

Article

Penalty and Shrinkage Strategies Based on Local Polynomials for Right-Censored Partially Linear Regression

Syed Ejaz Ahmed ¹, Dursun Aydın ²  and Ersin Yılmaz ^{2,*}

¹ Department of Mathematics and Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada

² Department of Statistics, Mugla Sıtkı Kocman University, 48000 Mugla, Turkey

* Correspondence: ersinyilmaz@mu.edu.tr

Abstract: This study aims to propose modified semiparametric estimators based on six different penalty and shrinkage strategies for the estimation of a right-censored semiparametric regression model. In this context, the methods used to obtain the estimators are ridge, lasso, adaptive lasso, SCAD, MCP, and elasticnet penalty functions. The most important contribution that distinguishes this article from its peers is that it uses the local polynomial method as a smoothing method. The theoretical estimation procedures for the obtained estimators are explained. In addition, a simulation study is performed to see the behavior of the estimators and make a detailed comparison, and hepatocellular carcinoma data are estimated as a real data example. As a result of the study, the estimators based on adaptive lasso and SCAD were more resistant to censorship and outperformed the other four estimators.

Keywords: local polynomial regression; lasso; elasticnet; SCAD; MCP; partially linear model; right-censored data



Citation: Ahmed, S.E.; Aydın, D.; Yılmaz, E. Penalty and Shrinkage Strategies Based on Local Polynomials for Right-Censored Partially Linear Regression. *Entropy* **2022**, *24*, 1833. <https://doi.org/10.3390/e24121833>

Academic Editor: Antonio M. Scarfone

Received: 7 November 2022

Accepted: 13 December 2022

Published: 15 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Consider the partially linear (or semiparametric) regression model

$$z_i = \mathbf{x}_i' \boldsymbol{\beta} + f(t_i) + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where z_i 's are the observations of the response variable, $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})$ is known k -dimensional vectors of explanatory variables, $t_i \in [a, b]$ is the value of an extra explanatory variable t , $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ is an unknown k -dimensional parameter vector to be estimated, $f(\cdot)$ is an unknown univariate smooth function, and ε_i 's are supposed to be uncorrelated independent random variables with mean zero and finite variance $\sigma_\varepsilon^2 = E(\varepsilon^2)$. Partially linear models through a nonparametric component are flexible enough to cover many situations; in fact, these models may be an appropriate choice when it is suspected that the response variable z is linearly dependent on \mathbf{x} , indicating parametric effects, but nonlinearly related to t_i denoting nonparametric effects. Note that model (1) can be expressed in matrix and vector form as

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{Z} = (z_1, \dots, z_n)'$, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]'$ is an $(n \times k)$ design matrix with $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})'$ denoting the i th k -dimensional row vector of \mathbf{X} , $\mathbf{f} = (f(t_1), \dots, f(t_n))'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ is a random error vector with $E(\boldsymbol{\varepsilon}) = 0$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$. For more discussions on model (1), see [1–3], among others.

In this paper, we are interested in estimating the parametric and nonparametric components of model (1) when the observations of the response variable are incompletely observed and right-censored by the random censoring variable c_i , $i = 1, 2, \dots, n$, but \mathbf{x}_i and t_i are completely observed. In the case where z_i 's are the censored from the right, then any estimation procedure cannot be applied to z_i due to censoring. To add the effect of the

ensorship to the model estimation process, it should be revealed with the help of auxiliary variables that the censorship problem should be solved accordingly. Therefore, instead of observing the values of the response variable z_i , we observe the dataset (y_i, δ_i) with

$$y_i = \min(z_i, c_i), \delta_i = I(y_i < c_i), i = 1, 2, \dots, n \quad (3)$$

where y_i 's are the observations of the updated new response variable according to censorship and δ_i 's are the values of the censor indicator related to y_i 's. If the i th observation of z_i is censored, we choose $y_i = c_i$ and $\delta_i = 0$; otherwise, we consider $y_i = z_i$ and $\delta_i = 1$. In this case, model (1) transforms into a semiparametric model with the right-censored data, which can also be updated in terms of the values of the new response variable.

In the literature, there are several studies about the right-censored linear model ($f(t) = 0$ in model (1)), including [4–7]. The right-censored nonparametric regression model ($\beta = 0$ in model (1)) has been studied by [8,9], among others. In addition, right-censored partially linear models have been studied by [10], who used smoothing splines based on Kaplan–Meier weights as an estimation procedure, and [11] considered censored partial linear models and illustrated the theoretical properties of the semiparametric estimators based on the synthetic data transformation. Aydin and Yilmaz [12] suggested three semiparametric estimators based on regression splines, kernel smoothing, and smoothing spline methods using synthetic data transformation. Regarding the partially linear models with penalty functions, in the case of the noncensored data, [13] studied the estimation of the semiparametric model based on the two absolute penalty functions, which are lasso and adaptive lasso with B-splines. Moreover, they conducted a technical analysis of the estimators meticulously with asymptotic properties.

This paper considers model (1) under a right-censored response variable and a large number of covariates in the parametric component. Notice that right-censored data cause biased estimates due to incomplete observations that manipulate the data structure. Therefore, if the censorship is ignored, inferences based on estimated models may be wrong or deviated. For instance, in clinical trials, some of the observed patients may withdraw from the study before it ends, or they may die from another reason, which makes the corresponding observation right-censored. In particular, in medical studies as in the given example, preventing information loss and obtaining less biased estimates are quite important. This paper, therefore, solves both variable selection and censorship problems. To achieve the variable selection, six different penalty functions are considered: ridge, lasso, adaptive lasso, SCAD, MCP, and elasticnet. Notice that a detailed study about penalty functions and shrinkage techniques is provided by [14]. Local polynomial regression is used as the smoothing method. Finally, the censorship problem is solved using the synthetic data transformation proposed by [6].

In the light of the information given, the main difference and most important contribution of this article from previous studies is that it proposes quasi-parametric estimators based on six different penalty functions with the local polynomial technique for the right-censored model (1). To the best of our knowledge, this kind of detailed study has not yet been made in the literature.

The paper is organized as follows: Section 2 introduces the right-censored data phenomenon and some preliminaries. Section 3 explains the local polynomial smoothing method, and the modified semiparametric estimators are introduced based on the six penalty functions with theoretical properties. In Section 4, the evaluation metrics are shown. Section 5 performs a Monte Carlo simulation study, and the results are presented. Section 6 presents an analysis of hepatocellular carcinoma data as a real data example. Finally, conclusions are given in Section 7.

2. Preliminaries and General Estimation Procedure

Let H , F , and G be distribution functions of the variables y_i , z_i and c_i , respectively. More precisely, let the probability distribution functions of these variables be

$$H(u) = P(y_i \leq u), F(u) = P(z_i \leq u), \text{ and } G(u) = P(c_i \leq u) \text{ for } u \in \mathbb{R}$$

and their corresponding survival functions are given by

$$\bar{H}(u) = 1 - H(u) = P(y_i > u), \bar{F}(u) = P(z_i > u), \text{ and } \bar{G}(u) = P(c_i > u).$$

The key idea here is to examine the effect of the explanatory variables on the response variable by estimating the expected value of $E(z|x, t)$ by the regression function. In the setting of a semiparametric regression problem, first, we need to make some identification conditions on the response variable, censoring, and explanatory variables and their dependence relationships. In other words, we take some assumptions to ensure that the model is identifiable.

Assumption 1. (i) z_i and c_i are conditionally independent given (x_i, t_i) ; (ii) $P(z_i \leq c_i | x_i, t_i, z_i) = P(z_i \leq c_i | z_i)$.

It should be emphasized that Assumption 1(i) and Assumption 1(ii) are commonly accepted assumptions regarding right-censored models and survival analysis (see [15,16]). Assumption 1(i) is an independency condition that provides identifiability for the model. Assumption 1(ii) indicates that covariates provide the same information about the response variable independent of the existence of censorship (see [17]).

Because of the censoring, the classical methods for estimating the parametric and nonparametric components of model (3) are inapplicable. The most important reason for this is that the censored observations z_i and updated random observations y_i have different expectations. This problem can be overcome by using so-called synthetic data, as in censored linear models. We refer, for example, to the studies of [6,12], among others. In this context, when the distribution G is known, we use synthetic data transformation

$$y_{iG}^* = \delta_i y_i \{1 - G(y_i)\}^{-1} = \delta_i y_i \{\bar{G}(y_i)\}^{-1} \quad (4)$$

where $\bar{G}(\cdot) = 1 - G(\cdot)$ and $G(\cdot)$ denotes the distribution functions of censoring variables C , as defined in the introduction to this section. The nature of the synthetic data method ensures that (Y_{iG}^*, X_i, W_i) , $i = 1, 2, \dots, n$ are independent random variables with $E(y_{iG}^* | x_i, t_i) = E(z_i | x_i, t_i)$, as described in Lemma 1.

Lemma 1. If, instead of response observations z_i , only $\{(y_i, \delta_i)\}_{i=1}^n$ is observed in the context of a semiparametric regression model and the censoring distribution G is known, then the regression function (or mean vector) $\mu = \mathbf{x}'\boldsymbol{\beta} + f(t_i)$ is a conditional expectation; that is, $E(y_{iG}^* | x_i, t_i) = E(z_i | x_i, t_i) = \mu$.

A proof of Lemma 1 is given in Appendix A.1.

Lemma 1 cannot be directly applied to the estimation $f(\cdot)$ if the distribution G is unknown. To overcome this difficulty, [6] recommends replacing G with its Kaplan–Meier estimator [18]:

$$\hat{G}(t) = 1 - \prod_{i=1}^n \left(\frac{n-i}{n-i+1} \right)^{I[y_{(i)} \leq t, \delta_{(i)}=0]}, \quad t \geq 0 \quad (5)$$

where $y_{(1)} \leq \dots \leq y_{(n)}$ are the order statistics of y_1, \dots, y_n and $\delta_{(i)}$ is the corresponding indicator related to $y_{(i)}$, as defined in previous sections. In this case, that is, when the distribution G is unknown, we consider the following synthetic data transformation:

$$y_{i\hat{G}}^* = \delta_i y_i \{1 - \hat{G}(y_i)\}^{-1}, \quad i = 1, 2, \dots, n \quad (6)$$

3. Local Polynomial Estimator

Consider the semiparametric regression model defined in (1). Here, we approximate the regression function $f(t_i)$ locally by a polynomial of order p (see [19]). Using a Taylor series expansion in t_i at a neighborhood of fixed t_{i0} , the p .th degree polynomial approximation of $f(t_i)$ yields

$$f(t_i) \approx \sum_{j=0}^p \frac{f^{(j)}(t_{i0})}{j!} (t_i - t_{i0})^j = \sum_{j=0}^p b_j (t_i - t_{i0})^j \tag{7}$$

Note that fixed t_{i0} is determined in the range $t_{i0} \in [t_i - \delta, t_i + \delta]$ for a small real-valued δ and used to estimate t_i locally (see [8] for details). The idea is to estimate the components of a semiparametric model, leading to the minimization of the local weighted least squares criterion:

$$\min_{b, \beta} \sum_{i=1}^n \left\{ y_{i\hat{G}}^* - \sum_{j=0}^p (t_i - t_{i0})^j b_j - \mathbf{x}'_i \beta \right\}^2 K\left(\frac{t_i - t_{i0}}{h}\right) \tag{8}$$

where $y_{i\hat{G}}^*$'s are the values of the synthetic variable, as defined in (6), $K(\cdot)$ is a kernel function assigning weights to each point, and h is the bandwidth parameter controlling the size of the local neighborhood of t_{i0} . Additionally, note that vector and matrix notation (8) can be written as follows:

$$\min_{b, \beta} (\mathbf{y}_{\hat{G}}^* - \mathbf{T}\mathbf{b} - \mathbf{X}\beta)' \mathbf{W}_n (\mathbf{y}_{\hat{G}}^* - \mathbf{T}\mathbf{b} - \mathbf{X}\beta) \tag{9}$$

where $\mathbf{y} = (y_1, \dots, y_n)'$, $\mathbf{b} = (b_0, \dots, b_p)'$, $\mathbf{W}_n = \text{diag}\left(K\left(\frac{t_i - t_{i0}}{h}\right)\right)$ is a $n \times n$ weights matrix whose properties are provided in Assumption 4. Note that the minimum problem (9) has a unique solution based on the following matrices:

$$\mathbf{T} = \begin{bmatrix} 1 & (t_1 - t_{i0}) & \dots & (t_1 - t_{i0})^p \\ 1 & (t_2 - t_{i0}) & \dots & (t_2 - t_{i0})^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (t_n - t_{i0}) & \dots & (t_n - t_{i0})^p \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

For technical convenience, we assume that β is known to be the true parameter. Then the solution to minimizing (9) is

$$\hat{\mathbf{b}} = (\mathbf{T}'\mathbf{W}_n\mathbf{T})^{-1} \mathbf{T}'\mathbf{W}_n(\mathbf{y}_{\hat{G}}^* - \mathbf{X}\beta) \tag{10}$$

It can be seen from the Taylor series expansion given in (7) that one needs to select the first element of the vector $\hat{\mathbf{b}} = (\hat{b}_0, \dots, \hat{b}_k)$ in order to obtain $\hat{f}(t_0) = \hat{b}_0$. Then, for the fixed neighborhood t_0 , the deconvoluted local polynomial estimator of the regression function can be written as

$$\begin{aligned} \hat{f}(t_0; h) &= \sum_{i=1}^n \omega_1'(t_i'W_{ni}t_i)^{-1} t_i'W_{ni}(\mathbf{y}_{i\hat{G}}^* - \mathbf{x}'_i\beta) \\ &= \omega_1'(\mathbf{T}'\mathbf{W}_n\mathbf{T})^{-1} \mathbf{T}'\mathbf{W}_n(\mathbf{y} - \mathbf{X}\beta) = \mathbf{S}_h(\mathbf{y}_{\hat{G}}^* - \mathbf{X}\beta) \end{aligned} \tag{11}$$

where $\mathbf{S}_h = \omega_1'(\mathbf{T}'\mathbf{W}_n\mathbf{T})^{-1} \mathbf{T}'\mathbf{W}_n$ denotes the deconvoluted local polynomial smoother matrix, $\omega_1' = (1, t, \dots, t^p) \in \mathbb{R}^{(p+1)}$ dimensional matrix having 1 in the first position and 0 otherwise, and the matrices \mathbf{T} and \mathbf{W}_n are as defined above.

