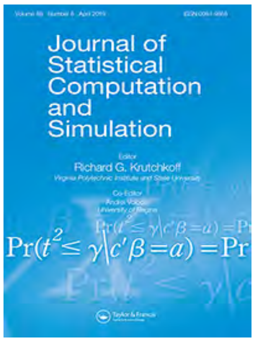

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Estimation of semiparametric regression model with right-censored high-dimensional data

Dursun Aydın, S. Ejaz Ahmed & Ersin Yılmaz

To cite this article: Dursun Aydın, S. Ejaz Ahmed & Ersin Yılmaz (2019) Estimation of semiparametric regression model with right-censored high-dimensional data, Journal of Statistical Computation and Simulation, 89:6, 985-1004, DOI: 10.1080/00949655.2019.1572757

To link to this article: https://doi.org/10.1080/00949655.2019.1572757

Published online: 28 Jan 2019.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2019, VOL. 89, NO. 6, 985–1004
https://doi.org/10.1080/00949655.2019.1572757

Estimation of semiparametric regression model with right-censored high-dimensional data

Dursun Aydın[a], S. Ejaz Ahmed[b] and Ersin Yılmaz[a]

[a]Department of Statistics, Faculty of Science, Mugla Sitki Kocman University, Mugla, Turkey; [b]Department of Mathematics and Statistics, Brock University, St. Catharines, ON, Canada

ABSTRACT

In this paper, we consider the estimation problem for the semiparametric regression model with censored data in which the number of explanatory variables p in the linear part is much larger than sample size n, often denoted as $p \gg n$. The purpose of this paper is to study the effects of covariates on a response variable censored on the right by a random censoring variable with an unknown probability distribution. It should be noted that high variance and over-fitting are a major concern in such problems. Ordinary statistical methods for estimation cannot be applied directly to censored and high-dimensional data, and therefore a transformation is required. In the context of this paper, a synthetic data transformation is used for solving the censoring problem. We then apply the LASSO-type double-penalized least squares (DPLS) to achieve sparsity in the parametric component and use smoothing splines to estimate the nonparametric component. A Monte Carlo simulation study is performed to show the performance of the estimators and to analyse the effects of the different censoring levels. A real high-dimensional censored data example is used to illustrate the ideas discussed herein.

ARTICLE HISTORY
Received 16 January 2019
Accepted 17 January 2019

KEYWORDS
High-dimensional data; right-censored data; smoothing spline; lasso; double-penalized least squares; semiparametric models

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
62N01; 62J07; 62H12

1. Introduction

In this paper, we are interested in a censored semiparametric model with a divergent number of covariates. In order to better understand the censoring mechanism, let y_i, c_i, and $\{\mathbf{x}_i,\ t_i\}$ be the survival times, the censoring times and their associated explanatory variables, respectively. Correspondingly, let $z_i = \min(y_i,\ c_i)$ be the observed survival times and $\delta_i = I(y_i \leq c_i)$ be the censoring indicator. Here, δ_i indicates whether the survival time (or lifetime) y_i corresponds to an event ($\delta_i = 1$) or is censored ($\delta_i = 0$), and z_i is equal to y_i, if the survival time is observed, and to c_i if it is censored. In this case, a convenient way to analyse the relationship between $\mathbf{y} =(y_1, \ldots, y_n)$ and $(\mathbf{x},\ t)$ in a statistical framework is required to consider the following observed data

$$\{(\mathbf{x}_i,\ t_i, z_i,\ \delta_i),\ i = 1, \ldots, n\} \tag{1}$$

CONTACT Ersin Yılmaz ✉ yilmazersin13@hotmail.com Department of Statistics, Faculty of Science, Mugla Sitki Kocman University, 48000 Mugla, Turkey

Given i.i.d observations (1), we suppose that the data can be described using a semiparametric model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + f(t_i) + \varepsilon_i, 1 \leq i \leq n \tag{2}$$

where $y_i{}'s$ are the observations of the response variable, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and $t_i{}'s$ are the observations of the explanatory variable, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown p-dimensional vector of parameters to be estimated, $f(.)$ is an unknown univariate smooth function, and $\varepsilon_i{}'s$ are supposed to be uncorrelated random variables with mean zero and a common variance σ^2, and independent of the explanatory variables. For notational simplicity, t_i is scalar and takes values in [0, 1] and the intercept term is not included. However, it is possible to achieve a model without intercept can by centring the variables. We should also note that the vector of response variable $\mathbf{y}$ depends parametric linearly on the vector of explanatory variables $\mathbf{x}_i$ and nonlinearly on a scalar variable t.

Generally speaking, when the number of parametric effect p is fixed (or $p < n$), the estimation of parametric and nonparametric components in model (1) with uncensored data have been studied in various investigations including smoothing spline [1–3], kernel smoothing [4], and regression spline [5] Similarly, a number of authors have studied the case of semiparametric regression model based on censored data. More detailed discussions are available in numerous studies, such as Orbe et al. [6], and Aydin and Yilmaz [7] among others.

With recent developments in science and technology, high-dimensional data has become of increasing importance, especially in medical studies, genomics and some areas of computational biology. In this context, many applications are constructed for possibly sparse models in high-dimensional settings when p is not fixed (often written as $p \gg n$). It is important to remember that when p increases with the increase of the sample size n, the sparsity of the true model is commonly assumed. Sparsity states that some explanatory variables do not contribute to the response variable, in the sense that some parametric coefficients in the model (2) are exactly zero. For example, Xie and Huang [8], Gao et al. [9], and Cheng et al. [10] are mainly focused on statistical inference for the coefficients in the linear part of the model (2). It should be noted that the studies given above use uncensored data.

In this paper, we study the high-dimensional semiparametric model with right-censored data. Our main contribution is to modify the LASSO-type penalty for high-dimensional censored data case with double-penalized least squares (DPLS), proposed in Ni et al. [11], and obtain an estimator that can deal with extra difficulties caused by the high-dimensional censored data and the nonlinear part of the model. It should be noted that this type censored data has drawn much attention in the past decade, especially for variable selection in a semiparametric model (see Ma and Du [12], for a detailed discussion of this topic). Furthermore, various penalization procedures have been proposed for uncensored data, such as the least absolute shrinkage and selection operator (LASSO, proposed in [13]), the smoothly clipped absolute deviation (SCAD, discussed in [14]), minimax concave penalty (MCP, examined in [15]), least angle regression (LARS, stated in [16]), and adaptive LASSO [17].

The rest of this paper is organized as follows: In Section 2, we discuss the required conditions and the model description and motivation. In Section 3, we derive the estimation of the right-censored high-dimensional semiparametric model using the DPLS method based on smoothing spline. Section 4 introduces the selection of the penalty parameters. The simulation results and a real data application are expressed in Section 5. Lastly, we present our concluding remarks and recommendations in Section 6.

2. Preliminaries

Suppose that the probability distribution functions of the survival times (y_i) and censoring times (c_i) are denoted with F and G, respectively. In other words, the unknown distribution function of y_i can be expressed as $F(t) = P(y_i \leq s)$ and c_i can be stated as $G(t) = P(c_i \leq s)$, respectively. The significance of the model depends on some specific assumptions on the response, censoring and explanatory variables which are defined by Stute [18] and explained as follows

Assumption 1: y_i and c_i are independent
Assumption 2: $P(y_i \leq c_i | y_i, \mathbf{x}_i, t_i) = P(y_i \leq c_i | y_i)$

Note that these assumptions are commonly used in survival analysis applications. Assumption 1 is an ordinary independence condition to support the accuracy of the model with censored data. If Assumption 1 is violated, then more information about the dataset is required to obtain a proper model. Assumption 2 is needed to allow for a dependency between (x_i, t_i) and c_i. More explicitly, Assumption 2 says that given time of death, covariates do not provide any further information whether the observation is censored or not. See Stute [19], Heuchenne and Van Keilegom [20] and Zhou [21] for more details on these assumptions of the survival data analysis.

As indicated in the introduction section of this paper, the response variable is observed incompletely, but the remaining other variables are observed completely. In this case, ordinary statistical methods cannot be applied directly to this type of observations, and data transformation is required. Under censorship, instead of using responses y_i alone, we consider the pairs of observations $\{(z_i, \delta_i), i = 1, \ldots, n\}$. For context, Koul et al. [22] denoted that when G is continuous and known, it is possible to adjust observed lifetimes z_i to yield an unbiased modification

$$y_{iG} = \frac{\delta_i z_i}{1 - G(z_i)}, i = 1, 2, \ldots, n \tag{3}$$

where y_{iG} has the same mean as y_i. In this sense, the aforementioned assumptions are also used to provide that $E[y_{iG}|\mathbf{x}_i, t_i] = E[y_i|\mathbf{x}_i, t_i] = \mathbf{x}_i\boldsymbol{\beta} + f(t_i)$. It should be noted that $\{y_{iG} = (y_{1G}, \ldots, y_{nG})'\} = \mathbf{y}_G$ is the vector of transformed responses. In most practices, however, distribution (i.e. G) of the censoring variable given in (3) is unknown. In order to solve this problem, Koul et al. [22] proposed to replace G by its Kaplan–Meier [23] estimator, given by

$$1 - \hat{G}(s) = \prod_{i=1}^{n} \left(\frac{n-i}{n-i+1} \right)^{\mathrm{I}[z_{(i)} \leq s, \delta_{(i)} = 0]}, \quad s \geq 0 \tag{4}$$

where $z_{(1)} \leq, \ldots, \leq z_{(n)}$ are the ordered values of observed response variable z and $\delta_{(i)}$ is the corresponding censoring indicator associated to $z_{(i)}$.

