Volume 47 • Number 9 • 2018
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Modified spline regression based on randomly right-censored data: A comparative study

Dursun Aydin & Ersin Yilmaz

To cite this article: Dursun Aydin & Ersin Yilmaz (2018) Modified spline regression based on randomly right-censored data: A comparative study, Communications in Statistics - Simulation and Computation, 47:9, 2587-2611, DOI: 10.1080/03610918.2017.1353615

To link to this article: https://doi.org/10.1080/03610918.2017.1353615

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2018, VOL. 47, NO. 9, 2587–2611
https://doi.org/10.1080/03610918.2017.1353615

Modified spline regression based on randomly right-censored data: A comparative study

Dursun Aydin and Ersin Yilmaz

Department of Statistics, Faculty of Science, Mugla Sitki Kocman University, Muğla, Turkey

ABSTRACT

In this paper, we propose modified spline estimators for nonparametric regression models with right-censored data, especially when the censored response observations are converted to synthetic data. Efficient implementation of these estimators depends on the set of knot points and an appropriate smoothing parameter. We use three algorithms, the default selection method (DSM), myopic algorithm (MA), and full search algorithm (FSA), to select the optimum set of knots in a penalized spline method based on a smoothing parameter, which is chosen based on different criteria, including the improved version of the Akaike information criterion (*AICc*), generalized cross validation (*GCV*), restricted maximum likelihood (REML), and Bayesian information criterion (*BIC*). We also consider the smoothing spline (SS), which uses all the data points as knots. The main goal of this study is to compare the performance of the algorithm and criteria combinations in the suggested penalized spline fits under censored data. A Monte Carlo simulation study is performed and a real data example is presented to illustrate the ideas in the paper. The results confirm that the FSA slightly outperforms the other methods, especially for high censoring levels.

ARTICLE HISTORY
Received 8 February 2017
Accepted 5 July 2017

KEYWORDS
Censored data; Kaplan–Meier estimator; Nonparametric regression; Penalized splines; Synthetic data

1. Introduction

In regression analysis, when the effect of a covariate on the response is unspecified parametrically, nonparametric regression methods are commonly used to explain the relationship between the response and covariate. Formally, the above situation can be described by the following nonparametric regression model. Let $\{(X_i, Y_i),\ 1 \leq i \leq n\}$ be a random sample satisfying

$$Y_i = f(X_i) + \varepsilon_i, \quad a = X_1 < \cdots < X_n = b, \tag{1}$$

where the Y_i's are observations of the response variable, the X_i's are the values of the covariate, f is some unspecified smooth regression function and the ε_i's are independent random error terms with mean zero and variance σ_ε^2.

In the case of uncensored response observations, a number of statistical methods have been developed to estimate model (1), for example, the studies by Eubank (1988), Hardle

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

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

(1990), Wahba (1990), Ruppert (2002), Ruppert, Wand and Carroll (2003), and Eilers and Marx (2010). In practice, the Y_i's may be incompletely observed and right censored by a random censoring variable C_i. Therefore, instead of observing $(Y_i,\ X_i)$, one observes the dataset $\{(X_i, Z_i, \delta_i),\ i = 1, 2, \ldots, n\}$ with

$$Z_i = \min\left(Y_i, C_i\right),\ \ \delta_i = \mathrm{I}\left(Y_i \le C_i\right) = \left\{1\ \mathit{if}\ \ (Y_i \le C_i)\ \ \mathit{and}\ 0\ \mathit{otherwise}\right\}, \tag{2}$$

where I(.) is an indicator function, and Z_i and C_i are the failure times (or observed lifetimes) and the censoring time, respectively, for the ith subject. In the presence of censoring, model (1) reduces to the censored nonparametric regression model.

Ordinary statistical methods cannot be applied directly to censored observations, and data transformation is required to explain the relationship between the response and covariate. In this context, several authors have proposed different data transformations when the regression function is linear (see Buckley & James, 1979; Koul et al., 1981; Leurgans, 1987). Furthermore, a data transformation exists when the form of the regression function is unspecified, for example, the local average transformation of Fan and Gijbels (1994) and the data transformation technique studied by El Ghouch and Van Keilegom (2008). Additionally, the consistency of the weighted estimate of the unknown regression function for nonparametric regression with right-censored data was studied by Kalbfleich and Prentice (1980) and Wang (1996) discussed the convergence properties of the weighted kernel estimate for nonparametric regression function; Yang (1999) examined the weighted kernel estimators of a nonparametric regression function with censored data. In addition to these authors, there are many studies in the literature on the estimation of nonparametric regression models with randomly right-censored data. Examples of these works include Zheng (1984), Cai and Betensky (2003), Dabrowska (1992), and Kim and Truong (1998).

In this paper, we focus on model (1) when the response observations are subject to random right censoring. For simplicity, we consider the transformed versions of the censored observations, called synthetic data, proposed by Koul et al. (1981). We apply a modified regression spline estimator to the synthetic data to study the knot-selection algorithms in combination with an appropriate smoothing parameter. The above estimator is a generalization of the well-known penalized spline estimator for model (1). For more details on penalized splines, see Eilers and Marx (1996), Ruppert et al. (2003), and Hall and Opsomer (2005). The main difference in our study is that we consider a randomly right-censored nonparametric regression model that is estimated by using several knot selection algorithms under simulation and real data settings. The basic idea is to find a useful selection algorithm that provides a good approximation $\hat{f}(X)$ to the function $f(X)$ and then to compare the performance of these algorithms by using different selection criteria. To the best of our knowledge, such a study has not yet been conducted.

The rest of this paper is organized as follows. In Section 2, the preliminaries required to understand the estimation method are expressed, and the regression spline method, synthetic data transformation and derivation of the proposed estimator are illustrated. The variance of the estimator and the relative efficiencies are obtained in Section 3. In Section 4, selection methods for the smoothing parameter are expressed, and in Section 5, the knot selection algorithms are illustrated. Then, a simulation experiment is conducted in Section 6, and the estimation method is applied to a dataset of patients with colon cancer. Finally, in the last section, conclusions and comments about the simulation and real data applications are presented.

2. The preliminaries and methodology

We assume that Y_i, C_i, and Z_i have distribution functions F, G, and K, respectively, and that (X, Y) and C are independent. These variables are assumed to be non-negative random with distribution functions

$$F(t|X=x)=P(Y_i\le t|X=x,\ (t\in R)),\ G(t|x)=P(C_i\le t|x),\quad \text{and } K(t|x)=P(Z_i\le t|x)$$

and corresponding survival functions

$\bar{F}(t|x)=1-F(t|x)=P(Y_i>t|x)$, $\bar{G}(t|x)=1-G(t|x)=P(C_i>t|x)$, and (because of the independence of Y and C)

$$\bar{K}(t|x)=1-K(t|x)=(1-(F(t|x)\times G(t|x)))=P(Z_i>t|x).$$

To ensure that the model is identifiable, we assume that

$$\begin{aligned} T_F &= \sup\left[\mathrm{t}:\bar{\mathrm{F}}(\mathrm{t}|\mathrm{x})>0\right];\quad T_G=\sup\left[\mathrm{t}:\bar{\mathrm{G}}(\mathrm{t}|\mathrm{x})>0\right];\\ T_K &= \sup\left[\mathrm{t}:\bar{\mathrm{K}}(\mathrm{t}|\mathrm{x})>0\right]=\min(\mathrm{T}_F,\mathrm{T}_G) \end{aligned} \tag{3}$$

Throughout this paper, we also assume that $\mathrm{T}_F<\infty$, G is continuous, F and G have no common jumps, and $G(\mathrm{T}_F)>0$. The assumption $G(\mathrm{T}_F)>0$ implies that $\mathrm{T}_F<T_G$; hence, it is easily seen that $\mathrm{T}_F=T_K$ by definition of T_K. Note that under assumption (3), an ordinary estimate of $f(.)$ can be defined by

$$f(x)=\int_0^{\infty}F(t|x)dt=\int_0^{T_F}F(t|x)dt=E(Y|X=x). \tag{4}$$

Because of the censoring, the traditional methods for estimating $f(x)$ are inapplicable. One reason for this restriction is that the censored observation Z_i and the true random variable Y_i have different expectations. This difficulty can be overcome by using the synthetic data method, as in censored linear models. We refer, for example, to the studies of Koul et al. (1981), Lai and Ying (1992), and Zhou (1992) for more details. The synthetic data method enables us, through some transformation, to modify the censored and uncensored observations, hence ensuring that a transformed observation has the same expectation as the random variable Y_i in principle (see Lemma 1). In this context, we perform data transformation

$$Z_{iG}=\frac{\delta_i Z_i}{1-G(Z_i)}=\frac{\delta_i Z_i}{\bar{G}(Z_i)}, \tag{5}$$

where $G(.)$ is the common distribution of the censoring variable C_i, as mentioned in the introduction to this section. Thus, model (1) transforms to the following censored nonparametric regression model

$$Z_{iG}=f(X_i)+\varepsilon_{iG},\quad \varepsilon_{iG}=Z_{iG}-f(X_i),\quad 1\le i\le n, \tag{6}$$

where the $\varepsilon_{iG}{}'s$ are random variables for a given G, and $E(\varepsilon_{iG})=0$ (see Appendix A4). Therefore, there is a distinct probability distribution for Z_{iG} at each point X_i; that is, (Z_{iG}, X_i), $i=1,\ldots,n$ is a sequence of random variables with the mean of distribution $E(Z_{iG}|X)=f(X_i)$. Additionally, the following Lemma 1 shows that Z_{iG} and Y_i have the same expected values, and the unknown regression function $f(X)$ becomes a problem of estimating the expectation from censored data.

Lemma 1. *If instead of Y_i only $\{(Z_i,\ \delta_i),\ 1\le i\le n\}$ are observed and the censoring distribution G is known, then the regression function $f(X)$ is a conditional expectation; that is, $E(Z_{iG}|X)=E(Y_i|X)=f(X_i)$.*

Proof. Lemma 1 can be easily verified by using the assumed independence of Y and C and the properties of conditional expectation:

$$\begin{aligned} E(Z_{iG}\,|X) &= E\left[\frac{\delta_i Z_i}{1-G(Z_i)}|X\right] = E\left[\frac{\delta_i Z_i}{\bar{G}(Z_i)}|X\right] \\ &= E\left[\frac{I(Y_i \le C_i)\min(Y_i, C_i)}{\bar{G}\,[\min(Y_i, C_i)]}|X\right] = E\left[I(Y_i \le C_i)\frac{Y_i}{\bar{G}(Z_i)}|X\right] \\ &= E\left[E\left[\frac{Y_i}{\bar{G}(Z_i)}I(Y_i \le C_i)|X, Y\right]|X\right] = E\left[\frac{Y_i}{\bar{G}(Z_i)}\bar{G}(Z_i)|X\right] = E(Y_i|X) = f(X_i) \end{aligned}$$

In survival applications, the censoring distribution G is usually unknown. Therefore, Lemma 1 for estimating $f(\cdot)$ cannot always be applied. To overcome this problem Koul et al. (1981) proposed replacing G with its Kaplan–Meier estimator

$$\hat{\bar{G}}(t) \equiv 1 - \hat{G}(t) = \prod_{i=1}^{n}\left(\frac{n-i}{n-i+1}\right)^{I[Z_{(i)}\le\, t,\ \delta_{(i)}=0]}, \quad (t \ge 0), \tag{7}$$

where $(Z_{(i)}, \delta_{(i)}),\ i = 1, 2, \ldots, n$, are the pairs of observations $(Z_{(i)}, \delta_{(i)})$ ordered on the $Z_{(i)}$, i.e., $Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(n)}$. Note that if G is chosen arbitrarily, some $Z_{(i)}$ may be identical. In this case, the ordering of $Z_1, \ldots, Z_n$ into $Z_{(1)}, \ldots, Z_{(n)}$ is not unique. However, the Kaplan-Meier estimator enables us to identify the ordering of Z uniquely. Additionally, $\hat{G}(t)$ has jumps only at the censored data points (see Peterson, 1977).