After the theoretical confirmation by giving Equations (10) and (11), the cases of both model parameters (β, \mathbf{f}) are unknown. To obtain the local polynomial-based estimates

$(\hat{\beta}_L, \hat{f}_L)$, the smoother matrix S_h given right after Equation (11) is used to calculate the following partial residuals in matrix form:

$$\tilde{X} = (I_n - S)X = x_i - \sum_{i=1}^n \omega_i'(t_i'W_{ni}t_i)^{-1}t_i'W_{ni}x_i = \tilde{x}_i \tag{12}$$

and

$$\tilde{y}_G^* = (I_n - S)y_G^* = y_{iG}^* - \sum_{i=1}^n \omega_i'(t_i'W_{ni}t_i)^{-1}t_i'W_{ni}y_{iG}^* = \tilde{y}_{iG}^* \tag{13}$$

where

$$S = \begin{pmatrix} (S_h)_{t_1} \\ \vdots \\ (S_h)_{t_n} \end{pmatrix} = \begin{pmatrix} (1\ 0\ \dots\ 0)(T'_{t_1}W_{t_1}T_{t_1})^{-1}T'_{t_1}W_{t_1} \\ \vdots \\ (1\ 0\ \dots\ 0)(T'_{t_n}W_{t_n}T_{t_n})^{-1}T'_{t_n}W_{t_n} \end{pmatrix}$$

Thus, we obtain a transformed set of data based on local residuals. Considering these partial residuals for the vector β yields the following least squares instead of criterion (9):

$$\min_{\beta} \sum_{i=1}^n \{ \tilde{y}_{iG}^* - \tilde{x}'_i\beta \}^2 = \|(\tilde{y}_G^* - \tilde{X}\beta)\|^2 \tag{14}$$

where \tilde{x}'_i is the i .th row of the matrix \tilde{X} . Under Assumptions 2–4, by applying the least squares technique to (14), we obtain a “modified local polynomial estimator” $\hat{\beta}_L$ for the parametric part of the semiparametric model (3), given by

$$\hat{\beta}_L = (\tilde{X}\tilde{X})^{-1}\tilde{X}'\tilde{y}_G^* \tag{15}$$

Correspondingly, a “modified local polynomial estimator” \hat{f}_L of the function $f(\cdot)$ for the nonparametric part in the semiparametric model (3) is defined as

$$\hat{f}_L = S(y_G^* - X\hat{\beta}_L) \tag{16}$$

The implementation details of the Equations (15) and (16) are given in Appendix A2. We conclude this section with the following assumptions necessary to obtain the main results. These assumptions are quite general and easily fulfilled.

Assumption 2. When the covariates (x_{ij}, t_i) are fixed design points, there exist continuous functions $h_j(\cdot)$ defined on $[0, 1]$ such that each component of x_i satisfies

$$x_{ij} = h_j(t_i) + v_{ij}, 1 \leq i \leq n, 1 \leq j \leq k$$

where $\{v_{ij}\}$ is a sequence of real numbers satisfying

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n v_{is}v'_{im} = c_{sm}, 1 \leq s \leq k, 1 \leq m \leq k$$

and $C = (c_{sm})$ is a $(k \times k)$ dimensional nonsingular matrix.

Assumption 3. The functions $f(\cdot)$ and $h_j(\cdot)$ are Lipschitz continuous of order 1 for $j = 1, \dots, k$.

Note that Assumption 2 generalizes the conditions of [20,21], where (x_{ij}, t_i) are fixed design points for a partially linear model with uncensored data. Assumption 3 is required to establish asymptotic normality with an observed value t_i .

Assumption 4. The weight functions W_n satisfy these conditions:

- (i.) $\max_{1 \leq i \leq n} \sum_{j=1}^n W_{ni}(t_j) = O(1)$
- (ii.) $\max_{1 \leq i, j \leq n} \sum_{j=1}^n W_{ni}(t_j) = O(n^{-2/3})$
- (iii.) $\max_{1 \leq i \leq n} \sum_{j=1}^n W_{ni}(t_j) I(|t_i - t_j| > a_n) = O(b_n),$

where $I(\cdot)$ is an indicator function, a_n satisfies $\limsup_{n \rightarrow \infty} n a_n^3 < \infty$, and b_n satisfies $\limsup_{n \rightarrow \infty} n b_n^3 < \infty$.

3.1. Ridge-Type Local Polynomial Estimator

In this paper, we confine ourselves to the local polynomial estimators of the vector parameter β and the unknown smooth function $f(\cdot)$ in a semiparametric model. For a given bandwidth parameter h , the corresponding estimators $\hat{\beta}$ and \hat{f} based on model (2) are described by (14) and (15), respectively. Multiplying both sides of model (2) by $(\mathbf{I}_n - \mathbf{S})$, we obtain

$$\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\beta + \tilde{\varepsilon} \quad (17)$$

where $\tilde{\mathbf{Z}} = (\mathbf{I}_n - \mathbf{S})\mathbf{Z}$, $\tilde{\varepsilon} = \tilde{\mathbf{f}} + \varepsilon^*$, $\tilde{\mathbf{f}} = (\mathbf{I}_n - \mathbf{S})\mathbf{f}$, and $\varepsilon^* = (\mathbf{I}_n - \mathbf{S})\varepsilon$, similar to (12) and (13).

This consideration turns model (17) into an optimization problem to obtain the estimator of the vector β corresponding parametric part of the semiparametric model in (2). In this context, this model leads to the following penalized least squares (PLS) criterion for the ridge regression problem:

$$PLS_{RL} = \arg \min_{\beta} \left(\tilde{\mathbf{Z}} - \tilde{\mathbf{X}}\beta \right)' \left(\tilde{\mathbf{Z}} - \tilde{\mathbf{X}}\beta \right) + \lambda \beta' \beta \quad (18)$$

where λ is a positive shrinkage parameter that controls the magnitude of the penalty. The solution to this minimization problem (17) provides the following Theorem 1.

Theorem 1. Ridge-type local polynomial estimator for β is presented by $\hat{\beta}_{RL}$ and is expressed based on the local polynomial smoothing matrix \mathbf{S} by

$$\hat{\beta}_{RL}(\lambda) = \left(\tilde{\mathbf{X}}' \tilde{\mathbf{X}} + \lambda \mathbf{I}_k \right)^{-1} \tilde{\mathbf{X}}' \tilde{\mathbf{y}}_{\hat{G}}^* \quad (19a)$$

where $\tilde{\mathbf{y}}_{\hat{G}}^*$ is a vector of updated response observations, as defined in Equations (6) and (13).

A proof of Theorem 1 is given in Appendix A.3.

As shown in Theorem 1, when $\lambda = 0$, the ridge-type local polynomial estimate reduces to an ordinary least squares estimate problem based on the local residuals defined in Equations (12) and (13). It should be noted that in order to estimate the unknown function \mathbf{f} , we imitate Equation (16) and define

$$\hat{\mathbf{f}}_{RL} = \mathbf{S} \left(\mathbf{y}_{\hat{G}}^* - \mathbf{X} \hat{\beta}_{RL}(\lambda) \right) \quad (19b)$$

Thus, the estimator (19b) is stated as the ridge-type local polynomial estimator of the unknown function \mathbf{f} in the semiparametric model (1.2).

3.2. Penalty Estimation Strategies Based on Local Polynomial

Several penalty functions are discussed for linear and generalized regression models in the literature (see [22]). In this paper, we study the minimax concave penalty (MCP), the least absolute shrinkage and selection operation (lasso), the smoothly clipped absolute deviation method (SCAD), the adaptive lasso, and the elasticnet method, which is a regu-

larized regression technique that linearly combines the L_1 and L_2 penalties of the lasso and the ridge regression methods, respectively.

In this paper, we suggest local polynomial estimators based on different penalties for the components of the semiparametric regression model. For a given penalty function $P_\lambda(\beta)$ and tuning parameter λ , the general form of the penalized least squares (PLS_G) of penalty estimators can be expressed as

$$PLS_G = \arg \min_{\beta} \left\{ \sum_{i=1}^n (\tilde{y}_{i\hat{G}}^* - \tilde{x}_i' \beta)^2 \right\} + P_\lambda(\beta) = (\tilde{\mathbf{y}}_G^* - \tilde{\mathbf{X}}\beta)' (\tilde{\mathbf{y}}_G^* - \tilde{\mathbf{X}}\beta) + P_\lambda(\beta) \quad (20)$$

Note that the vector $\hat{\beta}$ that minimizes (20) for lasso and ridge penalties is known as a bridge estimator, proposed by [23]. On the other hand, elasticnet, SCAD, MCP, and adaptive lasso involve different penalties, which are inspected in the remainder of this paper. It should be emphasized that in the mentioned four penalty functions, in the penalty term $P_\lambda(\beta) = \lambda \sum \|\beta_j\|_q^q$ satisfies the L_q norm of the regression coefficients β_j (see [24–26]). Thus, the different penalty estimators corresponding to the parametric and nonparametric components of the semiparametric model can be defined for different values of degree q and shrinkage parameter λ .

3.2.1. Estimation Procedure for the Parametric Component

From (20), we see that for $q = 2$, ridge estimates corresponding to the parametric component can be obtained by minimizing the following penalized residual sum of squares

$$\hat{\beta}_{RL} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (\tilde{y}_{i\hat{G}}^* - \tilde{x}_i' \beta)^2 + \lambda \sum_{j=1}^k |\beta_j|^2 \right\} \quad (21)$$

where $\tilde{y}_{i\hat{G}}^*$ is the i th synthetic observation of $\tilde{\mathbf{y}}_G^*$ and \tilde{x}_i' is the i .th row of the matrix $\tilde{\mathbf{X}}$. Notice that the solution (21) has the same regularization estimate stated in (19a). It should also be noted that when $q = 1$ in (20), we obtain the estimator known as the lasso.

Lasso: Proposed by [24], lasso, a penalized least squares method, is a regularization method for simultaneous estimation and variable selection that estimates with the L_1 penalty. The modified local polynomial estimators based on the lasso penalty can be defined as

$$\hat{\beta}_{LL} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (\tilde{y}_{i\hat{G}}^* - \tilde{x}_i' \beta)^2 + \lambda \sum_{j=1}^k |\beta_j| \right\} \quad (22)$$

Although Equation (22) may seem subtle, the absolute penalty term makes it impossible to find an analytical solution for the lasso. Initially, lasso solutions are obtained through quadratic programming.

Adaptive lasso: Zou [25] suggested modifying the lasso penalty by using adaptive weights on L_1 penalties on the regression coefficients. This weighted lasso, which has oracle properties, is referred to as the adaptive lasso. The local polynomial estimator $\hat{\beta}_{aLL}$ using the adaptive lasso penalty can be defined as follows:

$$\hat{\beta}_{aLL} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (\tilde{y}_{i\hat{G}}^* - \tilde{x}_i' \beta)^2 + \lambda \sum_{j=1}^k \hat{w}_j |\beta_j| \right\} \quad (23)$$

where w is a weight function given by

$$\hat{w}_j = \frac{1}{|\hat{\beta}_j^*|^q}, q > 0$$

It should be noted that $\hat{\beta}^*$ is an appropriate estimator of β here. For example, an ordinary least squares (OLS) estimate can be used as a reference value. To obtain the adaptive lasso estimates in (23), it is necessary to choose $q > 0$ and compute the weights after obtaining the OLS estimate.

SCAD: A disadvantage of the lasso method is that the penalty term is linear in the size of the regression coefficient, so it tends to give highly biased estimates for large regression coefficients. To account for this bias, [26] proposed a SCAD penalty obtained by replacing $|\beta_j|$ in (22) with $P_{\alpha,\lambda}|\beta_j|$. A modified local estimator $\hat{\beta}_{SL}$ based on the SCAD penalty can be described as

$$\hat{\beta}_{SL} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (\tilde{y}_{i\hat{G}}^* - \tilde{x}_i' \beta)^2 + \sum_{j=1}^k P_{\alpha,\lambda} |\beta_j| \right\} \tag{24a}$$

where $P_{\alpha,\lambda}(\cdot)$ is the SCAD penalty defined by

$$P_{\alpha,\lambda} = \lambda \left\{ \mathbf{I}(|\beta|) \leq \lambda + \frac{(\alpha\lambda - |\beta|)_+}{(\alpha - 1)\lambda} \mathbf{I}(|\beta| > \lambda) \right\}, \text{ and for } \lambda \geq 0 \tag{24b}$$

It should be stated that here $\lambda > 0$ and $\alpha > 2$ are the penalty parameters, $\mathbf{I}(\cdot)$ is the indicator function, and $(t)_+ = \max(t, 0)$. In addition, (24b) is equivalent to the L_1 penalty for $\alpha = \infty$.