For a given smoothing parameter $\lambda > 0$ and a positive-definite (symmetric) smoother matrix $\mathbf{S}_\lambda$, the corresponding smoothing spline (ss) estimators for $\boldsymbol{\beta}$, based on model (2) with censored data, can be defined as (see Aydin and Yilmaz [7] for a detailed discussion):

$$\hat{\boldsymbol{\beta}}_{ss} = (\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}_{\hat{G}} \tag{5}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ and $\mathbf{y}_{\hat{G}} = \{(y_{1\hat{G}}, \ldots, y_{n\hat{G}}) = y_{i\hat{G}}\} = \delta_i z_i / 1 - \hat{G}(z_i), i = 1, 2, \ldots, n$. We should also note that the response $\mathbf{y}_{\hat{G}}$ may also be called as synthetic response variable since the values of this variable are synthesized from the data (z_i, δ_i) to fit the semiparametric model $E[y_{i\hat{G}}|\mathbf{x}_i, t_i] = \mathbf{x}_i\boldsymbol{\beta} + f(t_i)$. In a similar fashion to the linear model case, the assumptions given above ensure that $E[y_{i\hat{G}}|\mathbf{x}_i, t_i] = E[y_i|\mathbf{x}_i, t_i] = \mathbf{x}_i\boldsymbol{\beta} + f(t_i)$.

Note that the ideas expressed in the above paragraph are designed for estimating the censored semiparametric model where p is assumed to be small relative to n. However, our claim is to establish statistical inference for the high-dimensional parametric coefficients $\boldsymbol{\beta}$ in presence of a univariate smooth function f. If the number of parametric effect p is larger than sample size n, ordinary statistical methods in general are not applicable to the semiparametric model with a high-dimensional parametric component. Obviously, when $p > n$, the estimator defined in (5) does not have a unique solution and its predictive accuracy will be low due to over-fitting, as in the linear regression case. Such problems need a form of complexity regularization to get the optimal solution. To overcome this problem, we follow the suggestions in the study of Ni et al. [11] by modifying the DPLS approach. It is understood that the resulting regularization problem can be solved by a LASSO-type DPLS method. Before proving this matter, we will briefly offer some ideas to solve a semiparametric regression problem.

2.1. Model specification and motivation

A formal connection between semiparametric and linear models can be constructed through a right-censored response variable y. When $f(.) = 0$ in the model (2) with high-dimensional parametric coefficients, this model reduces to the following linear regression model:

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \quad 1 \le i \le n \tag{6}$$

Note that model (6) contains the unknown high-dimensional parametric coefficients that need to be estimated in practice. We approximate $E[y_{i\hat{G}}|\mathbf{x}_i] = E[y_i|\mathbf{x}_i] = \mathbf{x}_i\boldsymbol{\beta}$ by LASSO, introduced by Tibshirani [13]. The LASSO estimates of the parametric coefficients in the model (6) are obtained by minimizing the L_1 -penalized objective function in

$$\hat{\boldsymbol{\beta}}(\lambda_2) = \underbrace{\text{argmin}}_{\beta}(\| y_{i\hat{G}} - \mathbf{x}_i\boldsymbol{\beta} \|_2^2 + \lambda_2 \| \boldsymbol{\beta}\|_1) \tag{7}$$

where $\lambda_2 \ge 0$ is a positive penalty parameter that controls the amount of shrinkage applied to the estimates. As $\lambda_2 \to \infty$, penalty dominates in (7) and the resulting LASSO estimates will be shrunk to zero. On the other hand, as $\lambda_2 \to 0$, penalty disappears and results in little shrinkage. Of course, for $\lambda_2 = 0$, there is no shrinkage at all. Also, Equation (7) suggests that the LASSO achieves variable selection and shrinkage at the same time. However, this result is limited in the parametric models.

In this paper, we are mainly interested in estimating the parametric and nonparametric components of a censored semiparametric model when the number of parametric variables p increases with the sample size n. Note that the estimation procedure for this type of a model is more challenging because it consists of several interrelated estimation and selection problems, such as nonparametric estimation, penalty parameter selection, and estimation for parametric linear variables. Müller and van de Geer [24] provide us with an appropriate estimator by altering the methods used in Mammen and van de Geer [25] for the low-dimensional case with the standard LASSO, to make them applicable uncensored data.

As stated in the previous sections, when the response variable is censored by a random variable c, the model (2) transforms to the following censored model

$$y_{i\hat{G}} = \mathbf{x}_i \boldsymbol{\beta}_n + f(t_i) + \varepsilon_{i\hat{G}}, \quad 1 \leq i \leq n \tag{8}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) = \mathbf{X}_n$ is an $n \times p$ matrix, $\boldsymbol{\beta}_n$ is the $p \times 1$ vector of parametric coefficients expressed before, and $\varepsilon_{i\hat{G}}{}'s$ are identical, but not independent, random error terms with unknown constant variance.

Remark 2.1: In this paper, we consider right-censored high-dimensional data; the number of parametric variables affecting the response variable is larger than the number of response observations. In this case, model (8) is considered as a sparse model. The idea behind this model is that p covariates are categorized into two groups: the important ones whose corresponding coefficients are nonzero and the trivial regression coefficients that actually are (nearly) zero and not present in the underlying model.

Note that the main purpose of this paper is to estimate the parametric effects and the unknown smooth function f by controlling the sparsity of the vector $\boldsymbol{\beta}_n$ in a high-dimensional setting**.** To achieve this, we follow an estimation procedure based on *DPLS* (proposed in Ni et al. [11]). It is emphasized that the estimators of $\boldsymbol{\beta}_n$ and $(f(t_1), \ldots, f(t_n))' = \mathbf{f}$ can be obtained by minimizing the penalized least squares objective function

$$L(\boldsymbol{\beta}_n, f(.)) = \sum_{i=1}^{n} \{y_{i\hat{G}} - \mathbf{x}_i\boldsymbol{\beta}_n - f(t_\mathbf{i})\}^2 + n\lambda_1 \int_0^1 \{f''(t)\}^2 \mathrm{d}t + 2n \sum_{j=1}^{p} \lambda_{2|}\beta_j| \tag{9}$$

In Equation (9), the first penalty term weighted by $\lambda_1 \geq 0$ denotes the roughness penalty and it imposes a penalty on the roughness of nonparametric fit $f(t)$. The second penalty term multiplied by $\lambda_2 \geq 0$ indicates a shrinkage penalty and it applies shrinkage to the slope coefficients of the regression model, but not the intercept. Note that λ_1 is a smoothing parameter that plays a key role in controlling the trade-off between the smoothness of $f(t)$ with fidelity to data, whereas λ_2 is a regularization parameter that controls the amount of shrinkage used in determining the parametric effects. To provide effective estimation it is necessary to select an optimum amount of these penalty parameters. These parameters are discussed in section 3.

In practice, there have been several studies on various regularization approaches, such as Elastic Net (discussed in [26]), Fused Lasso (studied in Tibshirani et al. [27], Adaptive lasso (examined in [17]), spline-lasso (discussed in [28]) to handle minimization problem (9) for $p \gg n$, and to avoid the over-fitting. In this paper, however, we use smoothing spline method to solve minimization of the L_1 penalty in (9). In this sense, the computation of the (9) can be achieved by a quadratic programming and an optimally designed algorithm, given in Section (4).

3. Solution of DPLS problem based on smoothing spline

We now introduce the smoothing spline solutions for $\boldsymbol{\beta}$ and $\mathbf{f}$ in the model (2) with right-censored high-dimensional data. Let $v_1 < v_2 < \ldots < v_q$ be the distinct and ordered values among $t_1, t_2, \ldots, t_n$. The connection between v's and t's is provided by nxq incidence matrix $\mathbf{N}$, with elements $N_{ij} = 1$ if $t_i = v_j$ and $N_{ij} = 0$ if $t_i \neq v_j$. In the light of these ideas, we also suppose that $\mathbf{f} = f(v_j) = (a_1, \ldots, a_q)$ is a vector. Then, in matrix and vector form, penalized least squares function (9) for estimating $\boldsymbol{\beta}_n$ and $\mathbf{f}$ can be rewritten as

$$L(\boldsymbol{\beta}_n, \mathbf{f}_n) = \| \mathbf{y}_{\hat{G}} - \mathbf{X}_n\boldsymbol{\beta}_n - \mathbf{N}\mathbf{f}_n \|_2^2 + n\lambda_1 \int_0^1 \{f''(t)\}^2 \mathrm{d}t + 2n\sum_{j=1}^{p} \lambda_2 |\beta_j| \tag{10}$$

Given $\lambda_1 > 0$, the smoothness of nonparametric component in (8) is regularized by a roughness penalty term $n\lambda_1 \int f''(t)^2 dt$ for $\lambda_1 > 0$.