Based on (6), several estimates of $f(X)$ can be performed, such as the smoothing spline, kernel smoothing and regression spline. For convenience, we use the regression spline method to estimate the unknown regression function in model (6). This estimation procedure is explained in the following section.

2.1. Derivation of the proposed estimator

We now consider the ideas described above to apply the regression spline (or penalized spline) method to the case of randomly right-censored data. To approximate the function $f(X)$ in (6), we use a pth-degree penalized spline with truncated polynomial basis

$$f(X; \boldsymbol{\beta}) = \beta_0 + \beta_1 X + \cdots + \beta_p X^p + \sum_{k=1}^{K}\beta_{p+k}(X-\kappa_k)_+^p,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p,\ \beta_{p+1}, \ldots, \beta_{p+K})'$ is a vector of unknown coefficients to be estimated, $p \ge 1$ is an integer, $(X-\kappa_k)_+^p = (X-\kappa_k)^p$ *if* $X > \kappa_k$ or zero otherwise and $\{\kappa_1, \ldots, \kappa_K\}$ is a set of fixed knots $\{\min(X_i) \le \kappa_1 <, \ldots, < \kappa_K \le \max(X_i)\}$(see Ruppert, 2002, for details about knot selection).

It follows from the above truncated polynomial that the censored regression model in (6) can be rewritten as

$$Z_{iG} = \left(f(X_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \sum_{k=1}^{K}\beta_{p+k}(X_i-\kappa_k)_+^p\right) + \varepsilon_{iG}, \quad 1 \le i \le n, \tag{8}$$

where the ε_{iG}'s error terms with mean zero and constant variance σ^2. In matrix and vector notation, model (8) is defined by

$$\mathbf{Z}_G = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}_G, \tag{9}$$

where $\mathbf{X}$ is the design matrix for the regression spline such that the ith row of matrix $\mathbf{X}$ is

$$\mathbf{X}_i = \left[\, 1 \; X_i \; \ldots \; X_i^p \; (X_i - \kappa_1)_+^p \; \ldots \; (X_i - \kappa_K)_+^p \,\right], \quad 1 \le i \le n,$$

and $\mathbf{Z}_G$ is a vector containing the values of the synthetic variable Z_{iG}. Then, the regression spline (or penalized spline) estimates of the coefficients vector $\boldsymbol{\beta}$ are obtained by minimizing the penalized residual sum of squares *(PRSS)* criterion

$$PRSS(\lambda;\ \boldsymbol{\beta}) = \sum_{i=1}^{n} (Z_{iG} - f(X_i))^2 + \lambda \sum_{k=1}^{K} \beta_{p+k}^2 = |\mathbf{Z}_G - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta}, \tag{10}$$

where $\mathbf{D} = \text{diag}(\mathbf{0}_{p+1},\ \mathbf{1}_K)$, that is, $\mathbf{D}$ is a diagonal matrix whose first $(p\ +\ 1)$ elements are 0 and whose remaining elements are 1.

We note that $\lambda\boldsymbol{\beta}'D\boldsymbol{\beta}$ in (10) is called a penalty term because it penalizes the curvature in function f, thus yielding a smoother result. The amount of penalty is controlled by smoothing parameter $\lambda > 0$. In general, large values of λ produce smoother estimators, while small values produce less smooth estimators. λ plays a key role in estimating model (8). An additional task is to select the optimal value of λ. This problem is discussed in Section 4.

In practice, however, since the censoring distribution G is unknown, criterion (10) cannot be used directly. Therefore, we replace G with Kaplan and Meier (1958) estimator $\hat{G}$ in (7). By introducing the $\hat{G}$, the model (8) transforms to $Z_{i\hat{G}} = f(X_i) + \varepsilon_{i\hat{G}}$. From the definition of $Z_{i\hat{G}}$ in (5) it can be seen that $\varepsilon_{i\hat{G}}$ (i $=1, \ldots,$ n) form a sequence of identically distributed (but not independent) random variables. Heuristically, $E(\varepsilon_{i\hat{G}}) \cong 0$ as n $\to \infty$. So we can treat the synthetic observation values (X$_i$,Z$_{i\hat{G}}$) (i $= 1, \ldots,$ n) if they come from a nonparametric regression model with errors $\varepsilon_{i\hat{G}}$ (i $= 1, \ldots,$ n). This heuristic argument will help us to determine estimates for f(.) (see Qin and Jing (2000) for details).Thus with the estimates of G, penalized criterion (10) can be modified as

$$PRSS(\lambda;\boldsymbol{\beta}) = \sum_{i=1}^{n} (Z_{i\hat{G}} - f(X_i))^2 + \lambda \sum_{k=1}^{K} \beta_{p+k}^2 = |\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda\boldsymbol{\beta}'D\boldsymbol{\beta}, \tag{11}$$

where $Z_{i\hat{G}} = \delta_i Z_i / 1 - \hat{G}(Z_i) = \delta_i Z_i / \hat{\bar{G}}(Z_i)$ is the estimate of the synthetic data (5), and $\mathbf{Z}_{\hat{G}}$ is the matrix form of synthetic variable $Z_{i\hat{G}}$.

As indicated above, we want to find estimates of vector $\boldsymbol{\beta}$ that minimize criterion (11). Theorem 1 gives these estimators and the regression spline fitted values for nonparametric regression model (9).

Theorem 1. *Let $Z_{\hat{G}} = \mathbf{X}\boldsymbol{\beta} + \varepsilon_{\hat{G}}$ where $\mathbf{X}$ is an $n \times h$ (where $h = p + K + 1$) matrix, $\boldsymbol{\beta}$ is a* $\text{h} \times 1$ *vector of unknown regression coefficients and $\varepsilon_{\hat{G}}$ is an $n \times 1$ vector of error terms with constant variance. The weighted penalized least squares estimator for $\boldsymbol{\beta}$ is indicated by $\hat{\boldsymbol{\beta}}$ and is defined as*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}'\mathbf{Z}_{\hat{G}} \tag{12}$$

The fitted values are

$$\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{Z}_{\hat{G}} \tag{13}$$

where $\mathbf{S}_\lambda$=$\mathbf{X}(\mathbf{X}'\mathbf{X}$+$\lambda\mathbf{D})^{-1}\mathbf{X}'$ is a smoother matrix. The proof of Theorem 1 is given in Appendix A1.

2.2. Finite sample properties with λ fixed

It follows that (12) is also a ridge-type regression estimator that shrinks the penalized spline towards the penalized least squares fit to a pth degree spline with a truncated polynomial. When the estimator is calculated over a grid of values with fixed smoothing parameter λ, this is very similar to linear ridge regression with a fixed design. Thus, as in ordinary ridge regression, we can approximate the variance matrix of $\hat{\boldsymbol{\beta}}$ using the following Sandwich formula

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2 n^{-1}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}. \tag{14}$$

Similarly, the variance matrix of the fitted values in (13) is given by

$$Cov(\hat{\mathbf{f}}_\lambda) = \sigma^2 n^{-1}(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}')(\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}')'. \tag{15}$$

Given a choice for λ, the only unknown in Eqs (14 and 15) is the variance σ^2, which can be replaced by the nearly unbiased estimator discussed in (20).

Now, we present results on the strong consistency and asymptotic normality for penalized least squares estimators of nonparametric regression models (see, Claeskens et al., 2007). The ideas are explained by the large sample properties of the estimator $\hat{\boldsymbol{\beta}}$ in the next section.

2.3. Large sample properties with $\lambda_n \to 0$

Denote λ by λ_n to imply that the value of λ depends on the sample size. When $n \to \infty$, the variance of $\hat{\boldsymbol{\beta}}$ goes to zero and λ_n tends to zero. However, if $\lambda_n \to 0$ as $n \to \infty$, then the bias of the estimator goes to zero, and asymptotic normality and consistency can be established in the following two theorems. The assumptions of these theorems are given in the Appendix. The proofs are similar to those in Yu and Ruppert (2002) and are not presented here to save space. These theorems are given follows.

Theorem 2. *(Yu and Ruppert, 2002) under assumption A1 given in the Appendix, let* $\{\hat{\boldsymbol{\beta}}_{n,\lambda_n}\}$ *be a sequence of penalized least squares estimators minimizing criterion (10). If the smoothing parameter* $\lambda_n = o(1)$*, then* $\hat{\boldsymbol{\beta}}_n$ *is a strongly consistent estimator of the true parameter* $\boldsymbol{\beta}_0$.

Theorem 3. *(Yu and Ruppert, 2002) Under assumptions A1 and A2 given in the Appendix, let* $\{\hat{\boldsymbol{\beta}}_{n,\lambda_n}\}$ *be a sequence of penalized least squares estimators minimizing criterion (10). If the smoothing parameter* $\lambda_n = o(n^{-1/2})$*, then*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{n,\lambda_n} - \boldsymbol{\beta}_0\right) \xrightarrow{D} N\left(0,\ \sigma^2\Omega^{-1}(\boldsymbol{\beta}_0)\right), \tag{16}$$

where $\Omega(\boldsymbol{\beta}_0) = \lim_{\mathrm{n}}(\mathbf{X}'\mathbf{X}/n)$*, i.e.,* $\Omega(\boldsymbol{\beta}_0)$ *is the almost certain limit of* $n^{-1}(\mathbf{X}'\mathbf{X})$.

In the case of large samples, we can obtain the “Sandwich formula”, which is similar to (14), for the asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$. Furthermore, as $\lambda \to 0$, $Var(\hat{\boldsymbol{\beta}})$ converges to $n^{-1}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

3. Estimating the variance, risk, and efficiency measures

We are now interested in a class of linear estimators for f, $B(\mathrm{H}) = [f_\lambda : \lambda \in \mathrm{H} | \lambda > 0]$, with H denoting any index set. The index parameter λ may be any scalar or vector. The main goal is to select an appropriate estimator of f from the elements $[f_\lambda : \lambda \in \mathrm{H}]$. To obtain such an estimator, we need to select an optimum value of λ. This value is obtained by minimizing the selection methods given in Section 4. According to B(H), an $n \times n$ smoother matrix $\mathbf{S}_\lambda$ can be obtained by each optimum parameter λ. Accordingly, for each selection criterion, Eq. (13) can be rewritten as

$$\hat{\mathbf{f}}_\lambda = (f_\lambda(X_1), \ldots, f_\lambda(X_n))' = \mathbf{S}_\lambda \mathbf{Z}_{\hat{G}} \tag{17}$$

There are some widely used and accepted performance measures that are used to determine the optimum estimator. One of these measures, the so-called P_2 risk, can be obtained as average value of the residual sum of squares $n^{-1}RSS(\lambda)$

$$RSS(\lambda) = \sum_{i=1}^{n} \left((\hat{f}_\lambda)_i - Z_{i\hat{G}} \right)^2 = (\hat{\mathbf{f}}_\lambda - \mathbf{Z}_{\hat{G}})'(\hat{\mathbf{f}}_\lambda - \mathbf{Z}_{\hat{G}}) = \mathbf{Z}'_{\hat{G}}(\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{Z}_{\hat{G}}, \tag{18}$$

where $\hat{\mathbf{f}}_\lambda$ is defined as in (17). The expected value of the squared residuals given in (18) is also known as the mean square error (MSE) of prediction and is given by

$$MSE\,(\lambda) = E\|\mathbf{Z}_{\hat{G}} - \hat{\mathbf{f}}_\lambda\|^2 = E\left\|(\mathbf{I}{-}\mathbf{S}_\lambda)\,\mathbf{Z}_{\hat{G}}\right\|^2 = \mathbf{f}'_\lambda(\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{f}_\lambda + \sigma^2 \left[n - 2\,(\mathbf{S}_\lambda) + (\mathbf{S}_\lambda{}'S_\lambda) \right]. \tag{19}$$

Details on the derivation of Eq. (19) can be found in Appendix A2.

