) and (ii) show fitted curves for $f_1(t_1)$ and $f_3(t_2)$ respectively.

Figure 5 investigates the effect of sample size (n) for fixed censoring levels in the upper and lower panels, particularly for $CL = 35\%$ in panels (c) and (d), while LLR-KMW and LLR-ST exhibit a slightly more pronounced response to increasing sample size compared with LLR-kNNI. This result is expected due to the nonparametric nature of kNNI. Furthermore, the changes observed in the fitted curves are more noticeable for the estimation of $f_1(t_1)$, as shown in Figure 4. Additionally, the differences between sample sizes for the lower censoring level ($CL = 5\%$) in panels (a)–(b) indicate that there is minimal variation between the fitted curves for both functions.

These trends are consistent with the findings reported by ref. [25], where a similar sensitivity of the ST basis estimator to sample size was identified in a related context. The reaction of the kNNI, KMW, and ST estimators to sample size fluctuations aligns with the observations made by ref. [26] reinforcing the notion that these estimators can exhibit greater flexibility in accommodating varying sample sizes.

To assess the performance of the introduced modified LLR estimators on real-world data and compare them with the simulation results, a real data example is presented in the following section, focusing on the hepatocellular carcinoma dataset.

6. Hepatocellular Carcinoma Data Example

In this section, the Hepatocellular Carcinoma dataset is modeled using the modified LLR estimators: LLR-ST, LLR-KMW, and LLR-kNNI. Their performances are compared with similar simulation configurations presented in Section 5. The dataset was originally presented by ref. [27] to investigate the gene expression of CXCL17 in hepatocellular carcinoma. Ref. [6] also studied this dataset, comparing parametric and semiparametric models on right-censored data. However, their study focused on a semiparametric model with a univariate nonparametric component using the covariate age. This paper considers a more realistic partially linear additive model (PLAM) that involves two nonparametric covariates.

The dataset consists of 227 data points and five explanatory variables: age, recurrence-free survival (RFS), CXCL17T (CXCT), CXCL17P (CXCP), and CXCL17N (CXCN). It should be noted that the logarithm of the response variable, overall survival time (OS), is used in this analysis. The parametric component of the PLAM is determined by the covariates

CXCL17T, CXCL17P, and CXCL17N. Additionally, *Age* and *RFS* are considered as nonparametric covariates due to their nonlinear structures, as depicted in Figure 6. The figure also illustrates the censored data points versus the transformed data points using the kNNI and ST solutions. Furthermore, panels (C) and (D) display hypothetical curves that represent the data structure and nonlinearity.

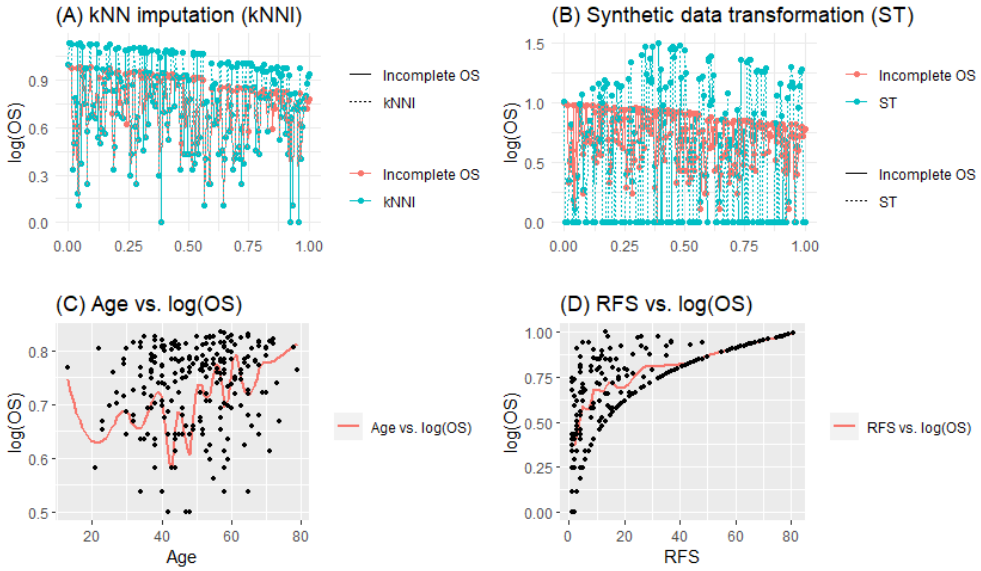

Figure 6. Descriptive plots for the Hepatocellular Carcinoma dataset.