Remark 3.1: If t is an $n \times 1$ dimensional vector (i.e. $t \in \mathbb{R}$), the L_2- norm of the second derivative $\int_{\mathbb{R}} (f''(t))^2 dt$ in Equation (10) satisfies the quadratic form $\mathbf{f}'\mathbf{K}\mathbf{f}$ (see [3] for a detailed discussion). This case denotes that the roughness penalty term is equal to the following notation:

$$\int_{\mathbb{R}} (f''(t))^2 dt = \mathbf{f}'\mathbf{K}\mathbf{f} \tag{11}$$

where $\mathbf{K}$ a symmetric $q \times q$ positive definite penalty matrix and its elements are computed by means of the knot points $v_1, \ldots, v_q$, and defined by

$$\mathbf{K} = \mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q} \tag{12}$$

where $\mathbf{Q}$ and $\mathbf{R}$ are the tri-diagonal matrices with dimensions $(q-2) \times q$ and $(q-2) \times (q-2)$, respectively. Their entries are obtained by $Q_{i,i} = 1/h_i$, $Q_{i,i+1} = -\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right)$, $Q_{i,i+2} = 1/h_{i+1}$, and $R_{i-1,i} = R_{i,i-1} = h_i/6$, $R_{i,i} = (h_i + h_{i+1})/3$ where $h_i = v_{i+1} - v_i, i = 1, \ldots, q-1$.

From these facts, it is easily seen that the DPLS criterion can be rewritten as

$$L(\boldsymbol{\beta}_n, \mathbf{f}_n) = \| \mathbf{y}_{\hat{G}} - \mathbf{X}_n\boldsymbol{\beta}_n - \mathbf{N}\mathbf{f}_n \|_2^2 + n\lambda_1 \mathbf{f}_n'\mathbf{K}\mathbf{f}_n + 2n\sum_{j=1}^{p} \lambda_2 |\beta_j| \tag{13}$$

By taking simple algebraic operations, one can see that given λ_1 and vector $\boldsymbol{\beta}_n$, the DPLS solution of nonparametric component $(\mathbf{f}_n = f(t_1), \ldots, f(t_n))'$ based on the smoothing spline can be obtained as

$$\hat{\mathbf{f}}_n(\boldsymbol{\beta}_n) = (\mathbf{N}'\mathbf{N} + n\lambda_1\mathbf{K})^{-1}\mathbf{N}'(\mathbf{y}_{\hat{G}} - \mathbf{X}_n\boldsymbol{\beta}_n) = \mathbf{S}_{\lambda_1}(\mathbf{y}_{\hat{G}} - \mathbf{X}_n\boldsymbol{\beta}_n) \tag{14}$$

where $\mathbf{S}_{\lambda_1} = (\mathbf{N}'\mathbf{N} + n\lambda_1\mathbf{K})^{-1}\mathbf{N}'$ is a positive-definite linear smoother matrix which depends on λ_1. It should be noted that when t_i are distinct and ordered already, $\mathbf{N} = \mathbf{I}$ and $\mathbf{S}_{\lambda_1}$ transforms to the following smoothing matrix: $\mathbf{S}_{\lambda_1} = (\mathbf{I} + n\lambda_1\mathbf{K})^{-1}$ where $\mathbf{I}$ is an $n \times n$ identity matrix. More specifically, it must be emphasized that the matrix $\mathbf{S}_{\lambda_1}$ is obtained from model (8) with $\boldsymbol{\beta}_n = 0$, and it transforms the vector of response observations into the fitted values $\hat{\mathbf{y}}_{\hat{G}} = \mathbf{S}_{\lambda_1}\mathbf{y}_{\hat{G}} = \{\hat{f}_{\lambda_1}(t_1), \ldots, \hat{f}_{\lambda_1}(t_\text{n})\} = \hat{\mathbf{f}}_{n(\lambda_1)}$.

When we substitute the $\hat{\mathbf{f}}_n(\boldsymbol{\beta}_n)$ into the criterion (13), we obtain the L_1 -penalized least squares function for only vector $\boldsymbol{\beta}_n$ **:**

$$L(\boldsymbol{\beta}_n) = \| \tilde{\mathbf{y}}_{\hat{G}} - \tilde{\mathbf{X}}_{\boldsymbol{n}}\boldsymbol{\beta}_n - \mathbf{N}\mathbf{f}_n \|_2^2 + 2n \sum_{j=1}^{p} \lambda_2 |\beta_j| \tag{15}$$

where $\tilde{\mathbf{X}}_n = (\mathbf{I} - \mathbf{S}_{\lambda_1})\mathbf{X}_n$ and $\tilde{\mathbf{y}}_{\hat{G}} = (\mathbf{I} - \mathbf{S}_{\lambda_1})\mathbf{y}_{\hat{G}}$. Or, equivalently, the for an appropriate parameter λ, Equation (15) can be rewritten as

$$L(\boldsymbol{\beta}_n) = \| \tilde{\mathbf{y}}_{\hat{G}} - \tilde{\mathbf{X}}_{\boldsymbol{n}}\boldsymbol{\beta}_n - \mathbf{N}\mathbf{f}_n \|_2^2 \quad \text{subject to } 2n \sum_{j=1}^{p} |\beta_j| \leq \lambda \tag{16}$$

As can be seen from Equations (15) and (16), the DPLS problem reduces to the standard LASSO-type regression problem. Note that the parameter λ in (16) controls the number of non-zero coefficients β_j, and the DPLS estimator results in fewer than p non-zero coefficients**.** In this case**,** the parameter λ is related to the sparse solutions of parametric coefficients vector $\boldsymbol{\beta}_n$.

The LASSO regression provides solutions to the penalized least squares function given in Equations (15) and (16). However, we expect that many of the LASSO estimates should be zero, and hence, seek a set of sparse solutions. Let $\hat{\beta}_j^{\text{ols}}$ be the full ordinary least squares estimates and let $\lambda_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{\text{ols}}|$. For example, if $\lambda_0 = \sum_{j=1}^{p} |\hat{\beta}_j^{\text{ols}}|$ or equivalently $\lambda = 0$, we obtain no shrinkage, and therefore obtain the least squares solutions. Additionally, the constraint $\sum_{j=1}^{p} |\beta_j| \leq \lambda$ in (5) denotes that we have a 'path' of solutions indexed by λ. This means that the values $\lambda < \lambda_0$ will cause shrinkage of the solutions leading to zero, and some coefficients may be exactly equal to zero. It should be noted that the path of LASSO solutions is indexed by a component of shrinkage penalty λ_0. For example, if $= \lambda_0/2$, the effect will be roughly similar to finding the best subset of size $p/2$, as indicated in Tibshirani [13]. For these reasons, it is very important to determine the estimation of parameter λ. We explain this case in more detail in Section 4.

As can be seen from Equations (15) and (16), the DPLS problem reduces to the standard LASSO-type problem. It should be noted that unlike the study of Ni et al. [11], we use ridge penalty instead of a SCAD penalty to determine the shrinkage penalties in Equations (15) and (16). In this paper, however, we have constantly emphasized that the number p of parameters is much larger than n. For this reason, we only seek to find a technique to eliminate most of the parameters, and reduce to a case with a low-dimensional structure that is useful for our estimation problem. That is to say, we want to explain a regression problem with large and complex structures, in which most of the parameters are unimportant, and focus instead on the subset of important regression parameters. Recent developments provide efficient variable selection algorithms, such as LASSO and LARS. Inspired by LASSO, we adopt a newly computational algorithm to obtain a solution of DPLS criterion described in (15).

Remark 3.2: In this paper, we consider the estimator $\hat{\boldsymbol{\beta}}_n$, which minimizes the least square objective function in Equations (15) or (16). Without loss of generality, we suppose that the true important coefficient index set $\mathrm{V} = \{1, 2, \ldots, q\}$, where q is an integer and $1 \leq q \leq p$. Therefore, based on the partition of the data matrix $\tilde{\mathbf{X}}_n = (\tilde{\mathbf{X}}_{1n}, \tilde{\mathbf{X}}_{2n})$ **,** we have true parametric coefficients vector $\boldsymbol{\beta}_n = (\boldsymbol{\beta}'_{1n}, \boldsymbol{\beta}'_{2n})'$, where $\boldsymbol{\beta}'_{1n}$ related to the $\tilde{\mathbf{X}}_{1n}$ contains the first q nonzero important coefficients, and $\boldsymbol{\beta}'_{2n}$ associated with $\tilde{\mathbf{X}}_{2n}$ contains the remaining unimportant parametric coefficients.