In practice, if σ^2 is known, $MSE(\lambda)$ can be used directly to assess the performance of the regression function $\mathbf{f}_\lambda$. However, σ^2 is generally unknown. In this case, an estimator for σ^2 can be developed based on the residual sum of squares (18).

As a result, σ^2 can be estimated as

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS(\lambda)}{n - p} = \frac{\mathbf{Z}'_{\hat{G}}(\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{Z}_{\hat{G}}}{tr(\mathbf{I} - \mathbf{S}_\lambda)^2} = \frac{\mathbf{Z}'_{\hat{G}}(\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{Z}_{\hat{G}}}{DF_{RES}}, \tag{20}$$

where

$$DF_{RES} = tr(\mathbf{I} - \mathbf{S}_\lambda)^2 = n - 2tr(\mathbf{S}_\lambda) + tr(\mathbf{S}_\lambda{}'\mathbf{S}_\lambda) \tag{21}$$

is called the residual degrees of freedom (DF_{RES}) for the pre-chosen λ with any selection criteria given in Section 4. Similar to the comment by Rupert et al. (2003), assuming that the bias term $\mathbf{f}_\lambda{}'(\mathbf{I} - \mathbf{S}_\lambda)^2 \mathbf{f}_\lambda$ in (19) is negligible, it follows that $\hat{\sigma}_\varepsilon^2 = \mathrm{E}(RSS(\lambda)/\mathrm{DF}_{\mathrm{RES}}$ is an asymptotically unbiased estimate of σ_ε^2.

Another performance measure is the R_2 risk, which measures the expected loss of a vector $\hat{\mathbf{f}}_\lambda$. The R_2 risk is given in Definition 3.1. Our application of the results of the simulation experiments is to approximate the risk in nonparametric regression models. Such approximations have the advantage of being simple to optimize the practical selection of smoothing parameters. For convenience, we work with the scalar-valued mean dispersion error.

Definition 3.1. The R_2 risk is closely related to the matrix-valued mean dispersion error (MDE) of an estimator $\hat{\mathbf{f}}_\lambda$ of $\mathbf{f}$. The scalar-valued version of the MDE matrix is specified as

$$SMDE(\hat{\mathbf{f}}_\lambda, \mathbf{f}) = E(\hat{\mathbf{f}}_\lambda - \mathbf{f})'(\hat{\mathbf{f}}_\lambda - \mathbf{f}) = tr(MDE(\hat{\mathbf{f}}_\lambda,\ \mathbf{f})) \tag{22}$$

Lemma 3.1. *Consider different estimators $\hat{\mathbf{f}}_\lambda$. The mean dispersion error (MDE) of these estimators is the sum of the covariance matrix and the squared bias vector*

$$SMDE(\hat{\mathbf{f}}_\lambda) = E\sum_{i=1}^{n} (f_i(X) - \hat{f}_{\lambda i}(X))^2 = E\|\mathbf{f} - \hat{\mathbf{f}}_\lambda\|^2 = \|(\mathbf{I} - \mathbf{S}_\lambda\,)\,\mathbf{f}\|^2 + \sigma_\varepsilon^2 tr[\mathbf{S}_\lambda{}'\mathbf{S}_\lambda]. \tag{23}$$

See Appendix A3 for the proof of Lemma 3.1.

As shown in Lemma 3.1, the *SMDE* matrix decomposes into a sum of the squared bias and variance of the estimator; hence, we can compare the quality of two estimators based on the ratio of their *SMDE* in (23). This ratio leads to the following definition concerning the superiority of any two estimators.

Definition 3.2. The relative efficiency of an estimator $\hat{\mathbf{f}}_{E1}(\lambda)$ compared to another estimator $\hat{\mathbf{f}}_{E2}(\lambda)$ is defined by

$$RE = \frac{R(\hat{\mathbf{f}}_{E2}(\lambda), \mathbf{f})}{R(\hat{\mathbf{f}}_{E1}\ (\lambda)\ ,\ \ \mathbf{f})} = \frac{SMDE(\hat{\mathbf{f}}_{E2}\ (\lambda))}{SMDE(\hat{\mathbf{f}}_{E1}(\lambda))}, \tag{24}$$

where $R(.)$ denotes the scalar risk, which is equivalent to Eq. (23). $\hat{\mathbf{f}}_{E2}\ (\lambda)$ is said to be more efficient than $\hat{\mathbf{f}}_{E1}\ (\lambda)$ if $RE\ < 1$.

Equation (24) is used to evaluate the efficiency of the obtained estimators. In this paper, the optimum values of λ are determined by the four election criteria, while the number and positions of the knots are based on three selection algorithms, which are presented Sections 4 and 5.

4. Selecting the smoothing parameter

Penalized spline is a linear estimator because it can be written in the form $\hat{\mathbf{f}}_{\lambda} = \mathbf{S}_{\lambda}\mathbf{Z}_{\hat{G}}$ with the smoother matrix $\mathbf{S}_{\lambda}$ in (13) being symmetric and positive definite, depending on λ but not on $\mathbf{Z}_{\hat{G}}$. Our task in this section is to select the optimum value of λ. The optimum λ is the value minimizes the average *MSE*. The *MSE* average is denoted by $\mathbf{T}(\lambda)$ and is given by

$$E(\mathbf{T}(\lambda)) = E\left(\frac{1}{n}\left\|(\mathbf{I}{-}\mathbf{S}_{\lambda})\mathbf{Z}_{\hat{G}}\right\|^{2}\right) = \frac{1}{n}\left\|(\mathbf{I}{-}\mathbf{S}_{\lambda})\ \mathbf{Z}_{\hat{G}}\right\|^{2} + \frac{\sigma^{2}}{n}tr(\mathbf{S}_{\lambda}{}'\mathbf{S}_{\lambda}). \tag{25}$$

The minimizer of Eq. (25) can be taken as a good value of λ; however, as shown by this equation, this is not practical since it depends on the unknown σ^2. Therefore, we only need to find a good estimate of the minimizer of (25) based on our dataset. In practice, this estimate can be achieved by using the smoothing parameter selection criteria. A reasonable value of λ can be chosen to minimize the selection methods. Examples of the most widely used automatic selection procedures are summarized as follows.

GCV Criterion: The generalized cross validation (*GCV*) score is specified as the minimizer of (29), defined by (see Craven and Wahba 1979)

$$GCV\,(\lambda) = n^{-1}\left\|(\mathbf{I}{-}\mathbf{S}_{\lambda})\ \mathbf{Z}_{\hat{G}}\right\|^{2}/\left[n^{-1}tr(\mathbf{I}-\mathbf{S}_{\lambda})\right]^{2},$$

where $\mathbf{S}_{\lambda}$, as is defined in (13), is a smoother matrix based on λ.

AICc: Note that the classic Akaike information criterion tends to overfit when the sample size is relatively small; therefore, Hurvich et al. (1998) suggested an improved version, called AIC_c, which is defined by

$$AIC_c(\lambda) = 1 + \log\left[\left\|(\mathbf{S}_{\lambda} - \mathbf{I})\mathbf{Z}_{\hat{G}}\right\|^{2}/n\right] + \left[\{2tr(\mathbf{S}_{\lambda}) + 1\}/n - tr(\mathbf{S}_{\lambda}) - 2\right].$$

BIC: Schwarz (1978) improved the Bayesian information criterion (*BIC*) by using Bayes estimators. Thus, the *BIC* is also called as Schwarz information criterion (*SIC*). The criterion is expressed as

$$BIC(\lambda) = 1/n\left\| (\mathbf{I} - \mathbf{S}_\lambda)\, \mathbf{Z}_{\hat{G}} \right\|^2 + (log(n)/n)tr(\mathbf{S}_\lambda)$$

REML Criterion: The restricted maximum likelihood (REML) criterion treats λ as a variance parameter. The REML and *GCV* have a similar form and provide identical values. Moreover, the derivatives of both the REML and *GCV* with respect to λ can be determined naturally in a common form (see Reis and Ogden, 2009).

The REML score can be specified as

$$\text{REML}(\lambda) = |(I - \mathbf{S}_\lambda)\mathbf{Z}_{\hat{G}}|^2/\, n - tr(\mathbf{S}_\lambda).$$

The selection criteria in this paper fall into two main groups. The first group attempts to minimize the model prediction error by optimizing the model selection-based criteria (or selectors) and includes *AICc*, *GCV*, and *BIC*. The second group treats smooth functions as random effects so that the smoothing parameter is a variance parameter that is estimated by a likelihood-based selector, such as REML (Wood, 2011). In smoothing parameter selection, although model-based selectors have better asymptotic results, likelihood-based methods converge faster to optimal values of the smoothing parameter (Hardle et al., 1988). For more details, see Wood (2011). Additionally, the *BIC* is closely related to the *AICc*. Although the prediction error can be decreased by adding parameters, excessive parameter can result in overfitting. To overcome this issue, the *BIC* includes penalty term, as do the selectors *AICc*, *GCV* and REML. The *BIC* penalty is larger than that of the *AICc*. Note that Hurvich et al. (1998) illustrated that *AICc* performs well for small sample sizes. For details of the comparison of the *AICc* and *BIC*, see Burnham and Anderson (2004).

These selection criteria balance the complexity of an estimate of f against how well the model fits the data. When using an adequate number of knots, the value of the smoothing parameter controls the influence of the penalty in (11).

5. Selecting the number of knots

A spline model with a basis function has knot points. The number of knots, K, is usually unknown and must be estimated. The key idea is to choose enough knot points to estimate the regression function with a penalized spline based on a truncated power basis. There are many studies in the statistical literature on selecting knot points (for example, see Ruppert, 2002; Ruppert et al., 2003; Ke and Wang, 2001).

As stated above, one important step in fitting penalized splines is selecting the number and locations of the knots. We use the three selection algorithms examined in Ruppert et al. (2003) to select the number of knots.

5.1. Default selection method (DSM)

The key idea is to choose enough knots to resolve the basic structure in the nonparametric regression model with censored data (1). The default knot location is defined as

$$\kappa_k = \left(\frac{k+1}{K+2}\right) th\ sample\ quantile\ of\ the\ unique\ X_i \quad \text{for } k = 1, 2, \ldots, K, \tag{26}$$

and a simple default selection of K is

$$K = \min\left(\frac{1}{4} \times number\ of\ unique\ X_i, 35\right). \tag{27}$$

5.2. Myopic algorithm (MA)

A myopic algorithm is an iterative process based on a smoothing parameter selection criterion. The algorithm explores a sequence of candidate values of knots and stops when there is no improvement in the selection criterion. Suppose that we have a sequence of candidate values of $K\{(K_1, \ldots, K_6) = (5,\ 10,\ 20,\ 40,\ 80,\ 120)\}$ for sample size $n \leq 120$. Additionally, assume that $\lambda = (\lambda_1, \ldots, \lambda_m)$ is a vector of values of the smoothing parameters. As in Ruppert et al. (2003), we use GCV as the selection criterion in the knot selection procedure. For $K_j,\ \ j = 1, \ldots, 6$ the algorithm works as follows:

(1) The penalized spline fit is computed by using λ_1, which is chosen to minimize $GCV(\lambda)$ for $K_1 = 5$.