Elasticnet: The elastic net, proposed by [27], is a penalized least squares regression technique that has been widely used in regularization and automatic variable selection to select groups of correlated variables. Note that the elastic net method linearly combines the L_1 penalty term, which enforces the sparsity of the elastic net estimator, and the L_2 penalty term, which ensures appropriate selection of correlated variable groups. Accordingly, the modified local estimator $\hat{\beta}_{ENL}$ using an elasticnet penalty is the solution of the following minimization problem:

$$\hat{\beta}_{ENL} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (\tilde{y}_{i\hat{G}}^* - \tilde{x}_i' \beta)^2 + \lambda_1 \sum_{j=1}^k |\beta_j|^2 + \lambda_2 \sum_{j=1}^k |\beta_j| \right\} \tag{25}$$

where λ_1 and λ_2 are the positive regularization parameters. Equation (25) ensures the estimates corresponding to the parametric part of the semiparametric regression model (2), as in the other methods.

MCP: Introduced by [28], MCP is an alternative method to obtain less biased estimates of the nonzero regression coefficients in a sparse model. For the given regularization parameters $\lambda > 0$ and $\alpha > 0$, the local polynomial estimator $\hat{\beta}_{MCL}$ based on the MCP penalty can be defined as

$$\hat{\beta}_{MCL} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (\tilde{y}_{i\hat{G}}^* - \tilde{x}_i' \beta)^2 + \sum_{j=1}^k P_{\alpha,\lambda}(|\beta_j|) \right\} \tag{26}$$

where $P_{\alpha,\lambda}(\cdot)$ is the MCP penalty given by

$$P_{\alpha,\lambda}(\beta) = \int_0^{|\beta|} \left(\frac{\lambda - x}{\alpha} \right)_+ dx = \left(\lambda |\beta| - \frac{\beta^2}{2\alpha} \right) \mathbf{I}(0 \leq |\beta| < \lambda\alpha) + \frac{\lambda^2 \alpha}{2} \mathbf{I}(|\beta| \geq \lambda\alpha)$$

3.2.2. The Estimation Procedure for the Nonparametric Component

Equations (21)–(26) provide modified local polynomial estimates based on different penalties for the parametric part of the semiparametric model in (2). Similar in spirit to (19b), the vector of estimated parametric coefficients $\hat{\beta}_{LL}$ given in (21) can be used to construct the estimation of the nonparametric part in the same model. In this case, we

obtain the modified local estimates based on the lasso penalty of the unknown function, given by

$$\hat{\mathbf{f}}_{LL} = \mathbf{S}\left(\mathbf{y}_G^* - \mathbf{X}\hat{\boldsymbol{\beta}}_{LL}\right) \tag{27}$$

as defined in the previous section.

Note that when $\hat{\boldsymbol{\beta}}_{aLL}$ defined in (23) is written instead of $\hat{\boldsymbol{\beta}}_{LL}$ in Equation (27), local estimates of the nonparametric part based on the adaptive lasso penalty are obtained and are stated as $\hat{\mathbf{f}}_{aLL}$ symbolically. Similarly, replacing $\hat{\boldsymbol{\beta}}_{LL}$ in (27) with $\hat{\boldsymbol{\beta}}_{SL}$, $\hat{\boldsymbol{\beta}}_{ENL}$, and $\hat{\boldsymbol{\beta}}_{MCL}$ yields modified local polynomial estimators, denoted as $\hat{\mathbf{f}}_{SL}$, $\hat{\mathbf{f}}_{ENL}$, and $\hat{\mathbf{f}}_{MCL}$ based on the SCAD, elasticnet, and MCP penalties, respectively, for the nonparametric part of the right-censored semiparametric model (2).

3.2.3. Some Remarks on the Penalties

Remarks on the penalties can be stated as follows:

- The regularization based on the L_1 norm produces sparse solutions as well as feature selection. However, the L_2 norm produces nonsparse solutions and does not have feature selection.
- Although all of the regularization methods shrink most of the coefficients towards zero, SCAD, MCP, and adaptive lasso apply less shrinkage to nonzero coefficients. This is known as bias reduction.
- As noted earlier, the tuning parameter α , used for SCAD and MCP estimations, controls how quickly the penalty rate goes to zero. This affects the bias and stability of the estimates, in the sense that there is a greater chance for more than one local minimum to exist as the penalty becomes more concave.
- As $\alpha \rightarrow \infty$, the MCP and SCAD penalties converge to the L_1 norm penalty. Conversely, as $\alpha \rightarrow 0$, the bias is minimized, but both MCP and SCAD estimates become unstable. Note also that lower values of the tuning parameter α for SCAD and MCP produce more highly variable, but less biased, estimates.
- The elasticnet penalty is designed to deal with highly correlated covariates more intelligently than other sparse penalties, such as the lasso. Note that the lasso penalty tends to choose one among highly correlated variables, while elasticnet uses them all.

4. Performance Indicators

Several performance measurements are described in this section with which to evaluate the performance of the modified six semiparametric local polynomial estimators based on penalty functions: ridge (*RL*), lasso (*LL*), adaptive lasso (*aLL*), SCAD (*SL*), MCP (*MCL*), and elasticnet (*ENL*). Note that the abbreviations given in parentheses here denote the estimators. The performance of the estimators are examined individually for the parametric component, nonparametric component, and overall estimated model. Accordingly, evaluation metrics are given by:

4.1. Measures for the Parametric Component

Root mean squared error (RMSE) of estimated regression coefficients ($\hat{\boldsymbol{\beta}}$). The calculation of *RMSE* is given by:

$$RMSE\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}\right) = \sqrt{k^{-1}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)} \tag{28}$$

where $\hat{\boldsymbol{\beta}}$ is the obtained estimate of $\boldsymbol{\beta}$ by any of the introduced six methods. It is replaced by $\hat{\boldsymbol{\beta}}_{RL}$, $\hat{\boldsymbol{\beta}}_{LL}$, $\hat{\boldsymbol{\beta}}_{aLL}$, $\hat{\boldsymbol{\beta}}_{SL}$, $\hat{\boldsymbol{\beta}}_{ENL}$, and $\hat{\boldsymbol{\beta}}_{MCL}$ to obtain the *RMSE* score for each estimator.

Coefficient of determination (R^2) for the estimated models. Note that R^2 allows us to see overall model performance of the six methods. It can be calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}, \quad i = 1, \dots, n \tag{29}$$

Sensitivity, specificity, and accuracy scores obtained from a confusion matrix. If true values of an interested variable are available, the confusion matrix can be obtained as in Table 1. This matrix allows us to measure the performance of the penalty functions for right-censored data. Accordingly, sensitivity, specificity, and accuracy values can be calculated as follows:

$$acc = (a + d) / (a + b + c + d); \text{sens} = a / (a + b); \text{spec} = d / (c + d) \tag{30}$$

Table 1. Confusion matrix for variable selection.

	Covariate of Interest	Irrelevant Covariate
Covariate of interest	a : no. true selection of covariate of interest	b : no. false selection of covariate of interest
Irrelevant covariate	c : no. false selection of irrelevant covariate	d : no. true selection of irrelevant covariate

G score calculated by geometric mean of sensitivity and specificity given in Equation (31):

$$G = \sqrt{\text{sens} \times \text{spec}} \tag{31}$$

4.2. Measures for the Nonparametric Component

Mean squared error (MSE) is used to measure the performance of the estimated nonparametric components by six methods: $\hat{\mathbf{f}}_{RL}$, $\hat{\mathbf{f}}_{LL}$, $\hat{\mathbf{f}}_{aLL}$, $\hat{\mathbf{f}}_{SL}$, $\hat{\mathbf{f}}_{ENL}$, and $\hat{\mathbf{f}}_{MCL}$. Assume that $\hat{\mathbf{f}}$ is the fitted nonparametric function obtained from any of the six methods. Accordingly, the MSE is computed as follows:

$$MSE(\hat{\mathbf{f}}) = n^{-1} \left(\sum_{i=1}^n f(t_i) - \hat{f}(t_i) \right)^2 = n^{-1} (\mathbf{f} - \hat{\mathbf{f}})^T (\mathbf{f} - \hat{\mathbf{f}}) \tag{32}$$

Relative MSE (ReMSE) is used to make a comparison between performances of the six methods on the estimation of the nonparametric component. The calculation of the ReMSE is given by

$$ReMSE(\hat{\mathbf{f}}_i) = n_m^{-1} \left[\sum_{i \neq j} MSE(\hat{\mathbf{f}}_i) / MSE(\hat{\mathbf{f}}_j) \right] \tag{33}$$

where n_m denotes the number of methods, which are six for this paper.

5. Simulation Study

We carried out an intense simulation study to evaluate the finite sample performance of the introduced six semiparametric estimators for a right-censored partially linear model. These estimators are compared with each other to evaluate their respective strengths and weaknesses in handling the right-censored problem. To obtain reproducibility, simulation codes with functions are provided in the following GitHub link: <https://github.com/yilmazersin13?tab=repositories> accessed on 31 August 2022. The estimators are computed using the

formulations in Section 3. The simulation design and data generation are described as follows:

Simulation Design: Two main scenarios are considered for generating the zero and nonzero coefficients of the model because the focus of the penalty functions is on making an accurate variable selection. In each scenario, simulation runs are made for

- Three sample sizes: $n = 50, 150, 300$
- Two censoring levels: $CL = 10\%, 30\%$
- Two numbers of parametric covariates $k = 1540$

All possible simulation configurations are repeated 1000 times. To evaluate the performance of the methods, the performance indicators described in Section 4 are used. The scenarios are defined in the data generation section below.

Data Generation: Regarding model (1), $z_i = \mathbf{x}_i' \boldsymbol{\beta} + f(t_i) + \varepsilon_i$, $1 \leq i \leq n$, each element of the model obtained as

$$\mathbf{x}_i \sim MN[\boldsymbol{\mu}_{k \times 1}, \boldsymbol{\Sigma}_{k \times k}]; t_i = 2.4(i - 0.5)/n, f(t_i) = -t_i \sin(-t_i^2)$$

The true values of regression coefficients are determined for both Scenarios 1 and 2 as follows:

$$\text{(Scenario 1)} \quad \beta_j = \begin{cases} 5 & \text{if } j = 1, \dots, 5 \\ -3 & \text{if } j = 11, \dots, 15 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{(Scenario 2)} \quad \beta_j = \begin{cases} 1 & \text{if } j = 1, \dots, 5 \\ -0.5 & \text{if } j = 11, \dots, 15 \\ 0 & \text{otherwise} \end{cases}$$

For both scenarios, there are 10 nonzero β_j 's to be estimated and $(k - 10)$ sparse coefficients. The main purpose of using these two scenarios is that it allows us to measure the capacity of the estimators on the selection of nonzero coefficients when β_j 's are close to zero. In addition, these scenarios make it possible to see how the censoring level (CL) affects their performances. *These scenarios allow us to inspect the convergence of the estimated coefficients to the true ones when the sample size is becoming larger practically under censored data, which can be counted as an important contribution of this paper.*

Regarding the censoring data, the censoring variable c_i is generated as $c_i \sim N(\mu_y, \sigma_y^2)$ independently of the initially observed variable y_i . An algorithm is provided by [29]. Another important point for this study is the selection of the shrinkage parameters and the bandwidth parameter for the local polynomial approach for the introduced six estimators. In this study, the improved Akaike information criterion (AIC_c), proposed by [30], is used. It can be calculated as follows:

$$AIC_c(\lambda; h) = \log(\sigma^2) + 1 + \left(\frac{2(\sum_{r < k} I(\hat{\beta}_r(\lambda, h) \neq 0) + 1)}{(n - \sum_{r < k} I(\hat{\beta}_r(\lambda, h) \neq 0) - 2)} \right)$$

where $\hat{\beta}_r(\lambda, h)$ is the estimated coefficient based on the shrinkage parameter $\lambda > 0$, and the bandwidth $h > 0$ and σ^2 is the variance of the model, and $\sum_{r < k} I(\hat{\beta}_r(\lambda, h) \neq 0)$ denotes the number of nonzero regression coefficients. Note that, due to the projection (hat) matrix, the introduced estimation procedures (except ridge regression) cannot be written; the number of nonzero coefficients is used instead of the hat matrix.

The results of the simulation study are presented individually for parametric and nonparametric components below. Before that, Figure 1 is presented to provide some information about the generated data. Figure 1 is formed by four panels. Panels (a) and (b) show the original, right-censored, and synthetic response values for $CL = 10\%$. Panels (c) and (d) show the same for $CL = 30\%$.