Computational Algorithm

Input: Data matrix $\in \mathbb{R}^{n\times p}$, data vector $t \in \mathbb{R}^{n\times 1}$, and response vector $\mathbf{y} \in \mathbb{R}^{n\times 1}$
Step 1. Solve Equation (3) to obtain the synthetic response vector $\mathbf{y}_{\hat{G}}$
Step 2. Select an appropriate roughness penalty λ_1 using the GCV criterion, and compute the smoother matrix $\mathbf{S}_{\lambda_1}$, as defined in (14): $\mathbf{S}_{\lambda_1} = (\mathbf{N}'\mathbf{N} + n\lambda_1\mathbf{K})^{-1}\mathbf{N}'$, and define the matrix and vectors based on residuals $\tilde{\mathbf{X}}_n = (\mathbf{I} - \mathbf{S}_{\lambda_1})\mathbf{X}_n$ and $\tilde{\mathbf{y}}_{\hat{G}} = (\mathbf{I} - \mathbf{S}_{\lambda_1})\mathbf{y}_{\hat{G}}$.
Step 3. Determine the penalty tuning parameter λ by GCV criterion given in (21)
Step 4. To eliminate unimportant variables in the L_1 -penalty constraint (16), follow the SAFE rule proposed by El Ghaoui et al. [29]:

(i) Discard the inactive predictor variables by using the condition

$$|\tilde{\mathbf{x}}_j'\tilde{\mathbf{y}}_{\hat{G}}| < \lambda - \parallel \tilde{\mathbf{X}}_n\|_2 \parallel \tilde{\mathbf{y}}_{\hat{G}}\|_2 \left(\frac{\lambda_{\max} - \lambda}{\lambda_{\max}}\right)$$

where $\tilde{\mathbf{x}}_j \in \mathbb{R}^n, j = 1, 2, \ldots, p$, the j -th column of $\tilde{\mathbf{X}}_n$ and $\lambda_{\max} = \max|\tilde{\mathbf{x}}_j'\tilde{\mathbf{y}}_{\hat{G}}| = \parallel \tilde{\mathbf{x}}_j'\tilde{\mathbf{y}}_{\hat{G}}\|_\infty$, which implies that all parametric coefficients estimates are zero (complete shrinkage to 0). Tibshirani et al. [30] modified this SAFE rule by replacing $\tilde{\mathbf{X}}_n\|_2 \parallel \tilde{\mathbf{y}}_{\hat{G}}\|_2/\lambda_{\max}$ with 1, making the equation read

$$|\tilde{\mathbf{x}}_j'\tilde{\mathbf{y}}_{\hat{G}}| < 2\lambda - \lambda_{\max}$$

This rule discards more predictor variables than the SAFE rule; this rule is used because in this study, the number of parameters p is considerable. Note that this rule provides substantial computational time savings for the estimation process.

(ii) After the ith case of step 4, partition the remaining variables in form $\tilde{\mathbf{X}}_n = (\tilde{\mathbf{X}}_{1n}, \tilde{\mathbf{X}}_{2n})$**,** as defined in Remark 3.2
(iii) Find the LASSO estimates of $\boldsymbol{\beta}'_{1n}$ associated with the $\tilde{\mathbf{X}}_{1n}$ contains the first q nonzero important coefficients.

Step 5. Estimate the nonparametric part of the censored semiparametric model:

$$\hat{\mathbf{f}}_n(\hat{\boldsymbol{\beta}}_{1n}) = (\mathbf{N}'\mathbf{N} + n\lambda_1\mathbf{K})^{-1}\mathbf{N}'(\mathbf{y}_{\hat{G}} - \mathbf{X}_{1n}\hat{\boldsymbol{\beta}}_{1n}) = \mathbf{S}_{\lambda_1}(\mathbf{y}_{\hat{G}} - \mathbf{X}_{1n}\hat{\boldsymbol{\beta}}_{1n})$$

Output: $\hat{\boldsymbol{\beta}}_n = \{\hat{\boldsymbol{\beta}}'_{1n}, \hat{\boldsymbol{\beta}}'_{2n}\} \in \mathbb{R}^{p\times 1}$ and $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\beta}}_n) = \{\hat{\mathbf{f}}_n(\hat{\boldsymbol{\beta}}'_{1n}), \hat{\mathbf{f}}_n(\hat{\boldsymbol{\beta}}'_{2n})\} \in \mathbb{R}^{n\times 1}$.

3.1. Asymptotical properties of DPLS estimator

In this section, we introduce a framework for establishing the asymptotic efficiency of the DPLS estimator in a high-dimensional setting. Asymptotic efficiency is first considered by van de Geer et al. [31], using linear models. In addition, Van Der Vaart [32], illustrates the efficiency bounds for a semiparametric model for fixed p (independent from n). Ni et al. [11], Jankova and van de Geer [33], study asymptotic properties of high-dimensional partially linear models based on L_1 -penalty.

A key feature of the estimation problem expressed in this paper is that the optimal rate can be achieved with respect to the sparsity parameter. Jankova and van de Geer [33], denoted that the minimax rates for the estimation (or *DPLS* estimator) of regression coefficients are shown to satisfy

$$\inf_{\hat{\beta}} \sup_{\beta} E|\hat{\beta}_i - \beta_i| \geq C\left(\frac{1}{\sqrt{n}} + s_n \frac{\log(p)}{n}\right), \quad i = 1, \ldots, p \tag{17}$$

where $C > 0$ is a constant, $\hat{\beta}_i$ is the estimator of the single regression coefficient β_i and s_n is the sparsity parameter that denotes the number of non-zero elements in the regression coefficients vector. Normally, Equation (17) implies that the DPLS method with a suitable selection of the smoothing parameter (λ_2) provides an optimal parametric rate of convergence $s_n \frac{\log(p)}{n}$ over the set of s_n -sparse regression coefficient vectors with $s_n \leq C\frac{n}{\log(p)}$. This means that the estimator $\hat{\beta}$ estimates the sparsity parameter s_n at minimax rate. Conversely, if there is deficient sparsity regime, the minimax lower bounds diverge, in particular when sparsity satisfies $s_n \gg n/log(p)$. This expression can be seen as the oracle inequalities for a such estimator under the condition $s_n = o(n/\log(p))$, which is actually necessary for asymptotically normal estimation. It is also noted that the optimal parametric rate cannot be provided in the moderate sparse region $\frac{\sqrt{n}}{\log(p)} \leq s_n < n/log(p)$. Furthermore, the upper bound parametric rate $\frac{1}{\sqrt{n}}$ can be obtained for estimation of single elements. As a consequence, the infimum in Equation (17) revealed that when sparsity of regression coefficients is of small order $\frac{\sqrt{n}}{\log(p)}$, parametric rate of order $\frac{1}{\sqrt{n}}$ is optimal.

In order to investigate the asymptotic behaviour of the *DPLS* estimator, we begin by introducing some notions. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' = (\boldsymbol{\beta}'_{1n}, \boldsymbol{\beta}'_{2n})$ be the true regression coefficients for the parametric component of the model where $\boldsymbol{\beta}'_{1n}$ is a q-dimensional nonzero coefficients vector and $\boldsymbol{\beta}'_{2n} = 0$ is a $r = (p\text{-}q)$-dimensional zero coefficients vector. Furthermore, we assume that $\mathbf{X}_{\boldsymbol{n}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ are independently and identically distributed with mean zero and positive definite covariance matrix

$$\mathbf{M} = (\tilde{\mathbf{X}}'_n \tilde{\mathbf{X}}_n)^{-1} = \begin{pmatrix} M_{11}^{-1} & M_{12}^{-1} \\ M_{21}^{-1} & M_{22}^{-1} \end{pmatrix} \tag{18}$$

We now provide the asymptotic theory for the DPLS estimator in terms of the estimation procedure. The study of Ni et al. [11] shows that if it is chosen the proper sequence of λ_1 and λ_2, then the DPLS estimator (i.e. $\hat{\boldsymbol{\beta}}_n$) is $\sqrt{n}$ -consistent. In other words, as $n \to \infty$, if $\lambda_1 \to 0$ and $\lambda_2 \to 0$, then there is a local minimizer estimator $\hat{\boldsymbol{\beta}}_n$ of $L(\boldsymbol{\beta}_n)$ such that $\| \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n \| = O_p\left(\sqrt{n}\right)$. They also illustrate the fact that as $n \to \infty$, if $\lambda_1 \to 0$, and $\lambda_2 \to 0$ then with probability tending to one, the local minimizer $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{1n}, \hat{\boldsymbol{\beta}}_{2n})^{\mathrm{T}}$ must satisfy: (i)

Sparsity: $\hat{\boldsymbol{\beta}}_{2n} = 0$. (ii) Asymptotic normality: $n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{1n} - \boldsymbol{\beta}_{1n}) \xrightarrow{d} N(0, \sigma^2 M_{11}^{-1})$, where σ^2 is the variance of error terms and M_{11}^{-1} is a $(q \times q)$ sub-matrix of $\mathbf{M}$, as defined in (18).