(2) The penalized spline fit is calculated based on λ_2, which is chosen to minimize $GCV(\lambda)$ for $K_2 = 10$.

(3) If $GCV(\lambda_2) > 0.98GCV(\lambda_1)$, then stop and use the number of knots corresponding to $\min(GCV(\lambda_1), GCV(\lambda_2))$. Otherwise, if $GCV(\lambda_2) \leq 0.98GCV(\lambda_1)$, repeat steps 1–3 for $j = 2, \ldots, 6$ until the stopping rule in step 3 is satisfied. For example, if the process is completed at $K_6 = 120$, then K_6 is considered to produce the best fit.

5.3. Full-search algorithm (FSA)

This algorithm is similar to the MA expressed in the previous section, but the full-search algorithm searches the entire sequence of possible knots and uses the value that minimizes the selection criterion. For $K_j,\ \ j = 1, \ldots, 6$, the algorithm proceed as follows:

1. The penalized spline fits are performed using the smoothing parameter λ_j, which is chosen by GCV for the knots $K_j,\ \ j = 1, \ldots, 6$.
2. The value of K_j that minimizes the $GCV(\lambda_j)$ criterion for $j = 1, \ldots, 6$ is selected.

We use the GCV criterion in the MA and FSA; however, any of the selection criteria described in the Section 4 can be used in the knot selection algorithms. Ruppert (2002) and Ruppert et al. (2003) provide more details about the knot selection methods.

6. Simulation experiment

In this section, we perform a simulation experiment to compare the knot selection algorithms when considering equally spaced knots for selecting λ and the number of knots (K). As indicated in Section 4, λ is selected by the GCV, $AICc$, BIC, and REML methods. We also use each of these parameter selection methods combined with each of the knot selection algorithms (DSM, MA, and FSA) in the presence of right censoring. Furthermore, we consider SS, which has a knot at each data point. The purpose of the simulation is to compare the relative efficiency and performance of the knot selection algorithms based on four smoothing parameter selection methods and to illustrate how well a selection approach works under different censoring conditions.

True survival times are generated by the following nonparametric regression model in generic form

$$Y_i = f(X_i) + \varepsilon_i = 2\sin(X_i) + 1.2\log(X_i^2 + 1) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{28}$$

where $\varepsilon_i \sim N(0,\ \sigma^2 = 0.5)$ and $X_i = 15[(i - 0.5)/n]$. The censoring time variable C is simulated using different normal distribution functions. For each simulation, we generate 1000 random samples of size $n = 50,\ 100,\ \mathit{and}\ 200$ based on the following conditions

Condition 1: $P(C) = |0.85 + 0.15\varphi|, \quad \mathit{if}\ \varphi <= 1\ \mathit{else}\ P(C) = 0.90$
Condition 2: $P(C) = |0.65 + 0.15\varphi|, \quad \mathit{if}\ \varphi <= 1\ \mathit{else}\ P(C) = 0.70$
Condition 3: $P(C) = |0.45 + 0.15\varphi|, \quad \mathit{if}\ \varphi <= 1\ \mathit{else}\ P(C) = 0.30$

where $\varphi_i = |X_i - 15|$. The censoring levels (CLs) corresponding to the preceding three conditions are approximately 10%, 30%, and 50%. Finally, based on the model with censored data in (28), we observe the simulated data $(Z_i,\ \delta_i)$ for $i = 1, \ldots, n$, where

$$Z_i = \min\left(Y_i,\ C_i\right)\ \mathit{and}\ \delta_i = I\left(Y_i \leq C_i\right).$$

Because of the censoring, the ordinary methods are not applied to estimate $f(X)$; therefore, we consider transformed response (or synthetic) data points, as described section in (2). Since these synthetic observation points depend on the unknown distribution of the censoring variable C, they are estimated by using the Kaplan and Meier (1958) estimator in (7). The estimate for the nonparametric regression model with censored data can then be obtained by minimizing criterion (11). In this step, theorem 1 serves as the basis for computing the fitted values, given by $\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{Z}_{\hat{G}}$ in (13).

6.1. Evaluation of the empirical findings

In this simulation study, many configurations are implemented to provide perspective of the adequacy of the above methods and approximations. Because 36 different configurations are analyzed, it is not possible to display the details of each configuration. Therefore, a selection of the simulation results, performed under varying conditions, is given in following tables and figures.

For each simulated dataset used in the experiments, we use the *MSE* values, which measure how close the predicted observations are to the real observations. Suppose we have a sample size of n; *MSE* can then be estimated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \hat{f}_\lambda(X_i)^2. \tag{29}$$

We choose regression models with small *MSE*, and boxplots of the replications of these MSEs are illustrated in Figure 1.

As shown in Figure 1, as the sample size n gets large, the range of penalized spline estimates decreases. The estimates from medium and large samples are more stable than those from small samples. On the other hand, although the values of the *MSE* replications differ for different algorithms, the general trend shows that as the censoring level increases, the range of the MSEs increases. Hence, censoring levels are far more efficient on large sample sizes.

Generally, censoring tends to increase the variance of the estimators. The precision decreases as the censorship level increases. In addition, the precision is improved as the sample size increases. To examine this case, the MSEs expressed in Eq. (29) are calculated from

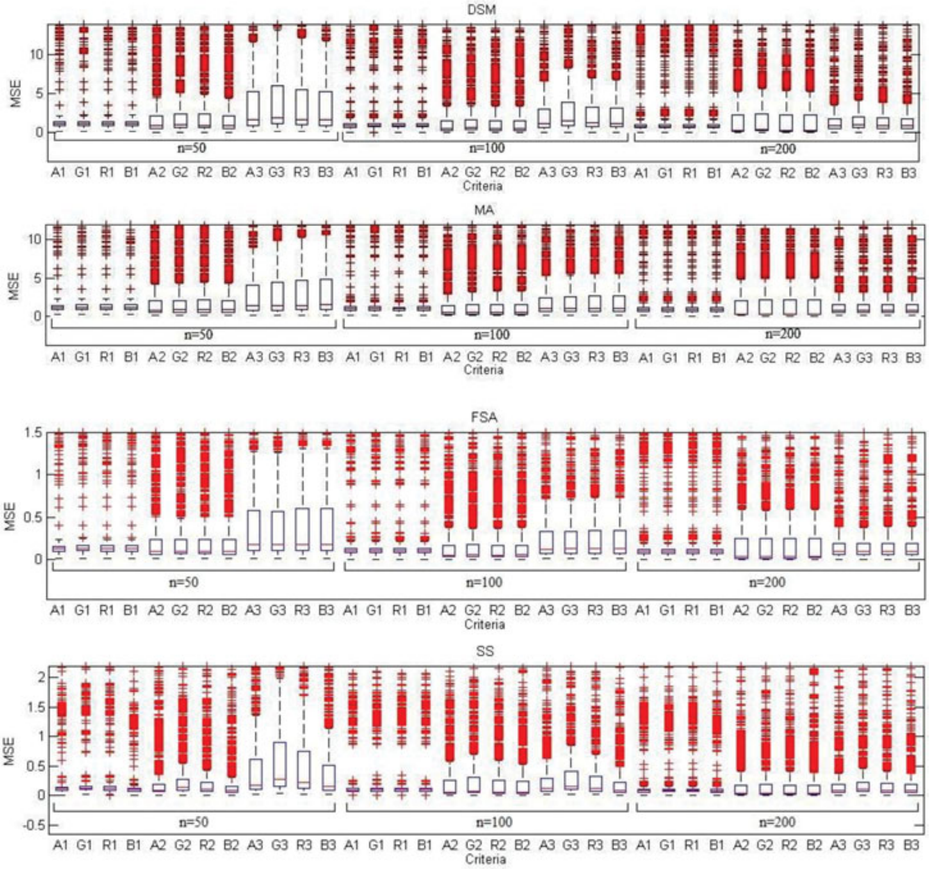

Figure 1. Boxplots of the MSEs from 1000 runs under different simulation scenarios. Upper panel: For each sample size A1, A2, and A3, the boxplots of the replications of the MSEs when penalized spline estimates based on the *AICc* criterion are constructed using the knot points determined by DMS for the three censoring levels of 10%, 30%, and 50%. In a similar fashion, G1, G2, and G3 represent the boxplots of the *MSE* replications based on the *GCV*, R1, R2, and R3 represent REML and B1, B2, and B3 denote the *BIC*. From top to bottom, the remaining panels are the same as the first panel but are for MA, FSA, and SS.

the penalized spline fits for each knot selection algorithm, criterion, sample, and censoring level. The outcomes from the simulation study are illustrated in Table 1.

The values in Table 1 are the means of the MSEs over 1000 simulation runs, with the average number of knots in parentheses. Furthermore, since DSM chooses $K = 12$, 25 and 50 knots for $n = 50$, 100 and 200, respectively, the number of knots from DSM is not given. As indicated in the introduction to Section 6, SS sets the number of knots equal to the sample size n; therefore, this information is not included in Table 1.

According to the results in Table 1, for all censoring levels with small sample sizes, the MA has better empirical performance than the other methods. However, the MA chooses 5 or 6 knot points in most of the simulation examples. MA with 5 knots is no better than FSA with the same number of knots, especially for medium and large sample sizes. The K provided by MA remains approximately unchanged as the sample size increases because the MA stops the node selection process prematurely. These findings are in accordance with those of Ruppert et al. (2003), who compared similar knot selection algorithms for uncensored data.

Table 1. MSE values and the number of knots chosen by the knot selection algorithms based on the criteria discussed in Section 4.

		$n = 50$			$n = 100$			$n = 200$		
	CLs	10%	30%	50%	10%	30%	50%	10%	30%	50%
DSM	*AIC*	0,236	0,391	0,619	0,177	0,203	0,344	0,156	0,202	0,251
	GCV	0,242	0,424	0,707	0,179	0,206	0,410	0,157	0,204	0,277
	REML	0,238	0,403	0,651	0,178	0,205	0,363	0,156	0,203	0,259
	BIC	0,238	0,394	0,624	0,178	0,205	0,347	0,156	0,203	0,253
MA	*AIC*	0,234	0,366	0,549	0,176	0,203	0,303	0,156	0,202	0,225
		(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)	(5)
	GCV	0,237	0,379	0,577	0,178	0,205	0,314	0,156	0,203	0,228
		(5,4)	(5,5)	(5,5)	(5,3)	(5,4)	(5,3)	(5,1)	(5,2)	(5,1)
	REML	0,237	0,379	0,583	0,178	0,204	0,315	0,156	0,203	0,229
		(5,4)	(5,4)	(5,3)	(5,3)	(5,3)	(5,2)	(5,1)	(5,2)	(5,1)
	BIC	0,238	0,380	0,584	0,178	0,204	0,317	0,156	0,203	0,229
		(5,4)	(5,5)	(5,5)	(5,3)	(5,4)	(5,3)	(5,1)	(5,2)	(5,1)
FSA	*AIC*	0,235	0,389	0,616	0,175	0,202	0,336	0,155	0,201	0,243
		(11)	(11)	(11)	(14)	(14)	(14)	(17)	(17)	(17)
	GCV	0,239	0,395	0,611	0,177	0,203	0,333	0,155	0,201	0,241
		(8,7)	(7,1)	(6,8)	(8,7)	(7,6)	(7)	(9)	(8,5)	(7,5)
	REML	0,238	0,397	0,628	0,176	0,209	0,34	0,155	0,201	0,244
		(9,6)	(8,9)	(8,6)	(10,7)	(10)	(9,6)	(11,9)	(11,2)	(10,8)
	BIC	0,238	0,395	0,627	0,176	0,203	0,34	0,155	0,201	0,245
		(5,4)	(5,5)	(5,5)	(5,3)	(5,4)	(5,3)	(5,1)	(5,2)	(5,1)
SS	*AIC*	0,248	0,448	0,647	0,207	0,233	0,355	0,191	0,206	0,248
	GCV	0,263	0,446	0,853	0,211	0,265	0,441	0,193	0,207	0,269
	REML	0,251	0,391	0,726	0,209	0,243	0,387	0,192	0,201	0,249
	BIC	0,232	0,388	0,653	0,165	0,226	0,349	0,151	0,199	0,248

By comparing the FSA to SS, we see that FSA has good performance, especially for sample sizes of 100 and 200. The FSA also provides a lower *MSE* than that of SS for almost all censoring levels. Additionally, the FSA has the advantage that it usually requires a shorter computation time than that of SS. In other words, it is extremely fast, especially for a censoring level of 50%, because an optimum estimate of $f(X)$ in (13) is achieved with a small number of nodes selected via FSA when λ is chosen based on the *AICc*.