The dataset contains 84 right-censored OS points, indicating a censoring level of $CL = 37\%$. This level of censorship can be classified as heavy censoring. Therefore, we expect that the results from the real data analysis may resemble the corresponding simulation configuration of $n = 200$ and $CL = 35\%$. Based on the information provided above, the partially linear additive model (PLAM) for the right-censored Hepatocellular Carcinoma dataset can be expressed as follows:

$$\log(OS_i) = \beta_0 + \beta_1 \text{CXCL17T}_\text{i} + \beta_2 \text{CXCL17P}_\text{i} + \beta_3 \text{CXCL17N}_\text{i} + f_1(Age_i) + f_2(RFS_i) + \varepsilon_i \tag{31}$$

where $i = 1, \ldots, 227$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ and $\mathbf{f} = (f_1, f_2)$. While estimating PLAM in (31), $\log(OS)$ is replaced by its ST version $\log(OS_{\hat{G}})$ and kNNI version $\log(OS_{imp})$. Also, KMW is applied. The outcomes of the Hepatocellular Carcinoma dataset with the modified LLR estimators are provided in Table 5.

Table 5. Performance scores of the introduced three estimators.

	LLR-ST	**LLR-KMW**	**LLR-kNNI**	**CPH**
$Bias(\beta_1;\beta_2;\beta_3)$	0.42;0.17;**0.08**	**0.30**;**0.16**;0.17	0.40;0.20;0.21	**0.24**;1.65;0.40
$Var(\beta_1;\beta_2;\beta_3)$	0.08;0.26;**0.05**	**0.05**;**0.24**;0.08	0.06;0.26;0.09	0.15;0.68;0.40
$SMDE$	0.220	**0.154**	0.256	1.341
$RMSE[f_1(Age)]$	**0.440**	0.533	0.491	-
$RMSE[f_2(RFS)]$	0.350	**0.168**	0.208	-
$ARMSE(f_1, f_2)$	0.395	**0.350**	0.350	1.822

Bold color denotes the best performance score.

Table 5 largely confirms the findings of the simulation study and demonstrates the superior performance of the LLR-KMW estimator in the estimation of the parametric component. However, in contrast to the simulation study, the LLR-ST estimator also provides results that are closer to the other two estimators, while the performance of LLR-kNNI is less satisfactory than expected. It should be noted that these conditions may be attributed to the relatively large sample size in terms of censored data. Additionally,

regarding the bias of β, as anticipated, both ST and KMW yield lower values compared with kNNI, as they theoretically promise less biased estimates. Overall, the performance evaluation in Table 6 confirms that LLR-KMW exhibits the best results, which are evident from the RE scores.

Table 6. Relative efficiencies; *REs*.

Estimator	LLR-ST	LLR-KMW	LLR-kNNI	CPH
LLR-ST	1.000	**0.699**	1.160	6.095
LLR-KMW	1.429	1.000	1.659	8.707
LLR-kNNI	**0.861**	**0.602**	1.000	5.238
CPH	**0.164**	**0.114**	**0.190**	1.000

Bold color denotes the best performance score.

In both Tables 5 and 6, the performance of benchmark CPH estimators is also provided and, as expected, it does not show a good performance, especially in the estimation of the nonparametric component. On the other hand, in terms of bias, Table 5 shows that CPH has satisfying bias values but with large variances that cause large SMDE scores. This poor performance is highly related to the lack of the ability of CPH to represent smooth functions. RE scores highly confirm this inference. Summing up the comprehensive assessment presented in Table 6, we encounter an unequivocal affirmation of the preeminent standing of the LLR-KMW estimator. This affirmation is elegantly illuminated by the notable RE scores, reflecting an ensemble of successful estimation endeavors.

In Figure 7, bar plots of the calculated relative efficiencies (RE) are presented. Consistent with the findings in Table 5, LLR-KMW exhibits lower RE scores compared with the other two estimators, which aligns with the results of the simulation study. It is worth noting that while the difference in performance between the estimators may appear significant, numerically they are relatively close to each other, with the RE values scattered around one.