In this paragraph, we discuss the asymptotic properties of the DPLS estimator in a high-dimensional case where the number of parametric covariates, p, goes to ∞ as $n \to \infty$. For any square matrix $\mathbf{A}$, indicate its minimum and maximum eigenvalues respectively by $\Lambda_{min}(\mathbf{A})$ and $\Lambda_{max}(\mathbf{A})$. In addition to the ideas expressed in the above paragraph, the following regularity conditions are introduced to show the asymptotic properties of the DPLS estimator (see [34] and [11], for more detailed discussions).

A1. The elements of $\beta_{1n,j}$'s of the vector $\boldsymbol{\beta}_{1n}$ have to be satisfied

$$\min\{|\beta_{1n,j}|, 1 \leq j \leq q_n\}/\lambda_2 \to \infty$$

A2. Let w_1 and w_2 be constants such that

$$0 < w_1 < \Lambda_{\min}(\mathbf{M}) \leq \Lambda_{\max}(\mathbf{M}) < w_2 < \infty.$$

Note that **A1** implies the ability of the DPLS estimator on the discrimination the regression coefficients from zero. **A2** confirms that $\mathbf{M}$ is positive definite and eigenvalues of $\mathbf{M}$ are uniformly limited. It should be emphasized that under the assumptions A1 and A2, as $n \to \infty$, if $\lambda_1 \to 0$, $\lambda_2 \to 0$ and $p \to \infty$, DPLS estimator $\hat{\boldsymbol{\beta}}_n$ is a $\sqrt{n/p}$ -consistent (see [11]).

4. Choice of penalty tuning parameters

In practice, penalty parameters in Equations (15) and (16) can be chosen by any selection criterion, such as cross-validation (CV), generalized cross-validation (GCV), Bayesian information criterion (BIC), and so on. In this paper, we use GCV criterion to determine optimum penalty parameter λ_2, or equivalently, to select the parametric coefficient λ in the L_1 penalty constraint (16), $\sum_{j=1}^{p} |\beta_j| \leq \lambda$. The key idea here is to determine the number of effective parameters in constrained estimates of $\boldsymbol{\beta}$.

A closed-form estimate for the parametric coefficients can be obtained by using the penalty $\sum_{j=1}^{p} |\beta_j|$ as $\sum_{j=1}^{p} (\beta_j^2/|\beta_j|)$. Thus, the constrained estimate vector of $\boldsymbol{\beta}$ in the Equation (16) can approximate the solution by a ridge regression of the form

$$\hat{\boldsymbol{\beta}}_n = (\tilde{\mathbf{X}}_{\boldsymbol{n}}' \tilde{\mathbf{X}}_n + \lambda \mathbf{W}^-)^{-1} \tilde{\mathbf{X}}_{\boldsymbol{n}}'' \tilde{\mathbf{y}}_{\hat{G}} \tag{19}$$

where $\mathbf{W}$ is a diagonal matrix with diagonal entries $|\beta_{nj}|$, and $\mathbf{W}^-$ denotes generalized inverse of the matrix $\mathbf{W}$. Consequently, the number of effective parameters (i.e. the coefficients vector $\boldsymbol{\beta}_{1n}'$) in the constrained Equation (16) fitted $\hat{\boldsymbol{\beta}}_n$ can be defined by the trace of the hat matrix

$$p(\lambda) = \mathrm{tr}\{\tilde{\mathbf{X}}_n (\tilde{\mathbf{X}}_{\boldsymbol{n}}' \tilde{\mathbf{X}}_n + \lambda \mathbf{W}^-)^{-1} \tilde{\mathbf{X}}_n\} = \mathrm{tr}(\mathbf{H}_\lambda) \tag{20}$$

Using Equation (20), we get the GCV function

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\{\mathrm{RSS}(\lambda) = (\tilde{\mathbf{y}}_{\hat{G}} - \tilde{\mathbf{X}}_n \hat{\boldsymbol{\beta}}_n)' (\tilde{\mathbf{y}}_{\hat{G}} - \tilde{\mathbf{X}}_n \hat{\boldsymbol{\beta}}_n)\} / \frac{1}{n}\mathrm{tr}(\boldsymbol{I} - \mathbf{H}(\lambda)) \tag{21}$$

where RSS(λ) denotes the residual sum of squares for the constrained fit with constraint λ. It should also be noted that the parameter λ which minimizes Equation (20) is selected as an optimum penalty tuning parameter. Accordingly, fitted values for the censored semiparametric model are obtained as

$$\hat{\mathbf{y}}_{\hat{G}} = \tilde{\mathbf{X}}_n\hat{\boldsymbol{\beta}}_n = \mathbf{H}(\lambda)\tilde{\mathbf{y}}_{\hat{G}} = \tilde{\mathbf{X}}_n(\tilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n + \lambda\mathbf{W}^{-})^{-1}\tilde{\mathbf{X}}_n\tilde{\mathbf{y}}_{\hat{G}} \tag{22}$$

5. Simulation experiment

In this section, we conduct Monte Carlo Simulation experiments to analyse the finite sample performance of the introduced DPLS method. For different values of sample size (n) and the number of variables (p), the response observations are generated from a partially linear model

$$y_i = \mathbf{x}_i\boldsymbol{\beta}_n + f(t_i) + \varepsilon_i, i = 1, \ldots, n, \quad \varepsilon_i \sim N(0, \sigma^2 = 0.5) \tag{23}$$

In this model, the covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ are constructed from a uniform distribution. We set the true regression coefficients $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{1n} = \{1, 2, -3, 0.5, -2, 1.5, 0.3, -1, 4, 0.4\}', \boldsymbol{\beta}_{2n} = \{0, \ldots, 0\}')$ with the variance–covariance matrix Σ, and the nonparametric component f (.) is determined by the function

$$f(t_i) = t_i(\sin(t_i^2) \text{ with } t_i = 4.3(i - 0.5)/n$$

To introduce right censoring, we generate the censoring variable c_i from the normal distribution with proportions at 10% and 40%. Finally, from the model (23), we define ith indicator as $\delta_i = \mathrm{I}(y_i \leq c_i)$ and then the observed response as

$$z_i = \min(y_i, c_i)$$

Because of the censoring, ordinary methods cannot be applied directly here to estimate the parameters of this model. For this reason, we consider transformed response observations (i.e., $y_{i\hat{G}}'s$), as described in (5), to estimate the components of the model (23).

It should be noted that we conducted simulations with n = 50,100, 200, p = 5, 300,

Table 1. Finite sample performances of the proposed estimator for the parametric part of the semiparametric model with CR = 10%, 40% and 12 different (n, p) combinations, respectively.

	CR = 10%			CR = 40%		
(n, p)	MSE$_y$	T Σ_{11}	q	MSE$_y$	Σ_{11}	q
(50,5)	0.029264	0.021871	5	0.40511	0.33346	5
(50, 300)	0.00669	0.00368	26	0.06350	0.00368	28
(50, 1000)	0.00697	0.00385	25	0.09200	0.00385	27
(50, 3000)	0.00803	0.00418	27	0.20783	0.00418	17
(100,5)	0.01051	0.01334	5	0.37619	0.30893	5
(100, 300)	0.00556	0.00217	41	0.05730	0.04491	47
(100, 1000)	0.00651	0.00226	54	0.08660	0.05589	43
(100, 3000)	0.00682	0.00403	52	0.13536	0.10503	45
(200,5)	0.00939	0.01077	5	0.33280	0.27154	5
(200, 300)	0.00370	0.00161	54	0.01020	0.05381	53
(200, 1000)	0.00519	0.00205	55	0.01999	0.05231	65
(200, 3000)	0.00442	0.00305	66	0.07500	0.05530	77

1000, 3000, and censoring rates (C.R.) $=10\%$, 40%, resulting in a total of 24 simulation scenarios for $p \gg n$. For each scenario, the reported experimental results are based on 1000 simulated data set. To get an idea of how well the fitted model describes the data, we consider the variance–covariance matrix of the regression coefficients $\boldsymbol{\beta}_n$ given by

$$\boldsymbol{\Sigma}(\hat{\beta}_n) = \hat{\sigma}_\varepsilon^2\mathbf{M} = \hat{\sigma}_\varepsilon^2[(\widetilde{\mathbf{X}}'_n\tilde{\mathbf{X}}_n)]^{-1} = \begin{pmatrix} \varSigma_{11} & \varSigma_{12} \\ \varSigma_{21} & \varSigma_{22} \end{pmatrix},$$

where $\varSigma_{11}$ is a $q \times q$ submatrix of the variance–covariance matrix $\boldsymbol{\Sigma}$ and $\hat{\sigma}_\varepsilon^2$ is the estimated variance of the errors with $\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{n} (\hat{y}_{i\hat{G}} - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{1n} - \hat{f}(t_i))^2/n - \boldsymbol{\beta}_{1n1}$. Note also that we consider the mean square error (*MSE*) to evaluate the goodness of fit for nonparametric estimations and fitted values from the model. For each simulated data set, the MSE values, which measure how close to predicted values are to real observations, are computed respectively by

$$\text{MSE}_{\text{f}} = \frac{1}{1000}\sum_{j=1}^{1000}\sum_{i=1}^{n}(\hat{f}(t_{ij}) - f(t_i))^2$$

and

$$\text{MSE}_y = \frac{1}{1000}\sum_{j=1}^{1000}\sum_{i=1}^{n}(\hat{y}_{ij\hat{G}} - y_{i\hat{G}})^2,$$

where $\hat{f}(t_{ij})$ shows the estimated value at the ith point of the function f in jth iterations and $\hat{y}_{ij\hat{G}}$ denotes the estimated fitted value at the ith point of the synthetic response variable $y_{\hat{G}}$ in jth replications.