Table 1 reveals that applying FSA to the observations with a censorship rate of 50% yields superior estimates to those obtained for censoring levels of 10% and 30% for all simulation examples. The FSA slightly outperforms the other node selection algorithms in terms of *MSE*. On the other hand, despite the differences in the number of knots, MA and DSM perform similarly. However, at high censoring levels, the FSA is faster.

As indicated in the introduction to this section, we use the MSEs to assess the quality of the regression functions. Additionally, we use paired Wilcoxon tests to determine whether the difference between the median of the MSEs of any two knot selection methods is statistically significant at a significance level of 5%. If the median *MSE* value of a method is significantly less than those of the remaining four methods, it is assigned a rank of 1. If the median *MSE* value of a method is significantly larger than one but less than those of the other three, it is assigned a rank of 2, and analogously for ranks 3-4. Methods with non-significantly different median values share the same averaged rank. The key idea is to specify which criteria (smoothing parameter selection methods) are most appropriate for the knot selection algorithms. We focus on only a selection of the simulation configurations related to FSA and SS due to a lack of space. Several examples of such configurations are shown in Figures 2 and 3. The numbers below the boxplots in Figure 2 graphically display the results from the summary in Table 2 for select simulation scenarios related to FSA and SS. Furthermore, the results of all 36 simulation experiments are given in Tables 2 and 3.

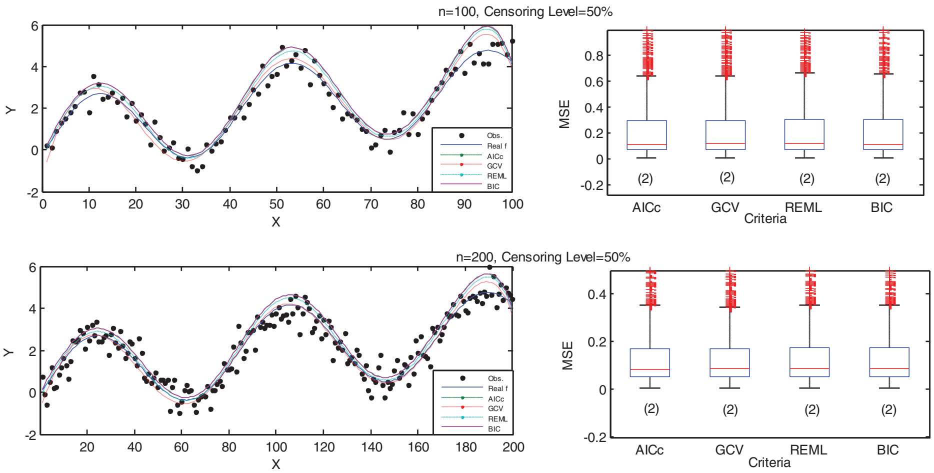

Figure 2. Top left panel shows the observations, true regression function, and four estimated curves from *AICc*, *GCV*, REML, and *BIC* for the knot points determined by FSA with CL = 50% and $n = 100$. The top right panel shows boxplots of the MSEs for these fitted curves. The numbers below the boxplots are the paired Wilcoxon test rankings. The bottom panels are similar to the upper panels but for $n = 200$.

The left panels of Figure 2 show four penalized spline fits of the simulated data for different values of λ and with knot points determined by the FSA. Each panel presents a single realization of simulated data and hence different fitted curves. The left panels show boxplots of the *MSE* replications of the obtained fitted values based on the different λ values chosen by the four criteria. The numbers under the boxplots indicate the Wilcoxon test rankings based on the median of the MSEs. From Figure 2, we see that all four penalized spline fits follow approximately the same pattern and are quite satisfactory. The four criteria (*AICc*, *GCV*, REML, and *BIC*) share the same performance orderings for medium and large sample sizes under a censoring level of 50%; therefore, the criteria for selecting λ have approximately equivalent performance in terms of MSEs when using knots of the FSA.

As shown in Figure 3, as the sample sizes get large, the penalized spline estimates approach the real regression function stated in (28). However, the boxplots of the *MSE* values in the

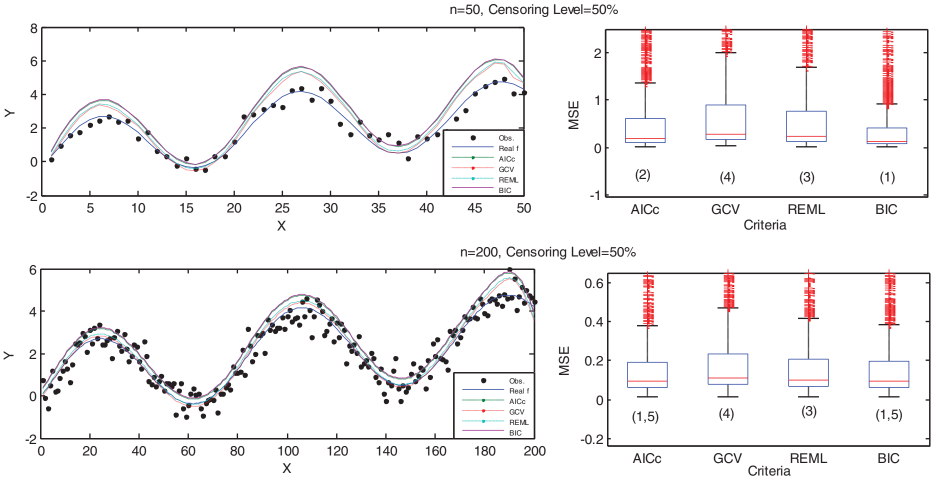

Figure 3. Similar to Figure 2 but for SS with sample sizes of $n = 50$ and 200.

Table 2. Wilcoxon test rankings related to the criteria and algorithms.

		$n = 50$			$n = 100$			$n = 200$			
	CLs	10%	30%	50%	10%	30%	50%	10%	30%	50%	Overall
DSM	*AICc*	1,5	1,5	1,5	2,0	1,5	1,0	2,0	1,5	1,5	1,56*
	GCV	4,0	4,0	4,0	2,0	4,0	4,0	2,0	4,0	4,0	3,56
	REML	1,5	1,5	1,5	2,0	1,5	2,5	2,0	1,5	1,5	1,72
	BIC	1,5	1,5	1,5	2,0	1,5	2,5	2,0	1,5	1,5	1,72
MA	*AICc*	2,0	2,0	1,0	2,0	2,0	1,0	2,0	2,0	1,0	1,67*
	GCV	2,0	2,0	2,5	2,0	2,0	2,5	2,0	2,0	2,5	2,17
	REML	2,0	2,0	2,5	2,0	2,0	2,5	2,0	2,0	2,5	2,17
	BIC	2,0	2,0	2,5	2,0	2,0	2,5	2,0	2,0	2,5	2,17
FSA	*AICc*	2,0	1,0	2,0	2,0	1,5	2,0	2,0	2,0	2,0	1,83*
	GCV	2,0	2,5	2,0	2,0	4,0	2,0	2,0	2,0	2,0	2,28
	REML	2,0	2,5	2,0	2,0	1,5	2,0	2,0	2,0	2,0	2,00
	BIC	2,0	2,5	2,0	2,0	1,5	2,0	2,0	2,0	2,0	2,00
SS	*AICc*	1,5	1,5	2,0	2,0	1,5	1,5	2,0	1,5	1,5	1,67
	GCV	4,0	4,0	4,0	2,0	4,0	4,0	2,0	4,0	4,0	3,56
	REML	1,5	3,0	3,0	2,0	3,0	3,0	2,0	3,0	3,0	2,61
	BIC	1,5	1,5	1,0	2,0	1,5	1,5	2,0	1,5	1,5	1,56*

(∗):Under each algorithm, indicates the optimum selection criteria having the best rankings.

top right panel of Figure 3 show that *BIC* provides the best fitted values, especially for small sample sizes. The bottom right boxplot indicates that *AICc* and *BIC* have approximately equal performance due to the effect of the simulation replications.

An important step in this stage is to determine which criteria to use for the knot selection algorithms. Table 2 is constructed based on the rankings of the median values of the MSEs (see the right panels of Figs. 2 and 3) under each censoring level and sample size. According to Table 2, when 10% of the data are censored, the criteria perform equally, especially for the MA and FSA. As the sample size and censoring level increase, these criteria are also more stable for the MA and FSA than for the other two algorithms. The criteria are not stationary for DSM and SS when the simulated datasets are censored at rates of 30% and 50%.

In summary, according to the overall Wilcoxon test rankings in Table 2, the *AICc* provides the desired smoothing parameter λ for the DSM, MA and FSA. For the remaining algorithm (SS), the optimum value of λ is chosen by the *BIC*. Thus, the *AICc* and *BIC* are the optimum selection criteria. In Table 2, these criteria are indicated with an asterisk for each algorithm.

From the information given above, by employing the *AICc*, we obtain the penalized spline fits at the knot points determined by the DSM, MA and FSA. Additionally, to obtain fitted values at all data points (that is, for SS), the smoothing parameter is obtained by the *BIC*. The spline fits for some of the simulation scenarios are displayed in Figure 4. Furthermore, the fits for $n = 50$ are similar to those for $n = 100$ and are not given here. The goal is to compare the

Table 3. Wilcoxon test scores obtained according to the knot selection methods.

n	CLs	DSM	MA	FSA	SS
50	10%	2,5	4,0	2,5	1,0
	30%	1,5	1,5	1,5	4,0
	50%	2,5	1,0	2,5	4,0
100	10%	2,5	4,0	2,5	1,0
	30%	1,5	1,5	1,5	4,0
	50%	2,5	1,0	2,5	4,0
200	10%	2,5	4,0	2,5	1,0
	30%	2,5	4,0	1,0	2,5
	50%	1,5	1,5	1,5	4,0
Average Scores		2,17	2,50	2,00	2,83

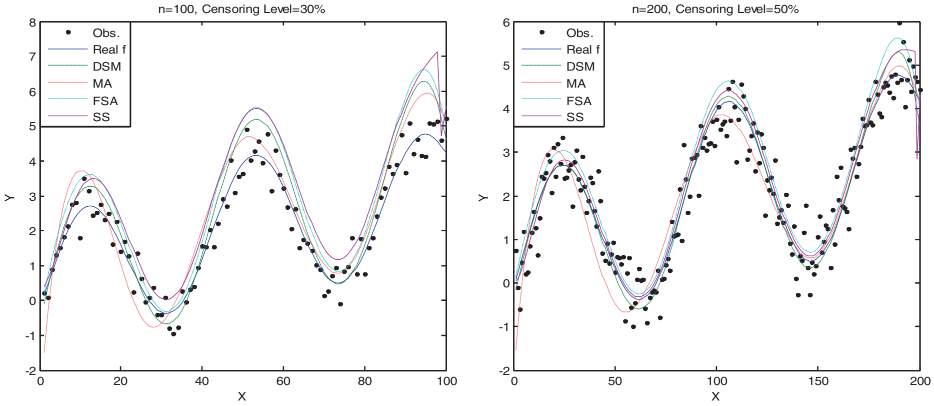

Figure 4. For medium and large sample sizes with CL = 30% and 50%, each panel shows the observations, real regression function, and fitted curves based on the optimum criteria for the DSM, MA, FSA, and SS.

fitted values from the algorithms for various censoring levels and samples sizes. The results appear to be quite reasonable for large sample sizes with high censoring levels; however, as noted previously, the estimated curves are not good for small sample sizes.