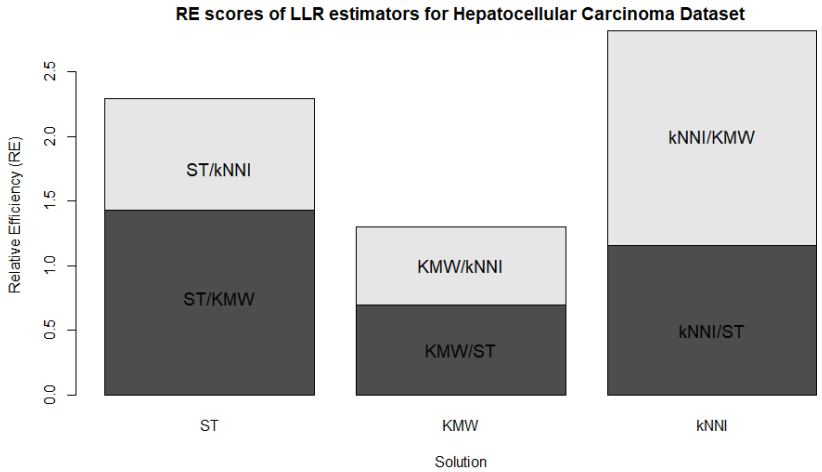

Figure 7. Bar plots of the REs for the modified LLR estimators based on the censorship solutions methods.

After assessing the estimation of the parametric component, Figure 8 presents the results of the estimation of the nonparametric components $f_1\ (Age)$ and $f_2\ (RFS)$. It is noteworthy that in this dataset, the relative failure of LLR-kNNI and the relative success of LLR-ST can be attributed to the structure of the nonparametric components. Both functions f_1 and f_2 exhibit favorable structures for the properties of LLR-ST, such as magnifying the magnitudes of uncensored data points and assigning zero to censored ones, as clearly observed in panel (ii) of Figure 8.

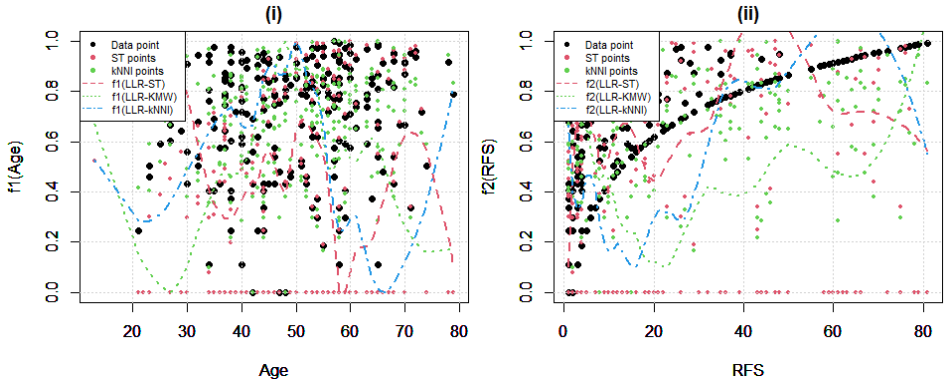

Figure 8. Fitted curves obtained for the Hepatocellular Carcinoma dataset. In panel (**i**) $f(Age)$ is shown and in panel (**ii**) involves $f(RFS)$.

To provide a more precise understanding of the solution procedures, the ST points and kNNI points are also included in the plots. These points illustrate why the fitted curves tend to lie below the region where all data points are scattered, especially in panel (ii). This is primarily influenced by the heavy censoring level, $CL = 37\%$. Additionally, in panel (i), one can observe the LLR-ST's fitted curve being pulled down by the zeros. As expected, LLR-KMW follows a balanced approach between the other two estimators, as shown in Table 5, yielding the smallest ARMSE scores in the estimation of the nonparametric component of the PLAM.