5.1. Evaluating the empirical results

Outcomes obtained from the simulation experiments are summarized in the following tables and figures. It should be noted that, in Tables 1 and 2, results of ($p = 5$) are given for comparing the introduced estimator with classical semiparametric estimation procedure which can be thought of as a benchmark case. In this sense, Table 1 gives the results obtained from the parametric component and fitted values of the model (23). In Table 2, T $\varSigma_{11}$ denotes the mean of the trace ($\varSigma_{11}$), and q indicates the number of nonzero regression coefficients. As can be seen from the data in Table 1, as the number of parameter in the model increases, the quality of the estimates decreases. Similarly, when the censoring rates increase, we get poor estimates. As expected, for larger sample sizes, we obtained good

Table 2. The MSE values for the nonparametric component of the semiparametric model with CR $=$ 10%, 40% and 12 various (n, p) scenarios, respectively.

		CR = 10%				CR = 40%		
Sample size (n)	$p = 5$	$p = 300$	$p = 1000$	$p = 3000$	$p = 5$	$p = 300$	$p = 1000$	$p = 3000$
50	0.2198	0.1980	0.1928	0.1816	0.4144	0.3841	0.3917	0.3968
100	0.1837	0.1295	0.1495	0.1445	0.3759	0.3140	0.3483	0.3602
200	0.1079	0.0940	0.1141	0.1175	0.3427	0.2706	0.2717	0.2743

results, which can be interpreted as a proof of asymptotical consistency. Asymptotic properties of DPLS are inspected by Ni et al. [11], in detail. Here, because of the smoothing spline method is used for estimating the model, findings for the high censoring level (40%) very different than from the low censoring level (10%). This case can be explained with a sensitivity of the smoothing splines to censoring. (See Aydın and Yılmaz [7], for a more detailed discussion.)

We also analyse the number of selected important explanatory variables are here. Stodden [35], states that in small sparsity levels – which means much selected explanatory variables – an increment in the error of the estimation can be seen; in addition, the model cannot be estimated correctly for less sparse cases. In this context, when Table 1 is inspected carefully, it should be emphasized that the models that contain more predictors have higher variances. The number of selected q-explanatory variables tends to change depending on the magnitude of both the number of parameters p and sample size n.

To better understand the performance of the estimation procedure from the parametric component, we use real observations of the response variable and their fitted values obtained from model (23) with different p covariates. To illustrate this point, Figure 1 offers four plot diagrams. To save space, only four combinations are presented in this figure, because there are many different situations and it would be both difficult and inefficient to present all of them. In each panel, three levels for the number of parameters are illustrated with three separate locations on the y-axis. The aim of Figure 1 is to see teh effects

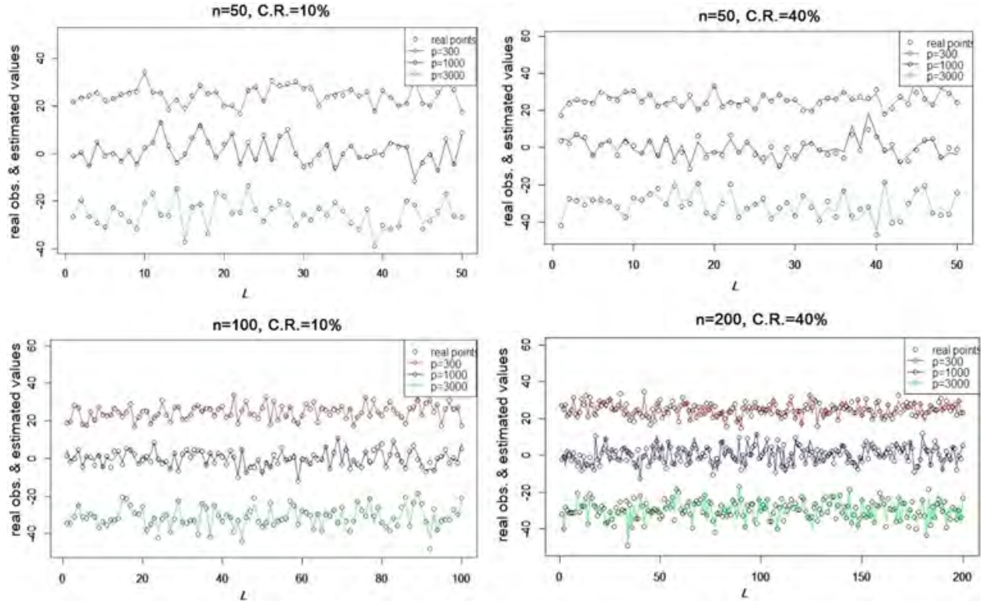

Figure 1. Real observations and fitted values, which were obtained from the parametric component of model (24), based on different simulation scenarios. The red line denotes the fitted values (i.e., $\tilde{X}'^{1n}\hat{\boldsymbol{\beta}}_{1n}$) for $p = 300$, where $\tilde{X}_{1n}$ represents the vector of the selected explanatory variables associated with the vector $\hat{\boldsymbol{\beta}}_{1n}$ of nonzero coefficients. Similarly, the blue line denotes the fits for $p = 1000$, and the green lines represents the fits for $p = 3000$. L denotes the ordered values range from 1 to sample size n.

of the censoring levels, sample sizes and the number of parameters on the estimation performance.

The upper two panels in Figure 1 show the real observations and their fitted values for $n = 50$, two different censoring rates and three different dimensions (p). The bottom-left panel of Figure 1 displays the fits obtained from the parametric component of the model (23) for sample of size $n = 100$, C.R. = 10% and three different dimensions, while the bottom-right panel of the same figure indicates the fits, but for $n = 200$ and C.R. = 40%. As expected, censoring level affected the performance of the estimator in a negative way for all sample sizes. It should also be noted that as the number of covariates p get large, the quality of the estimates declines. This case can be seen explicitly in the bottom-right panel of Figure 2.