For each sample size and censoring level, we compute the penalized spline fits based on the optimum criteria for the DSM, MA, FSA, and SS. The DSM, MA and FSA are compared to the estimates from SS using the optimum criteria (*AICc* and *BIC*). To this end, the MSEs of these algorithms are calculated, and the Wilcoxon test rankings of the median values of the MSEs are presented in Table 3.

According to the results in Table 3, the FSA is superior for all sample sizes and censoring levels. The DSM shows similar behavior to the FSA. As the sample size increases, the results become increasingly similar. In general, the MA method has the worst performance, and the average scores show that the FSA is the best knot selection algorithm in this simulation study.

6.2. Efficiency comparisons

To make inferences regarding the change in efficiency of the estimators due to the different algorithm and criterion combinations, we use the scalar-valued SMDEs in (23). As described in Definition 3.2, the relative efficiencies (REs) are constructed based on the *SMDE* ratios for datasets with different censoring levels. The key idea is to track the changes in the efficiencies of the combined estimators for different sample sizes. The obtained RE ratios are illustrated in Figures 5 and 6, and they are analyzed in terms of both the knot selection algorithms and the smoothing parameter selection criteria.

Figure 5 shows the overall relative efficiencies of the algorithms (DSM, MA, FSA, and SS). The DSM, MA, and FSA are used to select the optimum K knots in the penalized spline based on a smoothing parameter that is selected based on the *AICc*, *GCV*, REML, or *BIC*. Both panels of Figure 5 show that efficiencies of the DSM, MA and FSA are higher than those of the SS for samples with CL = 10%, but the efficiency of SS increases more rapidly as the sample size increases, especially for a censoring level of 50%.

Figure 6 shows the overall relative efficiency of each criterion within each knot selection algorithm in terms of the *SMDE*. The left panel of Figure 6 shows that the *AICc* dominates the other criteria for all sample sizes with CL = 10%. Additionally, the efficiency of the *BIC* shows similar behavior to that of the *AICc*. The right panel of Figure 6 reveals that for CL = 50%, the *AICc* remains superior; therefore, for censored data, the *AICc* has the best performance,

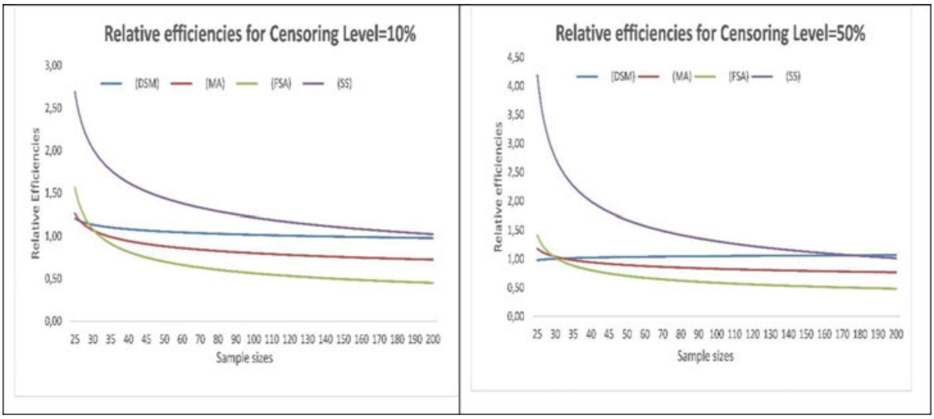

Figure 5. Overall relative efficiencies of the algorithms for different sample sizes with the censoring levels of 10% and 50%.

making it the ideal selection method for nonparametric regression using the penalized spline method. The simulated relative efficiencies of *GCV* are not good, especially for small sample sizes with CL = 10%. In a similar fashion, the relative efficiency of the *BIC* is not good for small sample sizes with CL = 50%.

The graphical results indicate a very good relative efficiency for the FSA for all censored datasets. Therefore, the FSA is an efficient knot selection algorithm, but it is not efficient on samples smaller than n = 30.

7. Real data example

We now present a real data example to compare the knot selection methods numerically. The data in panel (a) of Figure 7 are the survival time and albumin (i.e., the most common protein found in the blood) values of patients with colon cancer admitted to a hospital in Izmir, Turkey. In this example, the logarithm of the survival time is considered as the response variable (*SurvTime*), while the albumin value is used as an explanatory variable (*Albumn*). The key idea is to explain the relationship between these two variables using the model

$$\log(SurvTime_i) = f(Albumn_i) + \varepsilon_i, \quad i = 1, \ldots, 97. \tag{30}$$

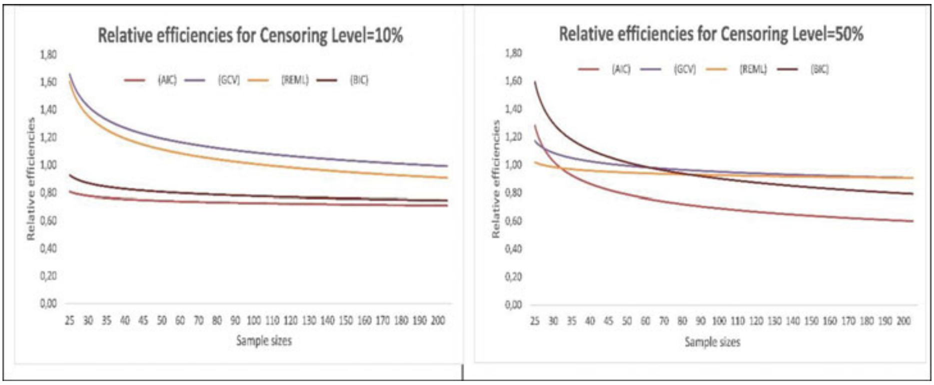

Figure 6. Overall relative efficiencies of the criteria for different sample sizes with censoring levels of 10% and 50%.

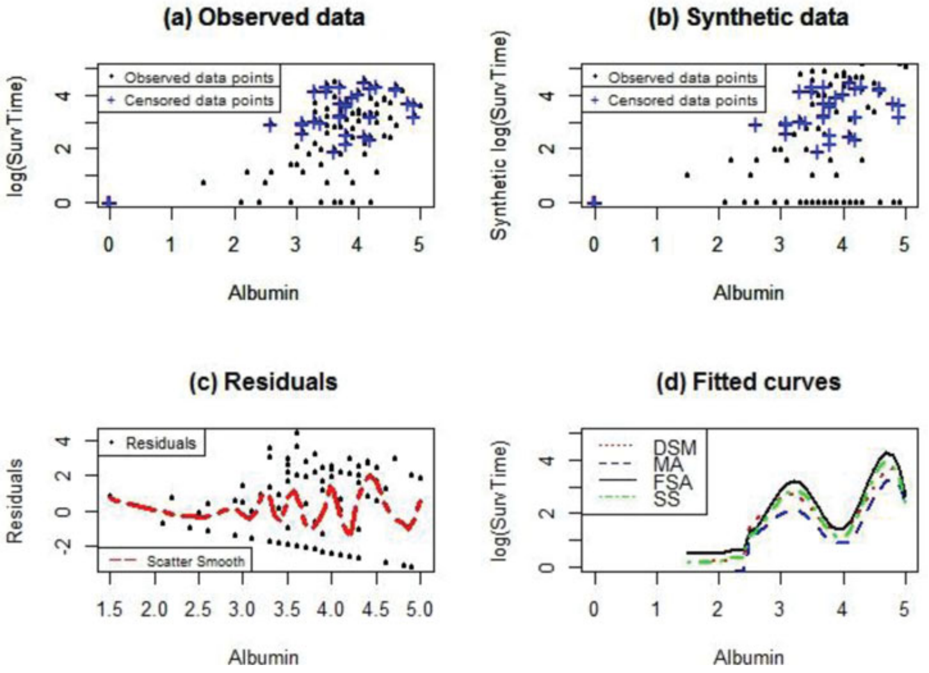

Figure 7. Colon cancer data: (a) Scatterplot of the $\log_{10}$ (survival times) against albumin values: Censored patients are denoted by "+", and the failure times (or observed lifetimes) are indicated by " • ". (b) Scatterplot of the synthetic data versus albumin values. The same symbols are used for the synthetic data. (c) Residuals from the regression of survival times on albumin together with the scatterplot smooth. (d) Real observations and fitted curves.

The 97 patients with complete records are used for the analysis. Of these 97 patients 32 are right censored for various reasons, including sudden death or withdrawal from the experiment, and the remaining 65 patients are dead, i.e., uncensored. The censoring rate is 32.99%.

The comparative outcomes of model (30) with the censored colon cancer data are summarized in the following Figures and Tables.

The scatterplot of *SurvTime* versus *Albumn* is shown in panel (a) of Figure 7. Panel (b) is similar to panel (a), showing the scatterplot of $SurvTime_{\hat{G}}$ against *Albumn*. Censored observations are denoted by "+", while uncensored data points are indicated by " • ". Additionally, the residuals are plotted against *Albumn* in panel (c) of Figure 7 to see the shape of the functional relationship between *SurvTime* and *Albumn*. The residuals are obtained from the results of the censored least squares regression fits to the data (see Qin and Jing, 2001). When a curve is added to the scatterplot smooth, it is clear that there is nonlinearity between the two variables, indicating that nonparametric regression will give more reasonable results. Finally, the penalized spline fits for the DSM, MA, FSA, and SS, based on the optimum criteria from the in simulation section, are depicted in panel (d). These fits show that the effect of albumin on the survival time depends on the value of albumin. In other words, the marginal effect of albumin is non-constant.