7. Conclusions

This paper introduces three modified LLR estimators based on different censorship solutions: ST, KMW, and kNNI, to model the right-censored PLAM. For the solution methods that have a theoretical background, such as ST and KMW, the statistical properties and some asymptotic properties of LLR-ST and LLR-KMW are presented. This paper focuses on two main objectives and successfully achieves them. The two purposes of this study are to combine the backfitting LLR estimator with the censorship solutions and to compare them, both theoretically and practically. The performances of the modified LLR estimators are observed through simulation and real data studies. The following conclusions have been drawn from this study:

- In the simulation study, the performance of the estimators is measured individually for both parametric and nonparametric components. Regarding the parametric component estimation, it is observed that LLR-KMW provides the best results, followed by LLR-kNNI. On the other hand, LLR-ST does not yield good results for any simulation configuration, and it is the estimator most affected by the censorship as its performance dramatically changes when the censoring level increases. In this case, LLR-KMW can be considered the most robust estimator, as it reacts to censorship in a more balanced way compared with the other two. In addition, the introduced estimators are also compared with the benchmark estimator for the survival model, CPH. It is observed that the LLR-basis estimators perform better than the CPH, as discussed in Section 6.
- In the estimation of the nonparametric components, the effects of sample size and censoring level are clearly different compared with the parametric component. However, similar to the parametric component, LLR-KMW exhibits dominant performance for both nonparametric functions. It is noteworthy that, as the sample size increases, all three estimators tend to provide closer performances in terms of fitted curves. Furthermore, it should be noted that the performance of the introduced estimators is highly dependent on the structure of the nonparametric component and its compatibility with the chosen censorship solution. Hence, this paper investigates the three different solutions in detail. Ultimately, because the CPH model lacks a smoother structural framework, it falls short when compared with the newly introduced estimators.

- The analysis of the Hepatocellular Carcinoma data serves as a real-world example in this study. This dataset is selected due to its censoring level and sample size, which align closely with one of the simulation configurations ($n = 200$ and $CL = 35\%$), enabling a more realistic comparison. The results of the real data modeling demonstrate that the three introduced modified LLR estimators effectively handle the estimation of the right-censored PLAM for both parametric and nonparametric components. They exhibit a good level of agreement with the corresponding simulation configuration, with some minor differences. As expected, LLR-KMW yields the best results. Also, CPH does not show a good performance except in the bias of regression coefficients, as observed in the simulation study. Notably, one important difference between the real data and the simulation study is that LLR-ST exhibits a surprisingly better performance than LLR-kNNI in the estimation of both parametric and nonparametric components. However, this discrepancy can be attributed to the relatively large sample size ($n = 227$), and it does not imply inconsistency with the simulation results. On the contrary, it indicates a close agreement among all performances.

Author Contributions: Conceptualization: S.E.A. and D.A.; Methodology: E.Y. and D.A.; Formal analysis and investigation: D.A. and E.Y.; Writing—original draft preparation: E.Y.; Writing—review and editing: S.E.A. and E.Y.; Data Curation: E.Y.; Visualization: E.Y.; Software: E.Y.; Supervision: S.E.A. and D.A.; Funding acquisition: S.E.A. and D.A.; Resources: S.E.A. and D.A.; Supervision: S.E.A. and D.A. All authors have read and agreed to the published version of the manuscript.

Funding: The research of Dursun Aydın was supported by the TUBITAK 1002 project with the project number: 122F045.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The Hepatocellular Carcinoma dataset is publicly available in R-package named "asaur".

Acknowledgments: The research of S. Ejaz Ahmed was supported by the Natural Sciences and the Engineering Research Council (NSERC) of Canada.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ruppert, D.; Wand, M.P.; Carroll, R.J. *Semiparametric Regression (No. 12)*; Cambridge University Press: Cambridge, UK, 2003.
2. Zhang, H.H.; Cheng, G.; Liu, Y. Linear or nonlinear? Automatic structure discovery for partially linear models. *J. Am. Stat. Assoc.* **2011**, *106*, 1099–1112. [CrossRef]
3. Hamilton, S.A.; Truong, Y.K. Local linear estimation in partly linear models. *J. Multivar. Anal.* **1997**, *60*, 1–19. [CrossRef]
4. Aydin, D.; Yilmaz, E. Modified estimators in semiparametric regression models with right-censored data. *J. Stat. Comput. Simul.* **2018**, *88*, 1470–1498. [CrossRef]
5. Orbe, J.; Virto, J. Penalized spline smoothing using Kaplan-Meier weights in semiparametric censored regression models. *Stat. Oper. Res. Trans.* **2022**, *46*, 95–114.
6. Yenilmez, I.; Yılmaz, E.; Kantar, Y.M.; Aydın, D. Comparison of parametric and semi-parametric models with randomly right-censored data by weighted estimators: Two applications in colon cancer and hepatocellular carcinoma datasets. *Stat. Methods Med. Res.* **2022**, *31*, 372–387. [CrossRef]
7. Opsomer, J.D.; Ruppert, D.; Wand, M.P.; Holst, U.; Hössjer, O. Kriging with nonparametric variance function estimation. *Biometrics* **1999**, *55*, 704–710. [CrossRef]
8. Ichimura, H.; Lee, S. Characterization of the asymptotic distribution of semiparametric M-estimators. *J. Econom.* **2010**, *159*, 252–266. [CrossRef]
9. Ahmed, S.E.; Aydın, D.; Yılmaz, E. A survey of smoothing techniques based on a backfitting algorithm in estimation of semiparametric additive models. *Wiley Interdiscip. Rev. Comput. Stat.* **2023**, *15*, e1605. [CrossRef]
10. Stute, W. Nonlinear censored regression. *Stat. Sin.* **1999**, *9*, 1089–1102.
11. Aydın, D.; Ahmed, S.E.; Yılmaz, E. Estimation of semiparametric regression model with right-censored high-dimensional data. *J. Stat. Comput. Simul.* **2019**, *89*, 985–1004. [CrossRef]
12. Koul, H.; Susarla, V.; Van Ryzin, J. Regression analysis with randomly right-censored data. *Ann. Stat.* **1981**, *9*, 1276–1288. [CrossRef]
13. Stute, W. Consistent estimation under random censorship when covariables are present. *J. Multivar. Anal.* **1993**, *45*, 89–103. [CrossRef]