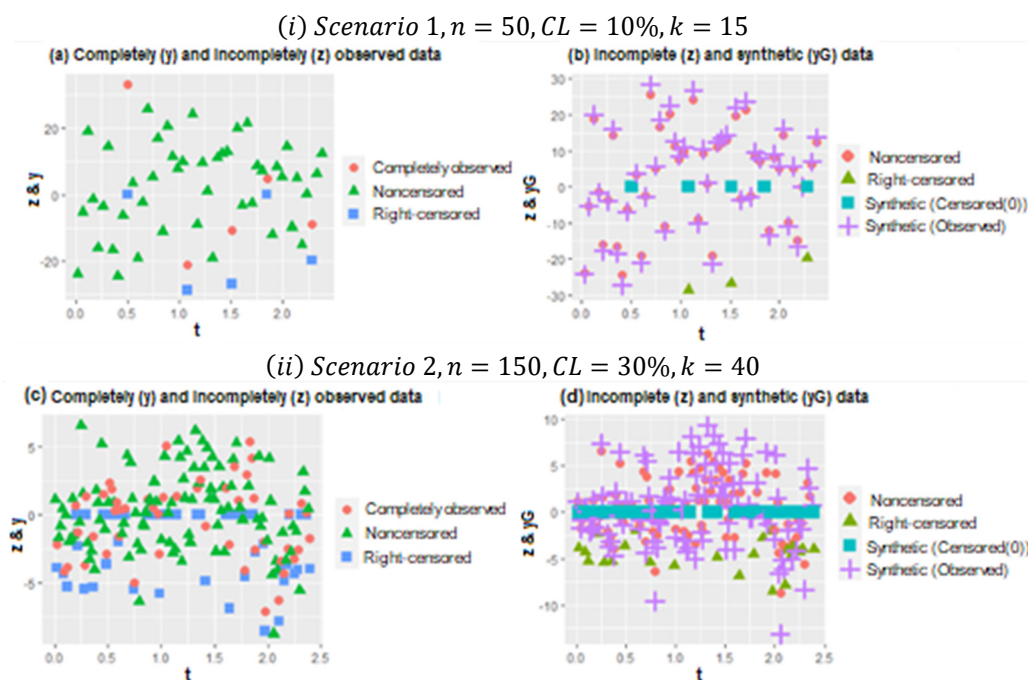


Figure 1. Overview for the generated data based on different simulation configurations.

In Figure 1, two plots (i – ii) are represented for the different configurations of Scenarios 1 and 2. In plot (i), scatter plots of the data points for $n = 50$, $CL = 10\%$, and $k = 15$ are given. In panel (b) of (i), the working procedure of synthetic data transformation can be seen clearly. It gives zero to right-censored observations and increases the magnitude of observed data points. Thus, it makes equal the expected value of the synthetic response variable and original response variable, as indicated in Section 2. Similarly, plot (ii) shows the scatter plots of data points for $n = 150$, $CL = 30\%$, and $k = 40$, which makes it possible to see heavily censored cases. In panel (b) of (ii), due to heavy censorship, the magnitude of the increments in the observed data points is larger than (i), which is the working principle of the synthetic data. This is a disadvantage because it significantly manipulates the data structure, although it still solves the censorship problem.

To describe the generated dataset further, Figure 2 shows the nonparametric component of the right-censored semiparametric model. In panel (a), the smooth function can be seen for the small sample size ($n = 50$), low censoring level ($CL = 10\%$), and low number of covariates ($k = 15$). Panel (b) shows the nonparametric smooth function for $n = 150$, $k = 15$, and $CL = 10\%$. It should be emphasized that the censoring level or number of the parametric covariates does not affect the shape of the nonparametric component. Thus, there is no need to show all possible generated functions here. Note that the nonparametric component affects the number of covariates and the censoring level indirectly in the estimation process.

As previously mentioned, this paper introduces six modified estimators based on penalty and shrinkage strategies. In Figure 3, the shrinkage process of the estimators, according to the shrinkage parameter “lambda,” is provided in panels (i) and (ii). In panel (i), shrunk regression coefficients are shown for scenario 1, $n = 150$, $CL = 30\%$, and $k = 40$. Panel (ii) is drawn for scenario 2, $n = 50$, $CL = 10\%$, and $k = 15$. When Figure 3 is inspected carefully, it can be seen that in panel (i), due to a high censoring level and many covariates, the shrinkage of the coefficients is more challenging than in panel (ii). One of the reasons for that is, in Scenario 2, coefficients are determined as smaller than the coefficients in Scenario 1 while generating data. For both panels, it can be observed that the SCAD and MCP methods behave similarly. As expected, they shrunk the coefficients quicker than the others. Additionally, lasso and ElasticNet seem close to each other for both

panels. However, adaptive lasso differs from the others in both panels. The reason for this is discussed with the results given in Section 5.1.

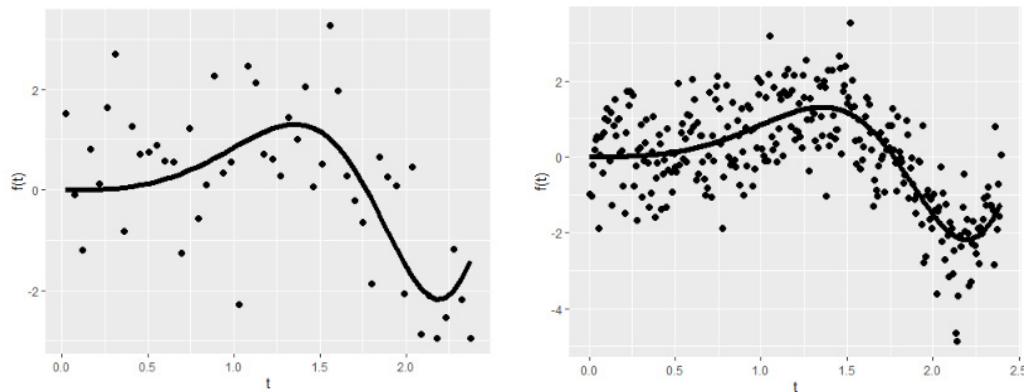


Figure 2. The generated nonparametric component of the semiparametric model for two simulation combinations.

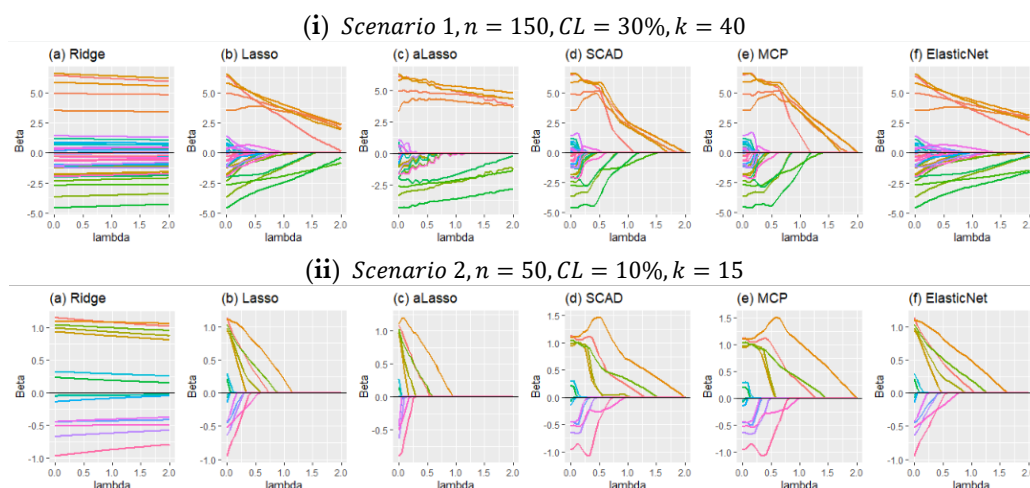


Figure 3. The behaviors of the introduced modified penalty functions according to shrinkage parameters for two different cases.

5.1. Analysis of Parametric Component

In this section, the estimation of the parametric component of a right-censored semi-parametric model is analyzed, and results are presented for all simulation scenarios in Tables 2–5 and Figures 4–7. The performance of the estimators is evaluated using the metrics given in Section 4: $RMSE$, R^2 , sensitivity, specificity, accuracy, and G -score. In addition to the performance criteria, the selection ratio of the methods is calculated for the estimators. The selection ratio can be defined as follows:

Selection Ratio: The ratio of the selected true nonzero coefficients by the corresponding estimator in 1000 simulation repetitions. The formulation can be given by:

$$Selection\ Ratio(\hat{\beta}_j) = \frac{1}{1000} \sum_{i=1}^{1000} \mathbf{I}(\hat{\beta}_j^{(i)} \neq 0),$$

where $\hat{\beta}_j^{(i)}$ is the estimated coefficient for i .th simulation by any of the introduced estimators. Results are given in Figures 6 and 7.

Table 2. RMSE and R^2 values obtained from all simulation runs (Scenario 1).

n	CL	k	$\hat{\beta}_{RL}$		$\hat{\beta}_{LL}$		$\hat{\beta}_{aLL}$		$\hat{\beta}_{SL}$		$\hat{\beta}_{MCL}$		$\hat{\beta}_{ENL}$	
			RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
50	0.1	15	0.779	0.908	0.782	0.905	0.758 *	0.906	0.743	0.909	0.787	0.908	0.777	0.907
		40	1.414	0.879	0.777 *	0.875	0.730	0.884	0.952	0.880	1.536	0.876	0.784	0.881
	0.3	15	1.602 *	0.653	1.630	0.646	1.670	0.640	1.620	0.651	1.617	0.651	1.581	0.653
		40	2.779	0.617	1.387 *	0.610	1.547	0.590	2.264	0.618	3.234	0.608	1.367	0.625
150	0.1	15	0.410	0.949	0.410	0.948	0.386 *	0.948	0.381	0.949	0.409	0.949	0.411	0.948
		40	0.423	0.942	0.312	0.945	0.236	0.950	0.294 *	0.950	0.423	0.942	0.379	0.944
	0.3	15	0.830	0.811	0.864	0.805	0.814	0.810	0.822 *	0.811	0.830	0.810	0.834	0.809
		40	0.854	0.780	0.692 *	0.787	0.621	0.799	0.783	0.789	0.855	0.778	0.741	0.791
300	0.1	15	0.263	0.968	0.267	0.967	0.237	0.968	0.239 *	0.969	0.261	0.968	0.262	0.968
		40	0.276	0.965	0.213	0.968	0.154	0.971	0.177 *	0.970	0.273	0.965	0.258	0.966
	0.3	15	0.562	0.895	0.583	0.893	0.550 *	0.895	0.545	0.896	0.562	0.895	0.568	0.895
		40	0.545	0.878	0.469 *	0.884	0.363	0.897	0.464	0.889	0.545	0.878	0.512	0.882

Bold ones are the best performance scores; *: the second-best score in RMSE scores.

Table 3. RMSE and R^2 values obtained for Scenario 2.

n	CL	k	$\hat{\beta}_{RL}$		$\hat{\beta}_{LL}$		$\hat{\beta}_{aLL}$		$\hat{\beta}_{SL}$		$\hat{\beta}_{MCL}$		$\hat{\beta}_{ENL}$	
			RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
50	0.1	15	0.211	0.890	0.207	0.884	0.204	0.885	0.205 *	0.889	0.209	0.889	0.203	0.889
		40	0.379	0.890	0.198	0.848	0.185 *	0.865	0.251	0.882	0.437	0.885	0.202	0.874
	0.3	15	0.363 *	0.635	0.365	0.617	0.384	0.614	0.365	0.629	0.365	0.630	0.355	0.632
		40	0.608	0.581	0.295 *	0.543	0.323	0.542	0.503	0.578	0.688	0.570	0.290	0.582
150	0.1	15	0.112	0.892	0.106	0.891	0.100	0.891	0.102 *	0.892	0.109	0.891	0.108	0.891
		40	0.114	0.899	0.083 *	0.893	0.064	0.899	0.083	0.900	0.110	0.896	0.101	0.899
	0.3	15	0.174	0.765	0.179	0.757	0.176	0.759	0.171	0.764	0.173 *	0.764	0.173 *	0.763
		40	0.189	0.737	0.151 *	0.733	0.141	0.743	0.174	0.740	0.187	0.734	0.169	0.743
300	0.1	15	0.074	0.897	0.069	0.897	0.064	0.897	0.066 *	0.897	0.071	0.897	0.071	0.897
		40	0.077	0.897	0.068	0.890	0.045	0.895	0.055 *	0.898	0.074	0.895	0.073	0.895
	0.3	15	0.125	0.834	0.126	0.831	0.121 *	0.832	0.120	0.834	0.123	0.833	0.124	0.833
		40	0.123	0.811	0.102 *	0.809	0.087	0.818	0.106	0.816	0.121	0.808	0.117	0.810

Bold ones are the best performance scores; *: the second-best score in RMSE scores.

Tables 1 and 2 include the RMSE scores for the estimated coefficients calculated from (4.1) and R^2 of the model for Scenarios 1 and 2. The best scores are indicated with bold text. If the two tables are inspected carefully, two observations can be made about the performance of the methods for both Scenarios 1 and 2. Regarding Scenario 1, when the sample size is small ($n = 50$), ENL and SL estimators give smaller RMSE scores and higher R^2 values than the other four methods. On the other hand, when the sample size becomes larger ($n = 150$, $n = 300$), aLL takes the lead in terms of estimation performance. The results for different censoring levels show that aLL is less affected by censorship than SL and the other methods. This can be observed in Table 1 clearly.