Figure 8 shows four curves fitted to the colon cancer data obtained by using (17) with smoothing parameter λ selected by the *AICc*, *GCV*, REML, and *BIC* under each knot selection algorithm. Each fit depends on the set of knot points and the smoothing parameter λ. Figure 8 shows that the fitted curves from the FSA, MA, and DSM are smoother than those obtained by SS, which uses all knot points. Although the knot points cover the range of

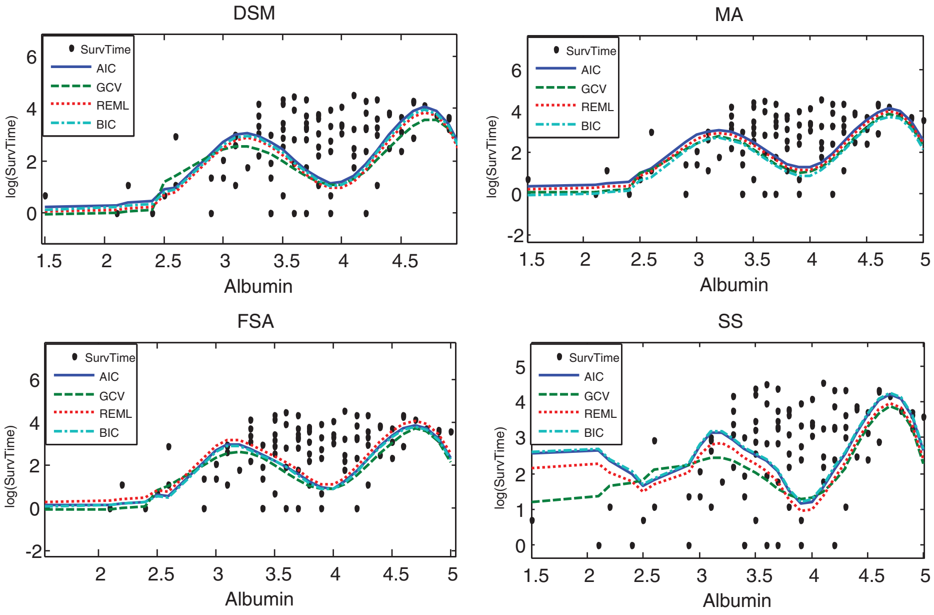

Figure 8. Penalized spline regression fits to the colon cancer data with different smoothing parameters for each algorithm.

Albumin values reasonably well, they do not have a substantial effect on the fitted curves. However, the smoothing parameter λ has a very strong effect, as shown in the bottom right panel of Figure 8, where the fits from the colon cancer data are shown for the values of λ chosen by the four criteria with the all knot points used.

Figure 8 also compares the knot selection algorithms with the smoothing parameter selection criteria on the censored colon cancer data. The estimates made when using all knot points are not good. In this context, the number of knots and the SMDEs from the fits for each algorithm and criterion are given in Table 4. As defined in Eq. (23), these *SMDE* values decompose into a sum of the squared bias and variance of the estimator. These variance and bias values are also presented in Table 5.

Table 4. The SMDE values and number of knots (*K*).

	DSM	*K*	MA	*K*	FSA	*K*	SS	*K*
AIC	1,680	6	1,321	6	1,065	5	2,974	97
GCV	2,059	6	2,451	5	1,321	8	3,083	97
REML	2,059	6	2,451	5	1,321	8	3,083	97
BIC	1,680	6	1,321	5	1,065	52	2,974	97
Means	1,869		1,886		**1,193**		3,029	

Table 5. The variance and biases of the algorithms.

	DSM		MA		FSA		SS	
	Variance	Bias	Variance	Bias	Variance	Bias	Variance	Bias
AIC	0,600	1,039	0,575	0,864	0,703	0,602	0,645	1,526
GCV	0,551	1,228	0,533	1,385	0,575	0,864	0,662	1,556
REML	0,551	1,228	0,533	1,385	0,575	0,864	0,662	1,556
BIC	0,600	1,039	0,575	0,864	0,703	0,602	0,645	1,526

For $K = 97$ knots, the *SMDE* of the SS is substantially higher than those of the other algorithms, which is not surprising because the SS considers all values of K. Table 4 shows that the *SMDE* values from FSA are equal for the *AICc* and *BIC*. Moreover, using the *AICc* or *BIC* to select a smoothing parameter produces smoother fits than using *GCV* and REML. Specifically, if the smoothing parameter is chosen via the *AICc*, one can obtain a fit with $K = 5$ knots selected by the FSA (see Table 4). This case indicates that the FSA is extremely fast. Additionally, according to the means given in Table 4, the FSA provides a better approximation than the other algorithms. As shown in Table 5, this means that the fitted values from the FSA using the *BIC* have more variance and less bias than those associated with *GCV* and REML; however, the fitted values will be approximately equally accurate in terms of the *SMDE* (see the means in Table 4).

8. Conclusions and recommendations

The simulation and real data results are satisfactory in general. As the sample size increases, the right-censored nonparametric model produces a closer fit to the real observations. The estimates from medium and large sample sizes are more stable than those from small sample sizes. Furthermore, as the censoring level increases, the range of MSEs of the fitted values becomes large; hence, the censoring levels are far more efficient on sample sizes.

The overall Wilcoxon test rankings show that the *AICc* gives the optimum smoothing parameter λ for the penalized spline fits based on knot points selected by the DSM, MA, and FSA, while the *BIC* provides the optimum λ for the SS. For penalized spline fits based on a smoothing parameter chosen by the *AICc*, the FSA slightly outperforms the other knot-selection algorithms in terms of SMDEs. Additionally, the fitted curves of the FSA, MA, and DSM are smoother than those of the SS, which uses all the knot points. Moreover, the smoothing parameter λ has a substantial effect in obtaining these smooth curves, but the knot points do not have a substantial effect.

The fitted curves using *GCV* and REML do not yield good performance in the estimation of the nonparametric regression model. Furthermore, the penalized spline fits based on these criteria give similar results.

Finally, based on the simulation and real data results, the following suggestions should be considered:

- Under response observations with right censoring, the *AICc* is recommended as the selection criterion for penalized spline regression fits, especially for knot points determined by the FSA.
- For SS, which considers all the data points, the *BIC* criteria is most appropriate.
- Although DSM works well in most of the censored data experiments, it does not use any information from the data except the sample size. Therefore, an algorithm that considers the data to select the number of knots should be used instead of the DSM.
- Since the knot selection process in the MA is based on an early stopping strategy, using the FSA, which selects the optimum number of knots and provides a much better fit, instead of the DSM, would be beneficial.

A. Appendix: Supplemental technical materials

The following assumption is required for the proof of strong consistency.

Assumption A1. The parameter space Θ is compact. For given G, the mean lifetime of the *ith* observation should be $E(Z_{iG}|\mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta} = \mu_i(\boldsymbol{\beta})$. It is also noted that this mean function $\mu(\boldsymbol{.})$

is continuous on Θ, $(1/n)\sum_{i=1}^{n}\{\mu_i(\boldsymbol{\beta})-\mu_i(\tilde{\boldsymbol{\beta}})\}^2$converges uniformly to some limit in $\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}} \in \Theta$ and $PRSS(\lambda;\, \boldsymbol{\beta}) = lim(1/n)\sum_{i=1}^{n}\{\mu_i(\boldsymbol{\beta}_0)-\mu_i(\boldsymbol{\beta})\}^2$ exists and has a unique minimum at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$.

Asymptotic normality can be established under the following additional assumption.

Assumption A2. The true parameter vector $\boldsymbol{\beta}_0$ is an interior point of Θ, the mean function $\mu(.)$ is twice continuously differentiable in a neighborhood of $\boldsymbol{\beta}_0$, and $\Omega(\boldsymbol{\beta}_0) = \lim(\frac{1}{n})\sum_{i=1}^{n}((\partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta})|_{\boldsymbol{\beta}_0})((\partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta})|_{\boldsymbol{\beta}_0})' = lim_n(1/n)\mathbf{X}'\mathbf{X}$ exists and is nonsingular and converges uniformly in $\boldsymbol{\beta}$ in an open neighborhood of $\boldsymbol{\beta}_0$.

A1. Proof of Theorem 1

Let's consider the model $Z_{\hat{G}} = \mathbf{X}\boldsymbol{\beta} + \varepsilon_{\hat{G}}$. The vector of ordinary least squares residuals can be written as $\boldsymbol{\varepsilon}_{\hat{G}} = (\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta})$ and hence

$$\boldsymbol{\varepsilon}_{\hat{G}}{}'\boldsymbol{\varepsilon}_{\hat{G}} = \left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right)$$

The ordinary least squares regression fits can be described as

$$\hat{\mathbf{Z}}_{\hat{G}} = \mathbf{X}\hat{\boldsymbol{\beta}} \text{ where } \hat{\boldsymbol{\beta}} \text{ minimizes } (\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta})$$

and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_{p+K})'$, with β_{p+k} the coefficient of the kth knot. As known, unrestricted estimation of the β_{p+k} leads to a wiggly fit. We assume that constraint on β_{p+k} the coefficient is $\sum\beta_{p+k}^2 < C$, where $C > 0$. Also, let $\mathbf{S}_\lambda$ be a positive definite and symmetric matrix, and $\mathbf{D}$ be a $(K+2)\times(K+2)$ diagonal penalty matrix whose first $(p+1)$ elements are 0, and the remaining elements are 1. Then, the criterion function (11) relating the model (9) can be written as

$$\text{minimizes } \left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right) \text{ subject to } \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \le C.$$

Using a Lagrange multiplier argument this expression is also equivalent to the criterion (11). As previously stated, this criterion is

$$PRSS(\lambda;\ \boldsymbol{\beta}) = \left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right) + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \tag{A1.1}$$

Simplifying

$$\begin{aligned} PRSS(\lambda;\ \boldsymbol{\beta}) &= \left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{Z}_{\hat{G}} - \mathbf{X}\boldsymbol{\beta}\right) + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \\ &= \mathbf{Z}'\mathbf{Z}_{\hat{G}} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Z}_{\hat{G}} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \end{aligned}$$

Since $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Z}_{\hat{G}}$is 1×1, $\boldsymbol{\beta}'\mathbf{X}'\mathbf{Z}_{\hat{G}} = (\boldsymbol{\beta}'\mathbf{X}'\mathbf{Z}_{\hat{G}})' = \mathbf{Z}'_{\hat{G}}\mathbf{X}\boldsymbol{\beta}$. By substitution,

$$\begin{aligned} PRSS(\lambda;\ \boldsymbol{\beta}) &= \mathbf{Z}'_{\hat{G}}\mathbf{Z}_{\hat{G}} - 2\mathbf{Z}'_{\hat{G}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X}\right)\boldsymbol{\beta} + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \\ &= \mathbf{Z}'_{\hat{G}}\mathbf{Z}_{\hat{G}} - 2\left(\mathbf{X}'\mathbf{Z}_{\hat{G}}\right)'\boldsymbol{\beta} + \boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X}\right)\boldsymbol{\beta} + \lambda\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \end{aligned} \tag{A1.2}$$

In order to minimize (A1.1), we could differentiate (A1.2) with respect to $\boldsymbol{\beta}$ and set the derivative equal to zero:

$$\frac{\partial\, PRSS(\lambda;\ \boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\,2\mathbf{X}'\mathbf{Z}_{\hat{G}} + 2\boldsymbol{\beta}\left(\mathbf{X}'\mathbf{X}\right) + 2\lambda\boldsymbol{\beta}\mathbf{D} = 0 \tag{A1.3}$$

Setting (A1.3) equal to zero and replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, we see that the penalized least squares normal equations are obtained as

$$\left(\mathbf{X}'\mathbf{X}+\lambda\,\mathbf{D}\right)\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Z}_{\hat{G}} \tag{A1.4}$$

To solve for $\hat{\boldsymbol{\beta}}$, multiply each side of the equation (A1.4) by $(\mathbf{X}'\mathbf{X}+\lambda\,\mathbf{D})^{-1}$ to obtain the penalized least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{D}\right)^{-1}\mathbf{X}'\mathbf{Z}_{\hat{G}}$$

Thus, fitted values are given by

$$\hat{\mathbf{Z}}_{\hat{G}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}+\lambda\mathbf{D}\right)^{-1}\mathbf{X}'\mathbf{Z}_{\hat{G}} = \mathbf{S}_{\lambda}\mathbf{Z}_{\hat{G}} = \hat{\mathbf{f}}_{\lambda}$$

as claimed.