14. Zhang, S. Nearest neighbor selection for iteratively kNN imputation. *J. Syst. Softw.* **2012**, *85*, 2541–2552. [CrossRef]
15. Ahmed, S.E.; Aydin, D.; Yılmaz, E. Nonparametric regression estimates based on imputation techniques for right-censored data. In *International Conference on Management Science and Engineering Management*; Springer International Publishing: Cham, Switzerland, 2019; pp. 109–120.
16. Cartwright, M.H.; Shepperd, M.J.; Song, Q. Dealing with missing software project data. In Proceedings of the 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (IEEE Cat. No. 03EX717), Sydney, Australia, 5 September 2004; IEEE: Piscataway, NJ, USA, 2004; pp. 154–165.
17. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; CRC Press: Boca Raton, FL, USA, 1990; Volume 43.
18. Härdle, W.; Müller, M.; Sperlich, S.; Werwatz, A. *Nonparametric and Semiparametric Models*; Springer: Berlin, Germany, 2004; Volume 1.
19. Buja, A.; Hastie, T.; Tibshirani, R. Linear smoothers and additive models. *Ann. Stat.* **1989**, *17*, 453–510. [CrossRef]
20. Opsomer, J.D.; Ruppert, D. A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comput. Graph. Stat.* **1999**, *8*, 715–732. [CrossRef]
21. Wei, C.H.; Liu, C. Statistical inference on semi-parametric partial linear additive models. *J. Nonparametr. Stat.* **2012**, *24*, 809–823. [CrossRef]
22. Kauermann, G.; Opsomer, J.D. Generalized cross-validation for bandwidth selection of backfitting estimates in generalized additive models. *J. Comput. Graph. Stat.* **2004**, *13*, 66–89. [CrossRef]
23. Chu, C.K. Bandwidth selection in nonparametric regression with general errors. *J. Stat. Plan. Inference* **1995**, *44*, 265–275. [CrossRef]
24. Hanley, J.A.; Parnes, M.N. Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics* **1983**, *39*, 129–139. [CrossRef]
25. Wang, Q.; Dinse, G.E. Linear regression analysis of survival data with missing censoring indicators. *Lifetime Data Anal.* **2011**, *17*, 256–279. [CrossRef]
26. Aydin, D.; Yilmaz, E. Semiparametric regression estimates based on some transformation techniques for right-censored data. *Eskişehir Tech. Univ. J. Sci. Technol. A—Appl. Sci. Eng.* **2019**, *20*, 1–12. [CrossRef]
27. Li, L.; Yan, J.; Xu, J.; Liu, C.-Q.; Zhen, Z.-J.; Chen, H.-W.; Ji, Y.; Wu, Z.-P.; Hu, J.-Y.; Zheng, L.; et al. CXCL17 expression predicts poor prognosis and correlates with adverse immune infiltration in hepatocellular carcinoma. *PLoS ONE* **2014**, *9*, e110064. [CrossRef] [PubMed]