Table 4. G-score and accuracy values obtained for all simulation combinations (Scenario 1).

n	CL	k	$\hat{\beta}_{RL}$		$\hat{\beta}_{LL}$		$\hat{\beta}_{aLL}$		$\hat{\beta}_{SL}$		$\hat{\beta}_{MCL}$		$\hat{\beta}_{ENL}$	
			Acc	G	Acc	G	Acc	G	Acc	G	Acc	G	Acc	G
50	0.1	15	0.667	0.082	0.772	0.699	0.859	0.874 *	0.845 *	0.900	0.715	0.455	0.679	0.190
		40	0.250	0.050	0.767	0.724 *	0.828	0.797	0.717	0.688	0.439	0.548	0.576	0.612
	0.3	15	0.667	0.082	0.759 *	0.753	0.761	0.728 *	0.735	0.646	0.673	0.215	0.687	0.311
		40	0.250	0.050	0.756	0.694 *	0.746 *	0.713	0.561	0.549	0.338	0.470	0.656	0.628
150	0.1	15	0.667	0.082	0.772	0.734	0.916	0.905 *	0.859 *	0.912	0.716	0.472	0.676	0.131
		40	0.250	0.050	0.707	0.690	0.952	0.930	0.722 *	0.698 *	0.377	0.527	0.343	0.481
	0.3	15	0.667	0.082	0.780 *	0.752	0.843	0.815	0.771	0.805 *	0.692	0.357	0.672	0.143
		40	0.250	0.050	0.712 *	0.692 *	0.868	0.826	0.536	0.593	0.313	0.472	0.388	0.531
300	0.1	15	0.667	0.082	0.768	0.572	0.963	0.976	0.911 *	0.943 *	0.725	0.549	0.669	0.112
		40	0.250	0.050	0.700	0.687	0.967	0.948	0.805 *	0.768 *	0.400	0.543	0.322	0.440
	0.3	15	0.667	0.082	0.768	0.673	0.887	0.844	0.809 *	0.855 *	0.696	0.388	0.668	0.097
		40	0.250	0.050	0.670 *	0.668 *	0.936	0.906	0.595	0.621	0.319	0.482	0.339	0.516

Bold ones are the best performance scores; *: the second-best score in G-scores and accuracy.

Table 5. G-score and accuracy metrics for all simulation combinations (Scenario 2).

n	CL	k	$\hat{\beta}_{RL}$		$\hat{\beta}_{LL}$		$\hat{\beta}_{aLL}$		$\hat{\beta}_{SL}$		$\hat{\beta}_{MCL}$		$\hat{\beta}_{ENL}$	
			Acc	G	Acc	G	Acc	G	Acc	G	Acc	G	Acc	G
50	0.1	15	0.667	0.082	0.740	0.489	0.837	0.823 *	0.809 *	0.838	0.692	0.299	0.671	0.127
		40	0.250	0.050	0.785 *	0.721 *	0.793	0.751	0.668	0.642	0.378	0.505	0.516	0.571
	0.3	15	0.667	0.082	0.705	0.585	0.723 *	0.656	0.729	0.642 *	0.679	0.283	0.675	0.286
		40	0.250	0.050	0.752	0.585 *	0.731 *	0.843	0.526	0.525	0.326	0.462	0.580	0.272
150	0.1	15	0.667	0.082	0.699	0.433	0.895	0.890	0.841 *	0.873 *	0.701	0.435	0.667	0.082
		40	0.250	0.050	0.757 *	0.729 *	0.928	0.895	0.665	0.660	0.343	0.517	0.296	0.404
	0.3	15	0.667	0.082	0.749	0.546	0.837	0.796	0.767 *	0.714 *	0.692	0.371	0.668	0.097
		40	0.250	0.050	0.721 *	0.693 *	0.859	0.816	0.517	0.585	0.307	0.470	0.316	0.472
300	0.1	15	0.667	0.082	0.709	0.512	0.953	0.969	0.860 *	0.897 *	0.724	0.592	0.668	0.097
		40	0.250	0.050	0.625	0.679 *	0.951	0.927	0.696 *	0.679 *	0.358	0.521	0.292	0.255
	0.3	15	0.667	0.082	0.709	0.353	0.892	0.916	0.788 *	0.800 *	0.688	0.311	0.667	0.082
		40	0.250	0.050	0.719 *	0.690 *	0.912	0.877	0.572	0.609	0.313	0.472	0.277	0.327

Bold ones are the best performance scores; *: the second-best score in G-scores and accuracy.

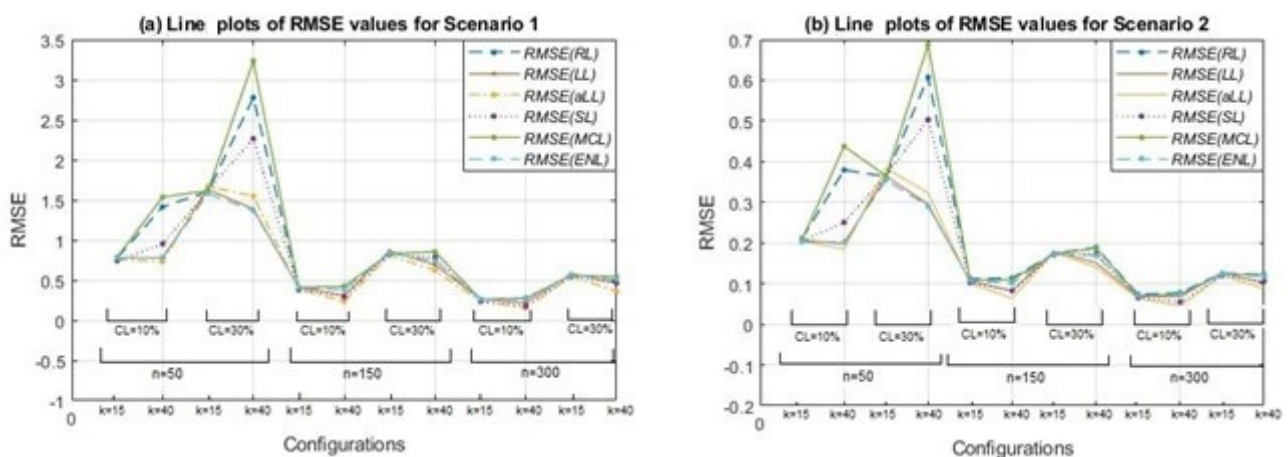


Figure 4. Line plots of RMSE scores given in Tables 2 and 3 for all simulation configurations.

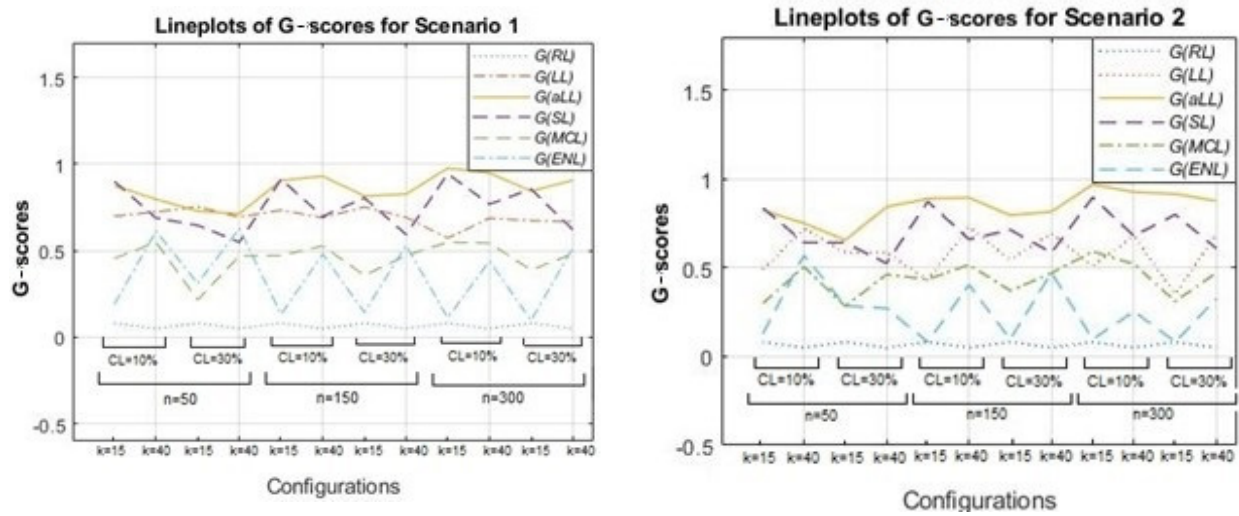


Figure 5. Line plots for G-scores given in Tables 4 and 5 for all simulation configurations.

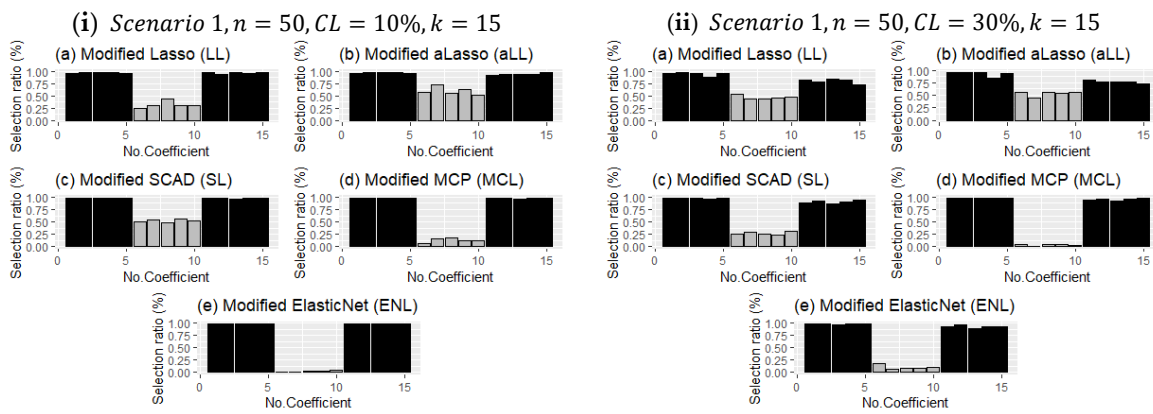


Figure 6. Comparison of the modified estimators regarding deciding the true coefficient for all simulation runs. The dark-colored bars denote the selection ratios of nonzero regression coefficients, and the gray-colored ones represent the ratios for sparse coefficients.

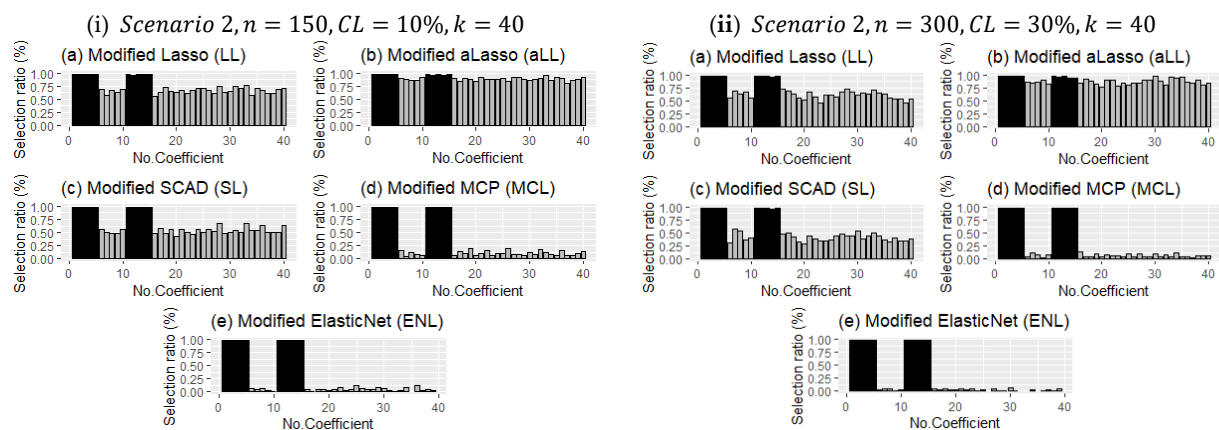


Figure 7. Comparison of the modified estimators as in Figure 6 but for different simulation configurations when $k = 40$.

In Table 2, RMSE and R^2 scores are provided for all simulation configurations of Scenario 2. The results can be distinguished from the results in Table 1 by the higher R^2 values obtained from the modified ridge estimator. However, the RMSE scores of the modified ridge estimator are the largest. This can be explained by the fact that the ridge

penalty uses all covariates, whether sparse or nonsparse. Therefore, the estimated model based on the ridge penalty has larger R^2 values. On the other hand, the RMSE scores prove that for small sample sizes ($n = 50$), ENL and aLL perform satisfactorily. Moreover, as in the case of Scenario 1, when the sample size becomes larger, aLL gives the most satisfying performance. SL- and LL-based estimators also show good performances in the general frame. If Table 2 is inspected in detail, it can be seen that when $k = 15$, the SL method comes to the front for both low ($CL = 10\%$) and high ($CL = 30\%$) censoring levels. The same is true for the aLL method regarding strength against censorship.

Figure 4 presents the line plots of the RMSE scores for all simulation cases. As expected, the negative effects of increment on the censoring level and the positive effect of growth on the sample size can be clearly observed from panels (a) and (b). For both scenarios, a peak can be seen when the sample size is small ($n = 50$) and the censorship level increases from 10% to 30%. The methods most affected by censorship are MCL, RL, and SL. The least affected are aLL, lasso, and ENL. Thus, Figure 4 supports the results and inferences obtained from Tables 2 and 3.

Together with the RMSE scores, one other important metric to evaluate the performance of the parametric component estimation is G -score, which measures the true selection made by the estimators for the sparse and nonzero subsets based on the confusion matrix given in Table 1. In this context, Figure 5 is drawn to illustrate the G -scores of the methods for all simulation combinations using line plots. Note that the G -score changes between the range $[0,1]$ and the lines of methods that are close to 1 are notated as better than the others in terms of successful determination of sparsity.

Figure 5 is formed by two panels: Scenario 1 (left) and Scenario 2 (right). As expected, for all methods except for RL (which does not involve any sparse subset and is therefore not shown in Figure 5 and Table 4), the G -scores diminish when the censoring level is high, and the number of covariates (k) is large. In addition, there is an increasing trend from the small to large sample sizes for LL, aLL, SL, and MCL. This trend is most evident for the aLL line, which makes aLL distinguishable. However, interestingly, ENL is not influenced by the change in sample size, and the G -scores of ENL do not take a value greater than 0.5. In general, aLL, SL, and LL provide the highest G -scores. All G -scores for the simulation study are provided in Tables 4 and 5 together with the accuracy values of the methods.