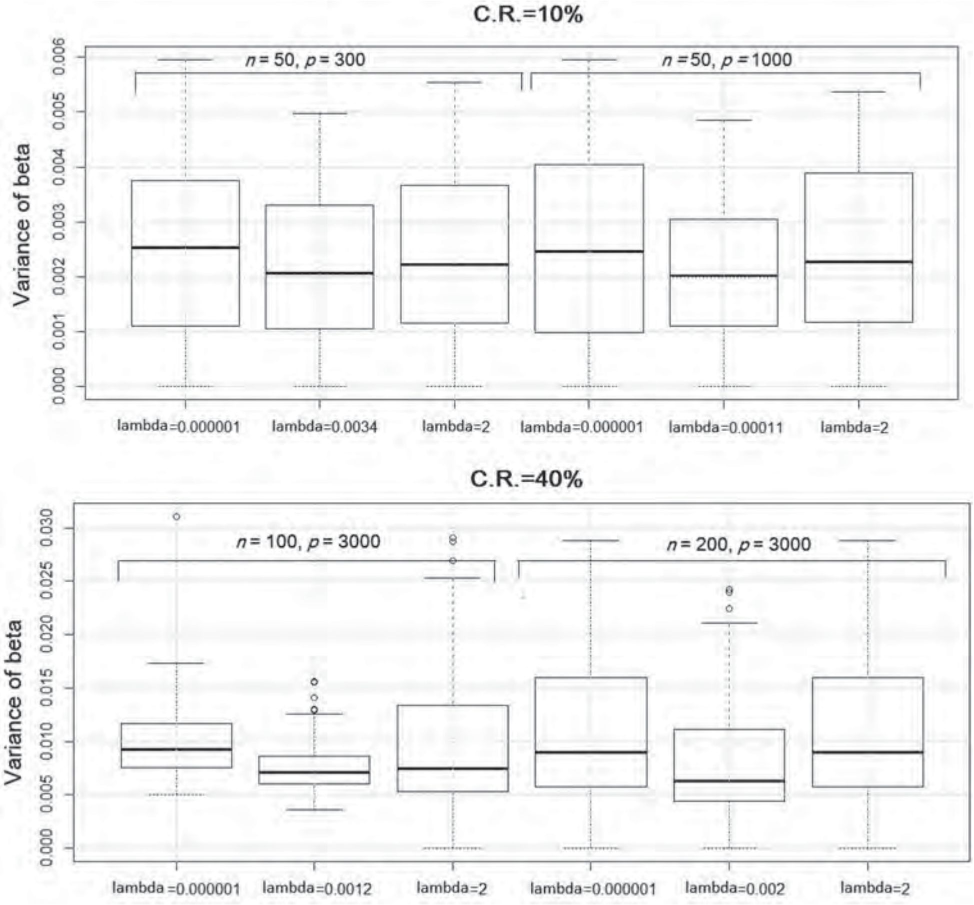

Figure 2. Boxplots of the variances of the estimated nonzero regression coefficients for different values of the shrinkage parameter λ. In each panel 'lambda = 0.000001and lambda = 2' denote the small and high values of shrinkage, respectively. All other values of lambda represent the shrinkage parameters selected by GCV. The upper panel shows the boxplots of the variances from the data with C.R. = 10% and $(n, p) = (50,300)$ and (50, 1000), respectively. The bottom panel presents the boxplots of the variances from the observations with C.R. = 40% and $(n, p) = (100, 3000)$ and (200, 3000), respectively.

Note that one of the most important issues in lasso-type estimation procedures is the over-fitting problem, resulting in noisy estimates. A careful inspection of the outcomes from the parametric component illustrated in Table 1 and Figures 1–2 indicates that the DPLS method produces estimates with satisfactory accuracy. The boxplots in Figure 2 show the averaged variance estimates of nonzero regression coefficients for different shrinkage parameters under various simulated data sets with censoring rates 10% and 40%. To save space, only four simulation combinations are illustrated in Figure 2. It is clear that the GCV method selects the optimum shrinkage parameter λ. It should be emphasized that the variances of nonzero regression coefficients based on parameter λ selected by GCV are optimal compared to the other shrinkage parameters (see Figure 2). This means that GCV provides a balance between the magnitude of error and degree of freedom.

The impact of the censoring rate and the number of parameters can be detected more easily in the results of the nonparametric component of the model. In order to depict this impact, Table 2 includes the MSE values from the nonparametric component of the model (23). Firstly, it should be noted that the results are comparatively good, considering the very problematic data from which they arise. Apart from these, the outcomes from the nonparametric part of the model are similar to the parametric component in terms of the magnitude of the censoring levels and the number of variables p. There is a remarkable point that needs to be explained in this study; normally, the smoothing spline method is a sensitive method for estimating censored data by using synthetic data, since all data points are used as node points. In this study, however, smoothing spline method appears to be less affected by censorship because it is used in conjunction with DPLS. As shown in Table 2,

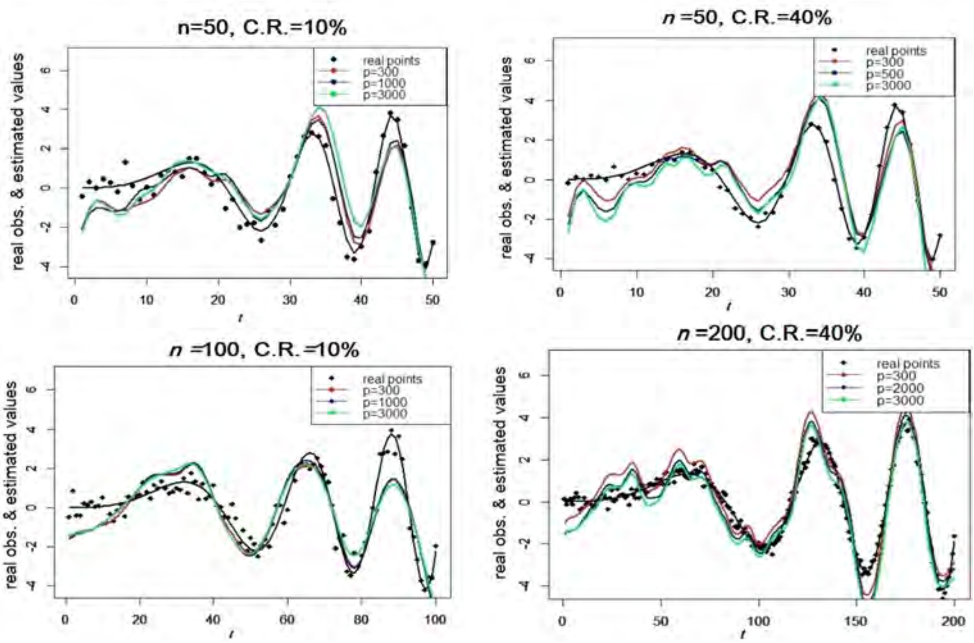

Figure 3. Real observations and their estimated curves for $f(.)$ for different sample sizes, censoring levels and number of parameters.

the MSE value is 0.1980 for the low censoring rate (10%) and $p = 300$, whereas the MSE is 0.3841 for the high censoring level (40%).

Figure 3 is designed for the nonparametric component; it is similar to Figure 1 and proves the outcomes given in Table 2. Here, the effect of the censoring rate can easily be seen in the top-right panel of this figure. Moreover, in each of the panels, estimated curves of the $p = 3000$ seem worse than the others. By looking at Figure 3, one can easily notice the improvement of the estimation when the sample size is getting larger.

It is worthwhile to note that some of the disruptions that can be seen in the estimated curves are heavily censored. One of the most important causes of this is synthetic data transformation, because synthetic data transformation increases the magnitude of the uncensored observations and replaces censored points with zero to provide the $E[y_{i\hat{G}}|\mathbf{x}_i, t_i] = E[y_i|\mathbf{x}_i, t_i]$.

6. Real data example

In this section, we used Norway/Stanford Breast Cancer (NSBC) data set to estimate the censored semiparametric regression model with high-dimensional. This data set is provided by Sorlie et al. [36], who studied the analysis of the patterns of the gene expressions to distinguish the subtypes of the breast tumours. This data set is also used by Li et al. [37], to obtain a parametric regression model for high-dimensional survival data.

The mentioned NSBC data set includes gene expression measurements of 115 malignant tumours obtained from women. Of the 115 patients, 33% (38) experienced an event during the study. In other words, censoring rate is 33%. It is also noted that the nonparametric part of the semiparametric model is composed of a univariate variable t, while the parametric part is constructed using 548 explanatory variables to estimate the *survival times* of the patients. For this example, a right-censored semiparametric model with high-dimensional data is specified by

$$y(\text{survival time})_{i\hat{G}} = \mathbf{x}_i\boldsymbol{\beta}_n + f(t_i) + \varepsilon_{i\hat{G}}, i = 1, \dots, 115 \tag{24}$$

where $\mathbf{x}_i = \{(x_{i1}, \dots, x_{ip})', i = 1, 2, \dots, n \quad \text{where } n = 115 \text{ and } p = 548\}$ denotes the vector-valued variables, $\boldsymbol{\beta}_n$ is p x 1 vector of regression coefficients, t_i is one point of the gene expression measurement data, and $f(.)$ is a nonlinear function of data points t_i. The results, which are graphically displayed in Figure 4, demonstrate that there is a nonlinear relationship between nonparametric and response variables.

Note also that the smoothing and penalty tuning (or shrinkage) parameters selected by GCV are λ_1 =0.00005 and $\lambda_2 = 0.00012$, respectively. Using these parameters, some of the outcomes obtained from the censored semiparametric regression analysed are summarized in Table 3 for the NSBC data set. As you can see, these results reveal that the semiparametric model (24) with a nonparametric component is reasonable for this data set.

When dealing with the high-dimensional problem, a key issue is to have a good insight into the variance of the estimator. The estimated averaged-variance of the regression coefficients is 0.14259 for this data set, as shown in Table 3. This value reveals that DPLS leads to a consistent variance estimation of parametric coefficients in the censored semiparametric model. In Figure 5, we present the nonparametric component of the model (12), through

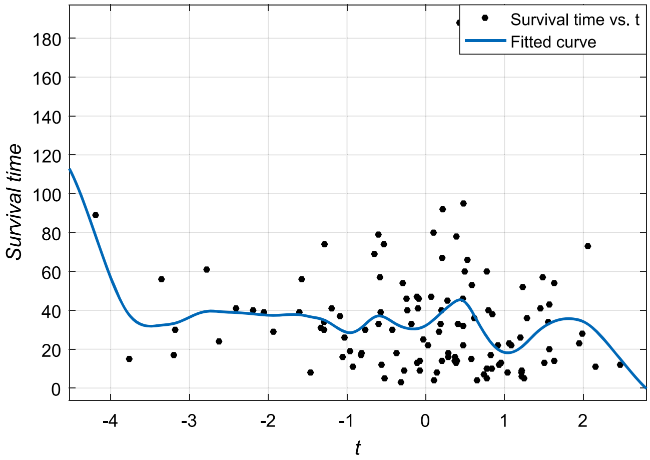

Figure 4. Nonlinear relationship between t_i and response variable y_i.