Tables 4 and 5 present the accuracy rates and G -scores for all methods and simulation configurations for both Scenarios 1 and 2. Note that, because the ridge penalty is unable to shrink the estimated coefficients towards zero, the specificity of RL is always calculated as zero. Thus, RL does not have a G -score. When the tables are examined, it can be clearly seen that the prominent methods are aLL, LL, and SL. The aLL produces satisfactory results for each simulation configuration. On the other hand, the other two methods, LL and SL, give good results under different conditions. When this situation is examined in detail, it can be seen that the SL method produces better results when $k = 15$, and the LL method when $k = 40$ with aLL. In addition, it can be said that the level of censorship and the sample size do not affect this situation, except for an increase or decrease in the values. These inferences apply to both scenarios. Here, the difference between the scenarios emerges in the size of the G -scores and accuracy values obtained. It can be said that they are slightly less than the values obtained for Scenario 2.

Unlike the evaluation criteria given in Section 4, the frequency of choosing the correct coefficients for each method in the simulation study is analyzed, and the results are presented for both scenarios in Figures 6 and 7 with bar plots. Figure 6 presents two panels (I and ii), which demonstrate the impact of censorship, one of the main purposes of this article. As expected, as censorship increases, the frequency of selection decreases for the nonzero coefficients. The point here is to reveal which methods are less affected by this. It can be observed in Figure 6 that the MCL and ENL methods are less affected by censorship in terms of the frequency of selection of nonzero coefficients. However, since these methods are less efficient than the SL, LL, and aLL methods in determining ineffective coefficients, their overall performance is lower (see Tables 4 and 5). On the other hand, the SL, LL,

and aLL methods make a balanced selection for both subsets (no effect-non-zero), which indirectly makes them more resistant to censorship.

Figure 7 presents the different configurations for Scenario 2. It shows both effects of sample size and censoring level increment and bar plots for $k = 40$. The detection performance of the methods of nonzero coefficients is less affected than in Figure 6. However, the selection of the ineffective set plays a decisive role in terms of the performance of each of the methods. For example, MCL and ENL performed poorly in the correct determination of ineffective coefficients when the censorship level increased, but LL, aLL, and SL were able to make the right choice under heavy censorship.

5.2. Analysis of Nonparametric Component

This section is prepared to show the behaviors of the estimation of nonparametric components by the introduced six estimators. Performance scores of the methods are given in Tables 6 and 7, and MSE and ReMSE metrics are used. Additionally, Figures 8 and 9 are provided to show the real smooth function versus all estimated curves for individual simulation repeats. These figures can provide information about the variation of the estimates according to both scenarios and censoring effects. Finally, in Figure 10, estimated curves obtained from all methods are inspected with four different configurations.

Table 6. Performance scores of fitted curves by the modified estimators for Scenario 1.

n	CL	k	\hat{f}_{RL}		\hat{f}_{LL}		\hat{f}_{aLL}		\hat{f}_{SL}		\hat{f}_{MCL}		\hat{f}_{ENL}		
			MSE	ReMSE	MSE	ReMSE	MSE	ReMSE	MSE	ReMSE	MSE	ReMSE	MSE	ReMSE	
50	0.1	15	1.26	1.06	1.21	1.01	1.19	0.98	1.17	0.97	1.21	1.01	1.18	0.98	
		40	1.26	0.97	1.50	1.20	1.14	0.84	1.22	0.94	1.50	1.20	1.20	0.92	
	0.3	15	4.17	1.00	4.29	1.04	4.11	0.98	4.15	1.00	4.13	0.99	4.14	0.99	
		40	4.73	1.12	4.59	1.08	4.17	0.90	3.50	0.77	5.26	1.26	4.07	0.93	
	150	0.1	15	0.75	1.12	0.67	0.98	0.62	0.89	0.68	1.00	0.69	1.02	0.69	1.02
			40	0.60	1.04	0.60	1.04	0.61	1.04	0.54	0.91	0.57	0.97	0.58	1.00
0.3		15	2.38	1.08	2.22	0.99	2.19	0.97	2.20	0.98	2.20	0.98	2.20	0.98	
		40	1.87	1.04	2.13	1.21	1.65	0.88	1.75	0.96	1.72	0.94	1.81	1.00	
300	0.1	15	0.45	1.07	0.43	1.01	0.41	0.95	0.43	1.01	0.42	0.98	0.42	0.98	
		40	0.39	1.07	0.35	0.94	0.36	0.97	0.36	0.97	0.37	1.01	0.38	1.04	
	0.3	15	1.29	1.01	1.28	1.00	1.24	0.97	1.29	1.01	1.28	1.00	1.27	1.00	
		40	1.12	1.06	1.14	1.08	1.01	0.93	1.06	0.99	1.04	0.97	1.06	0.99	

Bold ones are the best performance scores.

Tables 6 and 7 include the MSE and ReMSE values for the two scenarios. For Scenario 1, the aLL method gives more dominant values than others, followed by SL and LL. As expected, RL shows the worst performance; however, the difference from the others is small. Note that, when the sample size becomes larger, all methods begin to give similar results. Dependent on this similarity, the ReMSE scores become closer to one, which is an expected result. Thus, even if the censoring level increases, ReMSE scores may decrease. If the tables are inspected carefully, as mentioned in Section 5.1, aLL overcomes the censorship problem better than the others regarding Scenario 1, which means that contributions of covariates are high. However, in Scenario 2, SL shows better performance in high censoring levels, especially in small and medium sample sizes. Additionally, it is clearly observed that the number of covariates (k) affects the performances. In Table 7, when $k = 15$, the LL and SL methods show good performances.

Table 7. Performance scores of fitted curves by the modified estimators for Scenario 2.

n	CL	k	\hat{f}_{RL}		\hat{f}_{LL}		\hat{f}_{aLL}		\hat{f}_{SL}		\hat{f}_{MCL}		\hat{f}_{ENL}	
			MSE	ReMSE	MSE	ReMSE	MSE	ReMSE	MSE	ReMSE	MSE	MSE	ReMSE	MSE
50	0.1	15	0.85	1.02	0.83	0.99	0.81	0.95	0.78	0.92	0.87	1.05	0.88	1.06
		40	0.99	1.12	1.18	1.38	0.72	0.71	0.74	0.79	1.09	1.26	0.85	0.94
	0.3	15	2.14	0.94	2.27	1.01	2.30	1.03	2.22	0.99	2.29	1.02	2.25	1.00
		40	2.52	1.13	2.67	1.21	2.08	0.86	1.98	0.84	2.48	1.11	2.10	0.91
150	0.1	15	0.54	0.98	0.56	1.02	0.52	0.94	0.57	1.04	0.56	1.02	0.56	1.02
		40	0.56	1.21	0.50	1.06	0.41	0.81	0.45	0.93	0.48	1.01	0.49	1.03
	0.3	15	1.33	1.06	1.29	1.02	1.21	0.94	1.26	0.99	1.26	0.99	1.26	0.99
		40	1.14	1.09	1.04	0.98	1.06	0.97	0.94	0.86	1.09	1.03	1.12	1.07
300	0.1	15	0.52	1.02	0.50	0.98	0.51	1.00	0.50	0.98	0.52	1.02	0.51	1.00
		40	0.41	1.05	0.39	0.99	0.35	0.84	0.37	0.93	0.43	1.11	0.43	1.11
	0.3	15	0.87	1.02	0.84	0.98	0.85	0.99	0.85	0.99	0.87	1.02	0.86	1.00
		40	0.81	1.03	0.77	0.97	0.76	0.95	0.80	1.01	0.81	1.03	0.80	1.01

Bold ones are the best performance scores.

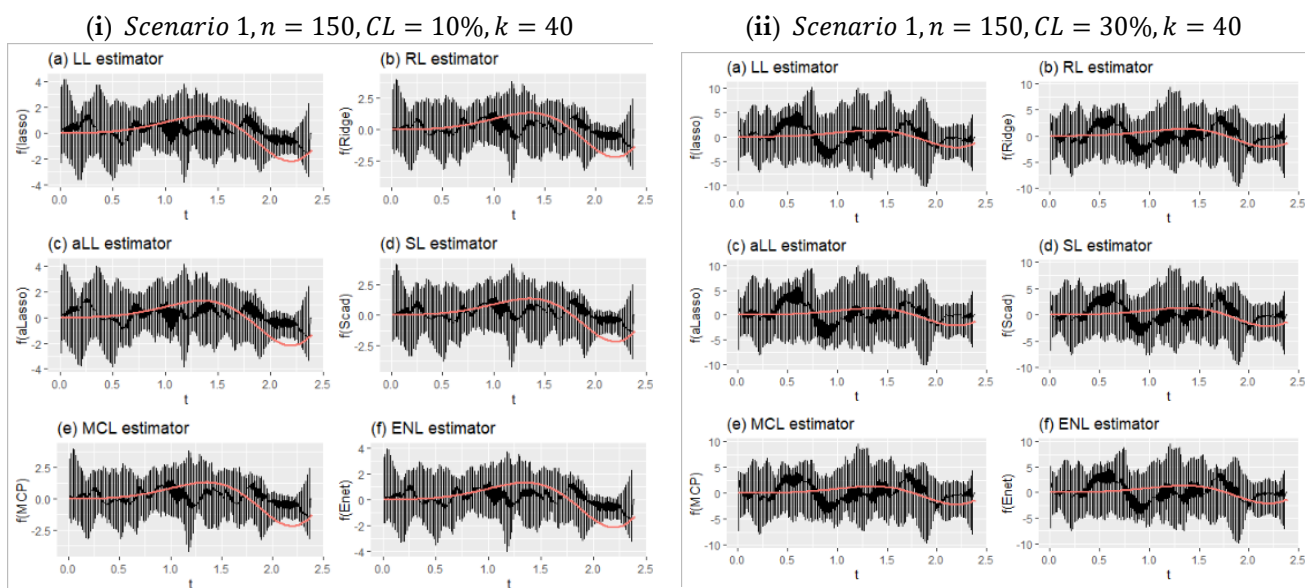


Figure 8. Obtained fitted curves for the methods from all simulation runs for Scenario 1.

Figure 8 shows two different simulation configurations for Scenario 1. The purpose of this figure is to illustrate the effect of censorship in curve estimation. Therefore, panel (i) is obtained for 10% censorship, and panel (ii) for 30% censorship. As can be seen at a glance, the minimum and maximum points of the prediction points obtained from all simulations around the real curve are shown with vertical lines. This reveals the range of variation of the estimators. Accordingly, when the difference between the effect of censorship panel (i) and panel (ii) is examined, it can be seen how the range of variation widens. It can be said, with the help of the values in Table 6, that the estimator with the least expansion is aLL and the method with the most is RL. It should also be noted that the SL and LL methods also showed satisfactory results.

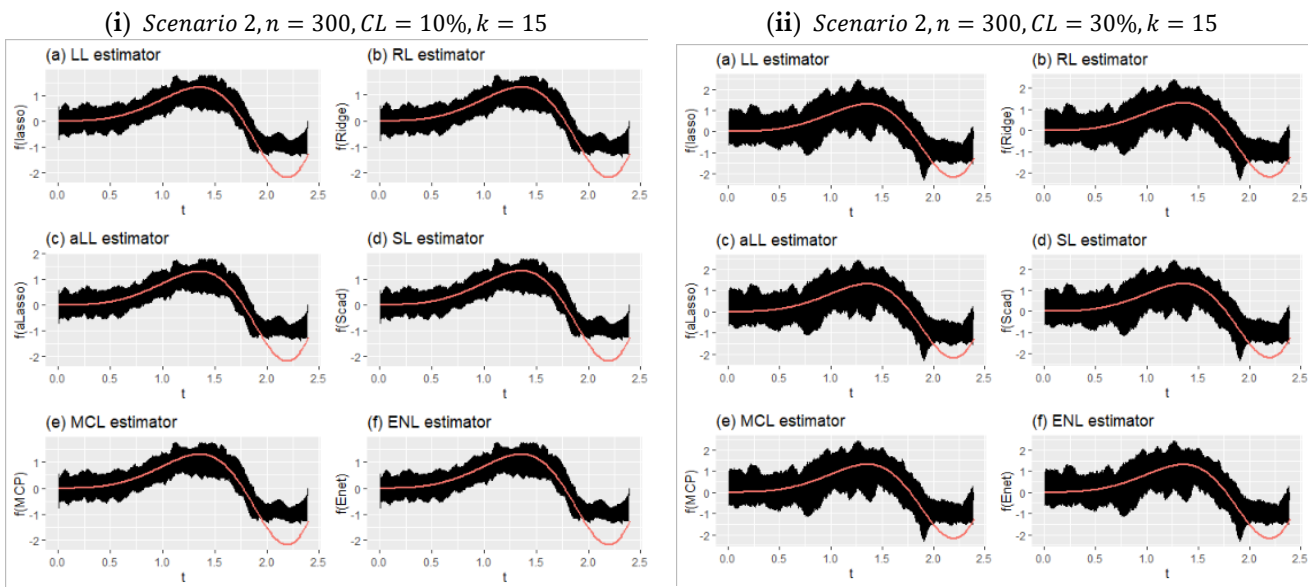


Figure 9. Fitted curves for Scenario 2 for the simulation settings given in (i) and (ii).