Table 3. The results from the estimated regression model

	MSE_y	MSE_f	T Σ_{11}	q
NSBCD set	3.00214	8.17324	0.14259	56

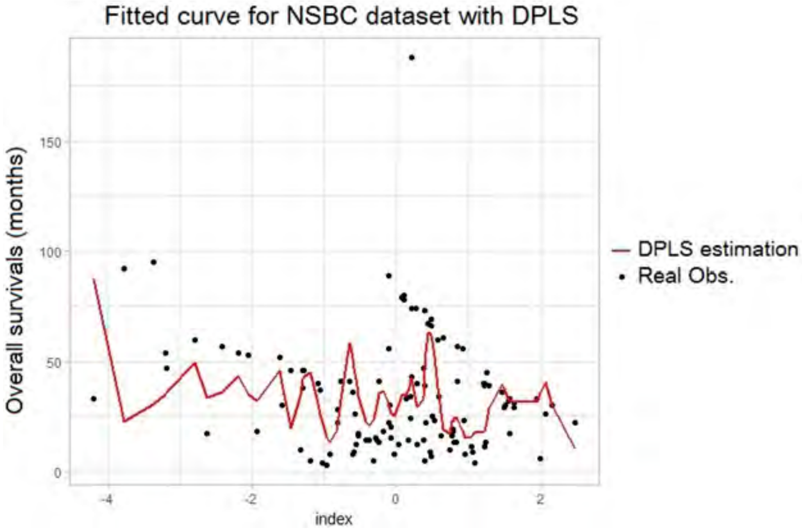

Figure 5. Real response observations and fitted curve, which are considered nonparametric components of the right-censored high-dimensional semiparametric model using DPLS.

which one can clearly see that the DPLS method also works well for the nonparametric part of the model in spite of the aforementioned censoring and high-dimensional problems.

7. Concluding remarks

In this paper, to estimate the semiparametric regression model with high-dimensional and right-censored data, we used the double-penalized least squares (DPLS) method, as indicated before. To better understand the method, simulation experiments and a real data example are carried out. We present the results obtained from the simulation study and the real data example in Figures 1–5 and Tables 1–3; the results that the DPLS method is both useful and feasible in the estimation procedure of the semiparametric regression model under censored high-dimensional data.

The empirical results of our study confirmed that the DPLS method generally performed well under high-dimensional censored data. Although the censoring level in the simulation is increased by up to 40%, the method has not lost its stability and accuracy. However, as the level of censorship increases, the quality of estimates decreases, as expected. In summary, based on the numerical simulation experiments and real data results, the following suggestions and conclusions should be considered:

- The DPLS method gives reasonable results for all censoring levels, sample sizes and the number of parameters. More specifically, one can see in Tables 1 and 2, that the performance of the method is affected by the number of parameters and the censoring rate. Under the condition of $p \gg n$, in general, as the number of model parameters increases, the performance of the model is decreased.
- Interestingly, the DPLS method is resistant to the censoring rate. When this ratio is set to 40%, we expected that the results would be much worse. However, when the results are compared with the classical ($p = 5$) results in Tables 1 and 2, it is clear that the DPLS estimator works reasonably well under the level of heavy censorship. This case proves that the SAFE rule stated in step 4 of the computational algorithm recovers the correct model and has an oracle property.
- In the real data example, we used the NSBC data set and obtained satisfactory results; these are presented in Table 3 and Figure 5. Outcomes of real data are in harmony with simulation study when $n = 100$ and $p = 1000$.
- For both studies, the estimated curves of the nonparametric component are shown in Figures 3 and 5. These outcomes denote that when the censorship ratio and the number of parameters increase, the curves begin to deteriorate, as in the results obtained from the parametric component of the model.

In conclusion, the overall results of two numerical studies demonstrated that the introduced DPLS method provides reasonable estimation procedure for semiparametric regression model with right-censored and high-dimensional data.

Acknowledgments

We would like to thank the editor, the associate editor, and the anonymous referee for beneficial comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Engle RF, Granger CWJ, Rice J, et al. Semiparametric estimates of the relation between weather and electricity sales. J Am Stat Assoc. 1986;81(394):310–320.

[2] Wahba G. Spline model for observational data. Philadelphia (PA): SIAM; 1990.

[3] Green PJ, Silverman BW. Nonparametric regression and generalized linear model. London: Chapman & Hall; 1994.

[4] Speckman P. Kernel smoothing in partial linear models. J Roy Stat Soc B (Method). 1988;50(3):413–436.

[5] Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. New York: Cambridge University Press; 2003.

[6] Orbe J, Ferreira E, Núñez-Antón V. Censored partial regression. Biostatistics. 2003;4(1): 109–121.

[7] Aydin D, Yilmaz E. Modified estimators in semiparametric regression models with right-censored data. J Stat Comput Simul. 2018;88(8):1470–1498.

[8] Xie H, Huang J. SCAD-penalized regression in high-dimensional partially linear models. Ann Stat. 2009;37(2):673–696.

[9] Gao X, Ahmet SE, Feng Y. Post selection shrinkage estimation for high dimensional data analysis. Appl Stoch Model Bus Ind. 2016;33:97–120.

[10] Cheng Y, Wang Y, Camps O, et al. The interplay between big data and sparsity in systems identification: some lessons from machine learning. IFAC-PapersOnLine. 2015;48(28):1285–1292.

[11] Ni X, Zhang HH, Zhang D. Automatic model selection for partially linear models. J Multivariate Anal. 2009;100(9):2100–2111.

[12] Ma S, Du P. Variable selection in partly linear regression model with diverging dimensions for right censored data. Stat Sin. 2012;22:1003–1020.

[13] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B. 1996;58:267–288.

[14] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1360.

[15] Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.

[16] Efron B, Hastie T, Johnstone I, et al. Least angle regression. Ann Stat. 2004;32(2):407–499.

[17] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–1429.

[18] Stute W. Nonlinear censored regression. Stat Sin. 1999;9:1089–1102.

[19] Stute W. The central limit theorem under random censorship. Ann Stat. 1995;23:422–439.

[20] Heuchenne C, Van Keilegom I. Nonlinear regression with censored data. Technometrics. 2007;49(1):34–44.

[21] Zhou M. Asymptotic normality of the synthetic data regression estimator for censored survival data. Ann Stat. 1992;20(2):1002–1021.

[22] Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. Annals Stat. 1981;9: 1276–1288.

[23] Kaplan E. L. M. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–481.

[24] Müller P, Van de Geer S. The partial linear model in high dimensions. Scand J Stat. 2015;42(2):580–608.

[25] Mammen E, Van de Geer S. Locally adaptive regression splines. Ann Stat. 1997;25(1): 387–413.

[26] Zou H, Hastie T. Regularization and variable selection via the Elastic Net. J Roy Stat Soc B. 2005;67:301–320.

[27] Tibshirani R, Saunders M, Rosset S, et al. Sparsity and smoothness via the fussed lasso. J Roy Stat Soc B. 2005;67(1):91–108.

[28] Guo J, Hu J, Jing B-Y, et al. Spline-Lasso in high-dimensional linear regression. J Am Stat Assoc. 2016;111(513):288–297.

[29] El Ghaoui L., Viallon V., Rabbani T. Safe feature elimination for the LASSO and sparse supervised learning problems. Pac J Optim. 2010;8(4):667–698.

[30] Tibshirani R, Bien J, Friedman J, et al. Strong rules for discarding predictors in lasso-type problems. J Roy Stat Soc B Stat Methodol. 2012;74(2):245–266.

[31] Van de Geer S, Bühlmann P, Ritov Y, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Stat. 2014;42(3):1166–1202.

[32] Van der Vaart A. Asymptotic statistics. Cambridge: Cambridge University Press; 2000.

[33] Jankova J, Van de Geer S. Semi-parametric efficiency bounds for high-dimensional models. Ann Stat. 2016;46(5):2336–2359.

[34] Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Ann Stat. 2004;32(3):928–961.

[35] Stodden V. Model selection when the number of variables exceeds the number of observations [Ph.D Thesis]. Department of Statistics, Stanford University; 2006.

[36] Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Nat Acad Sci. 2003;100(14):8418–8423.

[37] Li Y, Kevin SX, Chandan KR. (2016). Regularized parametric regression for high-dimensional survival analysis, Proceedings of the 2016 SIAM International Conference on Data Mining, doi:10.1137/1.9781611974348.86