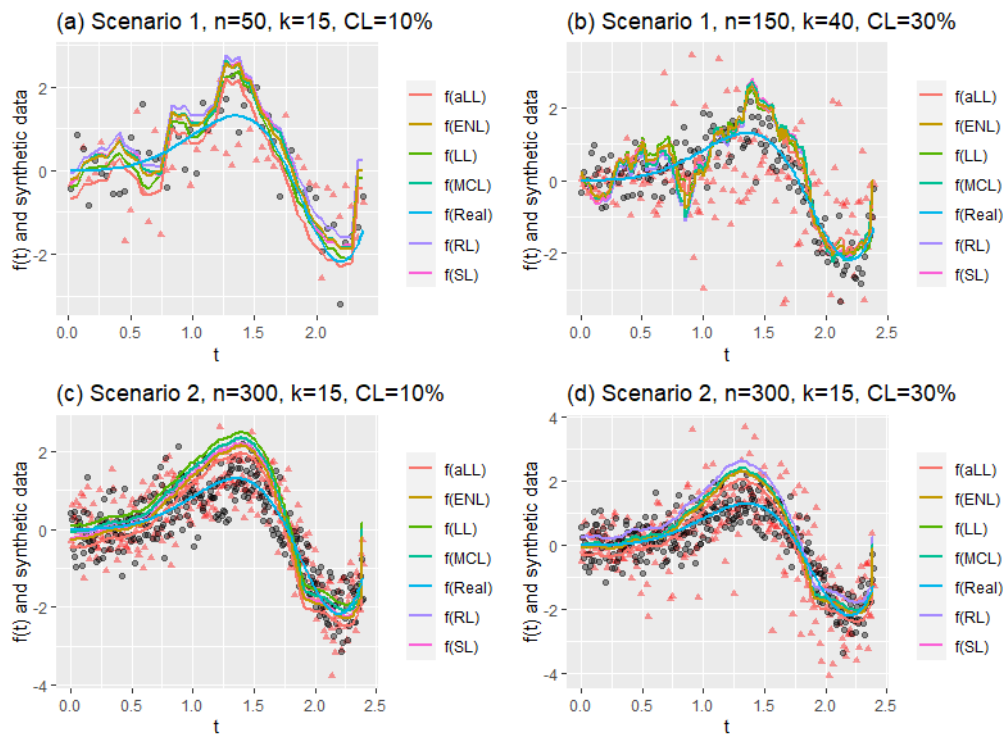


Figure 10. Mean of fitted curves obtained for the different simulation configurations. Red triangles (Δ) show the synthetic response values (t vs. $\hat{Y}_C - \hat{X}\beta$), and black dots (\cdot) denote the original data points t vs. $f(t)$.

Figure 9 shows the effect of censorship on the estimated curves for Scenario 2 with a large sample size and relatively few covariates ($k = 15$). Because there are too many data points, the lines appear as a black area. Compared with Figure 8, the effect of censorship is less, and the estimators obtain curves closer to the true curve. In addition, due to the large sample size, each method estimated very close curves. This can be clearly seen in Table 7. The obtained performance values were very close to each other. It can therefore be said that the introduced six estimators produce satisfactory results in high samples, and they are relatively less affected by censorship in this scenario.

Figure 10 consists of four panels (a)–(d) containing four different simulation cases. The first two panels (a and b) show the estimated curves of Scenario 1 for different sample sizes, different censorship levels, and different numbers of explanatory variables. It can be clearly seen that the curves in panel (a) are smoother than in panel (b). This can be explained by the messy scattering of synthetic data, which can be observed in all panels. The censorship level increases the corruption of the data structure. Similarly, panels (c) and (d) are obtained for Scenario 2, but only to observe the effect of the change in censorship level. However, the effect of the large sample size is clearly visible, and the curves appear smooth in panel (d), despite the deterioration in the data structure. If examined carefully, the aLL method gives the closest curve to the true curve. At the same time, the other methods have shown satisfactory results in representing the data.

6. Hepatocellular Carcinoma Dataset

This section contains the estimation of a right-censored partially linear model for real data, the Hepatocellular carcinoma dataset, by the introduced six estimators (RL, LL, aLL, SL, MCL, and ENL). Their performances are compared, and the results are presented in Table 8 and Figures 11–14. The dataset was collected by [31] to study CXCL17 gene expression for hepatocellular carcinoma.

Table 8. Scores of the evaluation metrics obtained from the hepatocellular carcinoma dataset.

	<i>RL</i>	<i>LL</i>	<i>aLL</i>	<i>SL</i>	<i>MCL</i>	<i>ENL</i>	<i>n</i>	<i>CL</i>	<i>k</i>
R^2	0.406	0.356	0.355 *	0.357 *	0.363	0.362			
No. ($\hat{\beta} = 0$)	0	3	5	4	3	1			
No. ($\hat{\beta} \neq 0$)	12	9	7	8	9	11	227	37%	12
$MSE(\hat{f})$	3.562	3.646	3.554 *	3.541	3.609	3.613			

Bold ones are the best performance scores; *: the second-best scores.

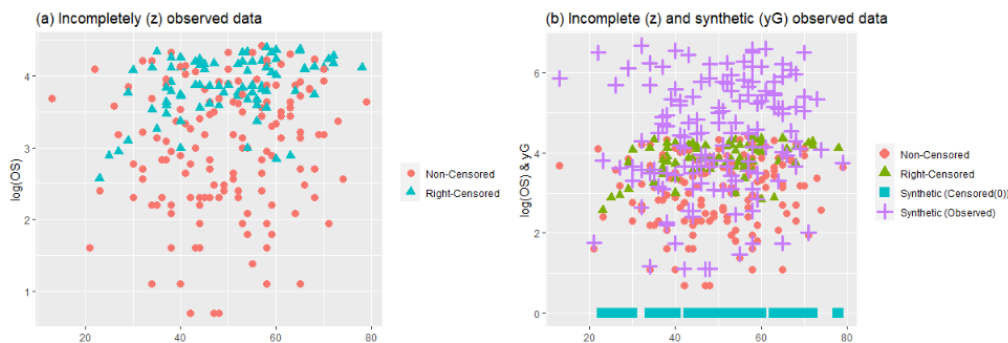


Figure 11. Descriptive plots for the right-censored hepatocellular carcinoma dataset.

The aforementioned dataset involves 227 data points and 13 explanatory variables, including age, recurrence-free survival (RFS_i), gender (Gen_i), and HBsAg (*surface antigen of the hepatitis B virus – HB_i*). Some variables that were obtained from blood tests to measure liver damage include ALT_i (*alanine aminotransferase*), AST_i (*aspartate aminotransferase*), and AFP_i (*alpha – fetoprotein*). The covariates of tumors detected in the liver are tumor size (TS_i), TNM_i (tumor node and metastasis), $BCLC_i$ (*Barcelona Clinic Liver Cancer Staging System*) and values of genes related to liver cancer: $CXCL17T$ ($CXCT_i$), $CXCL17P$ ($CXCP_i$), and $CXCL17N$ ($CXCN_i$). Note that the logarithm of the overall survival time (OS_i) is used as a response variable. Note also that the age variable is used as a nonparametric covariate because of its nonlinear structure. The remaining 12 explanatory variables are added to the parametric component of the model. Accordingly, the right-censored partially linear model can be written as follows:

$$\log(OS_i) = \mathbf{X}_i^T \boldsymbol{\beta} + f(\text{age}_i) + \varepsilon_i, 1 \leq i \leq 227 \tag{34}$$

where

$$\mathbf{X}_i^T = [RFS_i, Gen_i, HB_i, ALT_i, AST_i, AFP_i, TS_i, TNM_i, BCLC_i, CXCT_i, CXCP_i, CXCNi]$$

is the (228×12) -dimensional covariate matrix for the parametric component of the model, and $\beta = (\beta_1, \dots, \beta_{12})^T$ is the (12×1) -dimensional vector of the regression coefficients to be estimated. Note that in the estimation process, $\log(OS_i)$ cannot be used directly because of censoring. Therefore, synthetic data transformation is applied to $\log(OS_i)$ as in (6). Note also that the dataset includes 84 right-censored survival times, which means that the censoring level is $CL = 37\%$. This ratio can be interpreted as a heavy censoring level in the simulation study. Therefore, it is expected that the results of the real data example should be in harmony with the results of corresponding simulation configurations ($n = 150, 300, k = 15, CL = 30\%$).

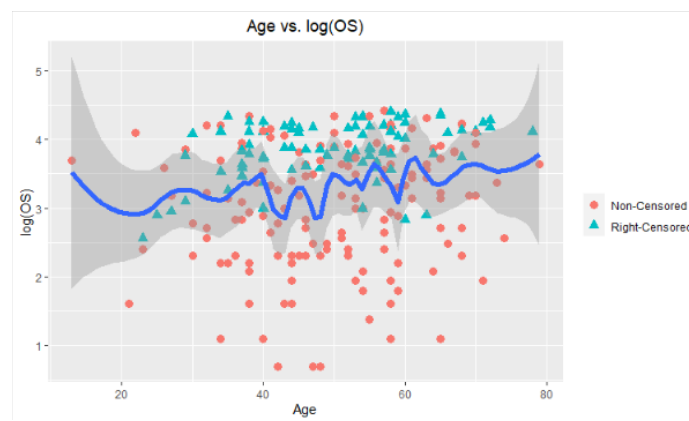


Figure 12. Plot for nonparametric covariate (age) with hypothetical curve (blue line) with confidence intervals (gray areas).

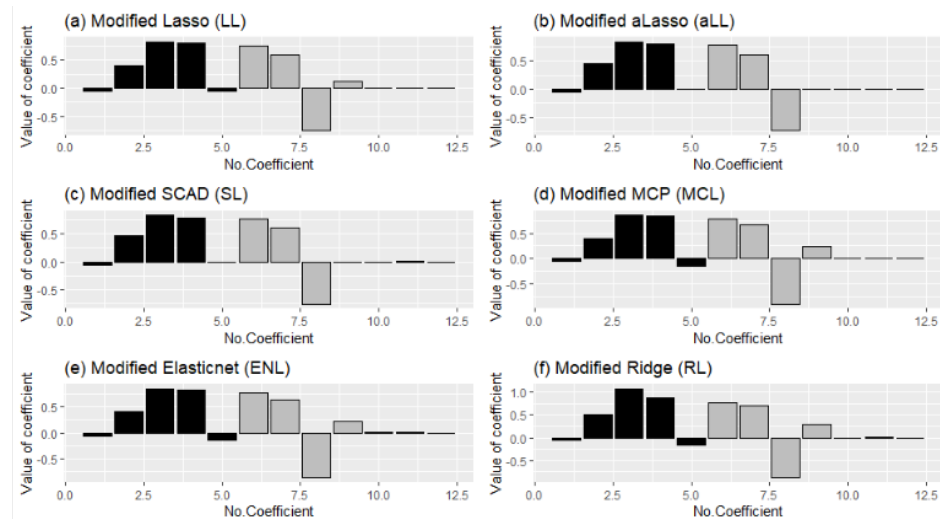


Figure 13. Bar plots of estimated coefficients obtained from the modified six estimators.

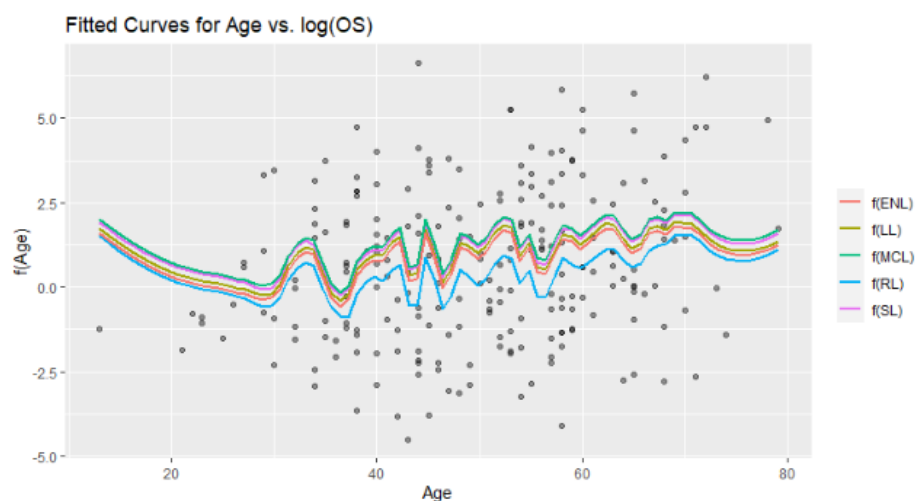


Figure 14. Fitted curves for a nonparametric component of the model obtained from the six modified estimators.

To describe the hepatocellular carcinoma dataset, Figures 11 and 12 are provided. Figure 11 is constructed by two panels, (a) and (b). In panel (a), a scatter plot of the data points can be seen with censored and noncensored points. As can be observed, there are a lot of right-censored points in the dataset. To solve this problem, synthetic data transformation is realized and is shown in panel (b). The synthetic data give zero value to right-censored points and increase the magnitude of the remaining data. Thus, it aims to make equal the expected values of $Y_{i\hat{G}}$ and completely observed response Y (but we do not know in real cases). Figure 12 presents the plot for response variable $\log(OS)$ versus nonparametric covariate age to show the nonlinear relationship between them. Accordingly, a hypothetical curve is presented, which proves our claim on the nonlinear relationship.

General outcomes for the analysis of the hepatocellular carcinoma dataset are presented in Table 8, which involves the performance scores of the six estimators. Note that, here, G -score cannot be calculated due to real regression coefficients being unknown. In Table 8, RL gives the highest R^2 value because it uses all 12 covariates in model estimation, and sparse and nonzero subsets are considered. The aLL and SL methods provide satisfying values with fewer covariates, especially aLL. Regarding the estimation of the nonparametric component, SL gives the best estimation, which supports our inference given before. In addition, aLL gives a smaller MSE value than the other four estimators.

The estimated coefficients are shown with bar plots in Figure 13 to illustrate how the methods work and to make a healthy comparison. In panels (b) and (c), the similar process of aLL and SL can be observed clearly. The ENL and RL methods also look similar to each other, which can be understood from Table 8.

Figure 14 involves the six fitted curves obtained by the introduced estimators. At first glance, all the fitted curves are very close to each other, which can be monitored in the MSE scores given in Table 8. However, the difference between RL and the other five methods can be easily observed. Due to the data structure having excessive variation, in the modeling process, the local polynomial method gains importance because it takes into account the local densities. This case can be counted as one of the important contributions of this paper.

7. Conclusions

The results of the paper obtained from the simulation study are given in Tables 2–7 and Figures 4–10. The analysis is made for both parametric and nonparametric components of the model individually. The advantage of the simulation study is knowing the real values of the regression coefficients; the accuracy and sensitivity of the estimators are thus evaluated, using the confusion matrix in Table 1. From the results, the aLL and SL estimators showed the best performance and gave satisfactory results for the model

estimation. In addition, the behaviors of the methods are inspected for three cases, which are sample sizes ($n = 50, 150, 300$), number of covariates ($k = 1540$), and censoring level ($CL = 10\%, 30\%$). These effects are also observed by the figures. A real data example using the hepatocellular carcinoma dataset is analyzed using the introduced estimators. The results of that dataset are compared with the related simulation configurations. By using the mentioned results, concluding remarks are given as follows:

- From the simulation results regarding the parametric component estimation, Tables 2–5 prove that the aLL and SL methods give satisfying results in terms of the metrics RMSE, R^2 , accuracy, and G-score. In more detail, for small sample sizes and low censoring levels, SL generally shows better performance than the other five methods. However, for the problematic scenarios, the aLL estimator is the best in both estimation performance and making a true selection between zero and nonzero subsets. Figures 4 and 5 support these inferences.
- In addition to introduced evaluation metrics, the selection frequency of the estimators is inspected for the simulation study, and results are shown in bar plots given in Figures 6 and 7. These figures demonstrate the consistency of the estimators in terms of their selection of the sparse and nonzero subsets for each coefficient. Under heavy censorship, it can be seen that LL, aLL, and SL gave the best performances. The ENL and MCL estimators did not show a good performance in this case.
- The introduced six estimators provide closer performances on the estimation of the nonparametric component. Corresponding results are given in Tables 6 and 7 and Figures 8–10. Note that Figures 8 and 9 are drawn to show the individual fitted curves obtained from each simulation, which provides information about the variation of the estimators. Although the estimators give similar evaluation scores and closer fitted curves (which is seen in Figure 10), aLL and SL are the best.
- In the hepatocellular carcinoma dataset analysis, the outcomes are found in harmony with the corresponding simulation scenarios. The results are provided in Table 8 and Figures 13 and 14. Similar to the simulation study, SL and aLL show the best performance. However, from Figure 14, it can be seen that the fitted curves are very close to each other, which can be explained by the large sample size. The real data study demonstrates that all six estimators show considerably good model estimates, which makes valuable the contribution of the paper.

Finally, from the results of both the simulation and real data studies, the introduced six estimators for the right-censored partially linear models based on penalty and shrinkage strategies are compared, and results are presented. It is found that the adaptive lasso (aLL) and SCAD (SL) methods are more resistant than the other four estimators against the effects of censorship and the number of covariates. In general, the ridge (RL) estimator showed poor performance. On the other hand, the lasso (LL), MCP (MCL), and elasticnet (ENL) methods provided good performance for both the parametric and nonparametric components. This study recommends the aLL and SL estimators for the problematic scenarios.

Author Contributions: Conceptualization: S.E.A. and D.A.; methodology: S.E.A. and D.A.; formal analysis and investigation: D.A. and E.Y.; writing—original draft preparation: D.A.; writing—review and editing: S.E.A. and D.A.; data curation: E.Y.; visualization: E.Y.; software: E.Y.; supervision: S.E.A. and D.A.; funding acquisition: E.Y.; 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 S. Ejaz Ahmed was supported by the Natural Sciences and the Engineering Research Council (NSERC) of Canada.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The simulation dataset can be regenerated by the codes provided in <https://github.com/yilmazersin13?tab=repositories> accessed on 31 August 2022. The hepatocellular carcinoma dataset is publicly available in R-package named “asaur”.

Acknowledgments: We would like to thank to the editor and referees. 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.

Appendix A

Appendix A.1. Proof of Lemma 1

Proof. From A(i) and A(ii) given in Assumption 1, we obtain

$$\begin{aligned} E(y_{iG}^* | X, t) &= E\left[\frac{\delta_i Y_i}{G(y_i)} | X, t\right] = E\left[\frac{I(Z_i \leq C_i) \min(Z_i, C_i)}{G[\min(Z_i, C_i)]} | X, t\right] \\ &= E\left[I(Z_i \leq C_i) \frac{Z_i}{G(y_i)} | X, t\right] \\ &= E\left[E\left[\frac{Z_i}{G(y_i)} I(Z_i \leq C_i) | X, t, Z\right] | X, t\right] \\ &= E\left[\frac{Z_i}{G(y_i)} \bar{G}(y_i) | X, t\right] = E(Z_i | X, t) \end{aligned}$$

This result shows that Lemma 1 has been proven. □

Appendix A.2. Derivation of Equations (15) and (16)

From Equation (14), the minimum criterion is rewritten as

$$L = [(\mathbf{I} - \mathbf{S})\mathbf{y}_G^* - (\mathbf{I} - \mathbf{S})\mathbf{X}\beta]^T [(\mathbf{I} - \mathbf{S})\mathbf{y}_G^* - (\mathbf{I} - \mathbf{S})\mathbf{X}\beta] \tag{A1}$$

By solving this equation with a partial derivative of (A1) with respect to β , Equation (15) can be obtained as follows:

$$L = \mathbf{y}_G^{*T} (\mathbf{I} - \mathbf{S}) \mathbf{y}_G^* - \mathbf{y}_G^{*T} (\mathbf{I} - \mathbf{S}) \mathbf{X} \beta - \beta^T \mathbf{y}_G^{*T} (\mathbf{I} - \mathbf{S}) \mathbf{X} + \beta^T \mathbf{X}^T (\mathbf{I} - \mathbf{S}) \mathbf{X} \beta$$

$$\frac{\partial L}{\partial \beta} = -2\mathbf{y}_G^{*T} (\mathbf{I} - \mathbf{S}) \mathbf{X} + 2\mathbf{X}^T (\mathbf{I} - \mathbf{S}) \mathbf{X} \beta = 0$$

$$\mathbf{X}^T (\mathbf{I} - \mathbf{S}) \mathbf{X} \beta = \mathbf{y}_G^{*T} (\mathbf{I} - \mathbf{S}) \mathbf{X}$$

$$\hat{\beta}_L = (\mathbf{X}^T (\mathbf{I} - \mathbf{S}) \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{S}) \mathbf{y}_G^*$$

Thus, Equation (15) is obtained. By using $\hat{\beta}$, and the corresponding smoothing matrix \mathbf{S}_h , a modified local polynomial estimator of \mathbf{f} is given by Equation (16).

Appendix A.3. Proof of Theorem 1

To obtain the ridge-penalty-based local polynomial estimators, the key point is to calculate the partial residuals and minimize (18). From that, as mentioned before, $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S})\mathbf{X}$ and $\tilde{\mathbf{y}}_G^* = (\mathbf{I} - \mathbf{S})\mathbf{y}_G^*$ are calculated, and the minimization of (18) is realized based on the synthetic response variable $\tilde{\mathbf{y}}_G^*$ as below:

$$\begin{aligned} PLS_{RL} &= \left(\tilde{\mathbf{y}}_G^* - \tilde{\mathbf{X}}\beta\right)' \left(\tilde{\mathbf{y}}_G^* - \tilde{\mathbf{X}}\beta\right) + \lambda \beta' \beta \\ &= \tilde{\mathbf{y}}_G^{*T} \tilde{\mathbf{y}}_G^* - \beta^T \tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_G^* - \beta^T \tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_G^* + \beta^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \beta + \lambda \beta' \beta \end{aligned}$$

If, $\frac{\partial L}{\partial \beta} = 0$

$$-2\tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_G^* + 2\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \beta + 2\lambda \beta = 0$$

$$\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \boldsymbol{\beta} + \lambda \boldsymbol{\beta} = \tilde{\mathbf{y}}_G^{*T} \tilde{\mathbf{X}}$$

$$\hat{\boldsymbol{\beta}}_{RL} = \left(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{y}}_G^*$$

References

- Speckman, P. Kernel smoothing in partial linear models. *J. R. Stat. Soc. Ser. B* **1988**, *50*, 413–436. [\[CrossRef\]](#)
- Green, P.J.; Silverman, B.W. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*; Crc Press: Boca Raton, FL, USA, 1993.
- Ruppert, D.; Wand, M.P.; Carroll, R.J. *Semiparametric Regression (No. 12)*; Cambridge University Press: Cambridge, UK, 2003.
- Miller, R.G. Least squares regression with censored data. *Biometrika* **1976**, *63*, 449–464. [\[CrossRef\]](#)
- Buckley, J.; James, I. Linear regression with censored data. *Biometrika* **1979**, *66*, 429–436. [\[CrossRef\]](#)
- Koul, H.; Susarla, V.; Van Ryzin, J. Regression Analysis with Randomly Right-Censored Data. *Ann. Stat.* **1981**, *9*, 1276–1288. [\[CrossRef\]](#)
- Lai, T.; Ying, Z.; Zheng, Z. Asymptotic Normality of a Class of Adaptive Statistics with Applications to Synthetic Data Methods for Censored Regression. *J. Multivar. Anal.* **1995**, *52*, 259–279. [\[CrossRef\]](#)
- Fan, J.; Gijbels, I. Censored regression: Local linear approximations and their applications. *J. Am. Stat. Assoc.* **1994**, *89*, 560–570. [\[CrossRef\]](#)
- Guessoum, Z.; Saïd, E.O. Kernel regression uniform rate estimation for censored data under α -mixing condition. *Electron. J. Stat.* **2010**, *4*, 117–132. [\[CrossRef\]](#)
- Orbe, J.; Ferreira, E.; Núñez-Antón, V. Censored partial regression. *Biostatistics* **2003**, *4*, 109–121. [\[CrossRef\]](#)
- Qin, G.; Jing, B.-Y. Censored Partial Linear Models and Empirical Likelihood. *J. Multivar. Anal.* **2001**, *78*, 37–61. [\[CrossRef\]](#)
- Aydin, D.; Yilmaz, E. Modified estimators in semiparametric regression models with right-censored data. *J. Stat. Comput. Simul.* **2018**, *88*, 1470–1498. [\[CrossRef\]](#)
- Raheem, S.E.; Ahmed, S.E.; Doksum, K.A. Absolute penalty and shrinkage estimation in partially linear models. *Comput. Stat. Data Anal.* **2012**, *56*, 874–891. [\[CrossRef\]](#)
- Ahmed, S.E. *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation (p. 2014)*; Springer: New York, NY, USA, 2014.
- Stute, W. Consistent Estimation Under Random Censorship When Covariables Are Present. *J. Multivar. Anal.* **1993**, *45*, 89–103. [\[CrossRef\]](#)
- Stute, W. Nonlinear censored regression. *Statistica Sinica* **1999**, *9*, 1089–1102.
- Wei, L.J.; Ying, Z.; Lin, D.Y. Linear regression analysis of censored survival data based on rank tests. *Biometrika* **1990**, *77*, 845–851. [\[CrossRef\]](#)
- Kaplan, E.L.; Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Stat. association* **1958**, *53*, 457–481. [\[CrossRef\]](#)
- Fan, J.; Gasser, T.; Gijbels, I.; Brockmann, M.; Engel, J. Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency. *Ann. Inst. Stat. Math.* **1997**, *49*, 79–99. [\[CrossRef\]](#)
- Heckman, N.E. Spline smoothing in a partly linear model. *J. R. Stat. Soc. Ser. B* **1986**, *48*, 244–248. [\[CrossRef\]](#)
- Rice, J. Convergence rates for partially splined models. *Stat. Probab. Lett.* **1986**, *4*, 203–208. [\[CrossRef\]](#)
- Ahmed, E.S.; Raheem, E.; Hossain, S. Absolute penalty estimation. In *International Encyclopedia of Statistical Science*; Springer: Berlin/Heidelberg, Germany, 2011.
- Frank, L.E.; Friedman, J.H. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* **1993**, *35*, 109–135. [\[CrossRef\]](#)
- Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [\[CrossRef\]](#)
- Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Stat. Assoc.* **2006**, *101*, 1418–1429. [\[CrossRef\]](#)
- Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [\[CrossRef\]](#)
- Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301–320. [\[CrossRef\]](#)
- Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [\[CrossRef\]](#) [\[PubMed\]](#)
- Aydin, D.; Ahmed, S.E.; Yilmaz, E. Right-Censored Time Series Modeling by Modified Semi-Parametric A-Spline Estimator. *Entropy* **2021**, *23*, 1586. [\[CrossRef\]](#)
- Hurvich, C.M.; Simonoff, J.S.; Tsai, C.-L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Ser. B* **1998**, *60*, 271–293. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)