A2. *Derivation of the equation (19)*

We begin by considering the general definition of quadratic form, Theorem and Lemmas for proof of the equation (19)

Definition A1.1: Let $\mathbf{S}_{\lambda}$=$[s_{ij}]$ be a positive semi-definite and symmetrical $n \times n$ matrix depending on the λ; and $\boldsymbol{\varepsilon}$ **=**$(\varepsilon_1, \ldots, \varepsilon_n)'$ be $n \times 1$ vector of random variables. Then

$$q = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\varepsilon_i\varepsilon_j = \boldsymbol{\varepsilon}'\mathbf{S}_{\lambda}\boldsymbol{\varepsilon} \tag{A1.5}$$

is a called a quadratic form in $\boldsymbol{\varepsilon}$ and $\mathbf{S}_{\lambda}$ is a called the matrix of a quadratic form.

Theorem A1.1: If $E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = Cov(\boldsymbol{\varepsilon}) = \sum$ **=**(σ_{ij}), and $E(\boldsymbol{\varepsilon}) = 0$, then

$$E\left(\boldsymbol{\varepsilon}'\mathbf{S}_{\lambda}\boldsymbol{\varepsilon}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\,\sigma_{ij} = tr\left(\mathbf{S}_{\lambda}\sum\right)$$

where $tr(A)$ denotes trace of the matrix (A)
Proof

$$E\left[(\boldsymbol{\varepsilon}-E(\boldsymbol{\varepsilon}))'\mathbf{S}_{\lambda}\,(\boldsymbol{\varepsilon}-E(\boldsymbol{\varepsilon}))\right] = E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\left(\varepsilon_i - E(\varepsilon_j)\right)'\left(\varepsilon_i - E(\varepsilon_j)\right)\right]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}E\left[\left(\varepsilon_i - E(\varepsilon_j)\right)'\left(\varepsilon_i - E(\varepsilon_j)\right)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}Cov\left(\varepsilon_i, \varepsilon_j\right) = tr\left(\mathbf{S}_{\lambda}\sum\right)$$

as claimed. □

Theorem A1.2: Let $\boldsymbol{\varepsilon}$ be an $n \times 1$ random vector with $E(\boldsymbol{\varepsilon}) = \boldsymbol{\mu}$ and $Cov(\boldsymbol{\varepsilon}) = \sum = (\sigma_{ij})$. Let $\mathbf{S}_{\lambda}$ be a $n \times n$ constant matrix. Then, the expected value of the equation (A1.5) is

$$E(\boldsymbol{\varepsilon}'\mathbf{S}_{\lambda}\boldsymbol{\varepsilon}) = \boldsymbol{\mu}'\mathbf{S}_{\lambda}\,\boldsymbol{\mu} + tr\left(\mathbf{S}_{\lambda}\sum\right) \tag{A1.6}$$

Proof. It is well known that for $i \neq j$,

$$\sigma_{ij} = E\left(\varepsilon_i\varepsilon_j\right) - \mu_i\mu_j$$

and that for $i = j$,

$$\sigma_{ij} = \sigma_{ii} = E\left(\varepsilon_i^2\right) - \mu_i^2 = \sigma_i^2$$

According to (A1.6), the expected value of the quadratic form $\boldsymbol{\varepsilon}'\mathbf{S}_{\lambda}\boldsymbol{\varepsilon}$ in expanded form is

$$E\left(q\right) = E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\varepsilon_i\varepsilon_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} E\left(s_{ij}\varepsilon_i\varepsilon_j\right)$$

Since $\sigma_{ij} = E(\varepsilon_i\varepsilon_j) - \mu_i\mu_j$, we obtain $E(\varepsilon_i\varepsilon_j) = \sigma_{ij} + \mu_i\mu_j$
Substituting,

$$E\left(q\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}E\left(\varepsilon_i\varepsilon_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\left(\sigma_{ij} + \mu_i\mu_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\,\sigma_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\,\mu_i\mu_j \tag{A1.7}$$

Note also that the terms σ_{ij} are the elements of the variance-covariance matrix $\sum$. This matrix is a symmetric matrix whose ith element is the variance of ε_i and whose (ij)th off-diagonal element is the covariance between ε_i and ε_j.

It follows from (A1.6), and theorem 1 that the equation (A1.7) is equivalent to

$$E\left(\boldsymbol{\varepsilon}'\mathbf{S}_\lambda\boldsymbol{\varepsilon}\right) = tr\left(\mathbf{S}_\lambda\sum\right) + \boldsymbol{\mu}'\mathbf{S}_\lambda\,\boldsymbol{\mu}$$

This completes the proof of theorem A1.2.

Again, let's consider equation (13)

$$RSS\left(\lambda\right) = \left(\hat{\mathbf{f}}_\lambda - \mathbf{Z}_{\hat{G}}\right)'\left(\hat{\mathbf{f}}_\lambda - \mathbf{Z}_{\hat{G}}\right) = \mathbf{Z}_{\hat{G}}\left(\mathbf{I} - \mathbf{S}_\lambda\right)^2\mathbf{Z}_{\hat{G}}$$

Thus, from Theorems A1-A2 connected with quadratic form, the expected value of the $RSS(h)$ is stated as

$$\begin{aligned}
E\left[RSS(\lambda)\right] = MSE(\lambda) = E\left\|\hat{\mathbf{f}}_\lambda - \mathbf{Z}_{\hat{G}}\right\|^2 &= E\left\|(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}_{\hat{G}}\right\|^2 \\
&= E\left\|\mathbf{Z}_{\hat{G}}(\mathbf{I} - \mathbf{S}_\lambda)(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}_{\hat{G}}\right\| = \mathbf{f}'_\lambda(\mathbf{I} - \mathbf{S}_\lambda)^2\mathbf{f}_\lambda + \sigma^2\left(\mathrm{tr}(\mathbf{I} - \mathbf{S}_\lambda)^2\right) \\
&= \mathbf{f}'_\lambda(\mathbf{I} - \mathbf{S}_\lambda)^2\,\mathbf{f}_\lambda + n\sigma - 2\sigma^2\mathrm{tr}(\mathbf{S}_\lambda) + \sigma^2\mathrm{tr}(\mathbf{S}'_\lambda\mathbf{S}_\lambda) \\
&= \mathbf{f}'_\lambda(\mathbf{I} - \mathbf{S}_\lambda)^2\,\mathbf{f}_\lambda + \sigma^2\left[n - 2\mathrm{tr}(\mathbf{S}_\lambda) + \mathrm{tr}(\mathbf{S}'_\lambda\mathbf{S}_\lambda)\right]
\end{aligned}$$

as defined in the equation (19). □

A3. *Proof of the Lemma 3.1*

$SMDE = E\|\mathbf{f} - \hat{\mathbf{f}}_\lambda\|$, where $\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda\mathbf{Z}_{\hat{G}}$ Then the scalar valued version of the MDE matrix can be specified as

$$\begin{aligned}
SMDE\left(\hat{\mathbf{f}}_\lambda\right) &= \sum_{i=1}^{n}\left[\left(f_i(X) - E\left(\hat{f}_{\lambda i}(X)\right)\right)\right]^2 + Cov\left[\hat{f}_{\lambda i}(X)\right] \\
&= \sum_{i=1}^{n}\left[f_i(X) - E\left(\mathbf{S}_\lambda\mathbf{Z}_{\hat{G}}\right)_i\right]^2 + Cov\left[\left(\mathbf{S}_\lambda\mathbf{Z}_{\hat{G}}\right)_i\right] \\
&= \sum_{i=1}^{n}\left[f_i(X) - E\left(\mathbf{S}_\lambda\mathbf{Z}_{\hat{G}}\right)_i\right]^2 + Cov\left[\left(\mathbf{S}_\lambda\mathbf{Z}_{\hat{G}}\right)\right]_{ii} \\
&= \|(\mathbf{I} - \mathbf{S}_\lambda)\,\mathbf{f}\|^{\,2} + tr\left[Cov\left(\mathbf{S}_\lambda\mathbf{Z}_{\hat{G}}\right)\right] \\
&= \|(\mathbf{I} - \mathbf{S}_\lambda)\,\mathbf{f}\|^{\,2} + tr\left[\mathbf{S}_\lambda Cov\left(\mathbf{Z}_{\hat{G}}\right)\mathbf{S}'_\lambda\right]
\end{aligned}$$

Assume that $Cov\left(\mathbf{Z}_{\hat{G}}\right) = \sigma_\varepsilon^2\mathbf{I}_n$ yields

$$SMDE\left(\hat{\mathbf{f}}_\lambda\right) = \|(\mathbf{I} - \mathbf{S}_\lambda)\,\mathbf{f}\|^{\,2} + \sigma_\varepsilon^2 tr\left[\mathbf{S}_\lambda{}'\mathbf{S}_\lambda\right]$$

A4. *Derivation of the* $E\left(\varepsilon_{iG}\right) = 0$

In view of Lemma, we may write $E(Z_{iG}|X_i) = f(X_i)$. Note that G is the cumulative distribution function of the censoring variable C_i, as pointed out before. When G is known, the expectation of the error terms

$\varepsilon_{1G}, \ldots, \varepsilon_{nG}$ are obtained by

$$\begin{aligned} E(\varepsilon_{iG}) &= E\left[Z_{iG} - E\left(Z_{iG}|X_i\right)\right] \\ &= E(Z_{iG}) - E\left(E(Z_{iG}|X_i)\right) = 0 \end{aligned}$$

as claimed. Thus, the data set $(Z_{iG},\ X_i)$ can be considered as observations drawn from a nonparametric model with errors ε_{iG}. This argument helps us calculate the estimate of $f(X)$.

When G is unknown, a natural solution is to replace G by its Kaplan-Meier (1958) estimator $\hat{G}$. It is obvious that errors occur caused by this estimation procedure. In this case, we cannot expect that the expected value of the errors will also obtain zero. Heuristically, however, they should converge to zero for a sample size tending to infinity.

Acknowledgments

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

References

Buckley, J., James, I. (1979). Linear regression with censored data. *Biometrika* 66(3):429–436.

Cai, T., Betensky, R. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* 59:570–579.

Burnham, K. P., Anderson, D. R. (2004). Multimodel inference understanding AIC and *BIC* in model selection. *Socialogical Methods and Research* 33(2):261–304.

Claeskens, G., Krivobokova, T., Opsomer, J. D. (2007). Asymptotic properties of penalized spline estimators. Technical Report, Faculty of Business and Economics, Katholieke Universiteit Leuven.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* 31:377–403.

Dabrowska, D. M. (1992). Nonparametric quantile regression with censored data. *Sankhya: The Indian Journal of Statistics Series A* 54(2):252–259.

Eilers, P. H. C., Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Scinece* 11(2):89–102.

Eilers, P. H. C., Marx, B. D. (2010). Splines, knots and penalties. *Computational Statistics* 2(6):637–653.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Macal Dekker.

Fan, J., Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *Journal of the American Statistical Association* 89(426):560–570.

Ghouch, A. E., Keilegom, I. V.(2008). Nonparametric regression with dependent censored data. *Scandinavian Journal of Statistics* 35:228–247.

Hall, P., Opsomer, J. D. (2005). Theory for penalized spline regression. *Biometrika* 92(1):105–118.

Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hardle, W., Hall, P., Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association* 83(401):86–95.

Hurvich, C. M., Simonoff, J. S., Tasi, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of Royal Statistical Society. Series B (Statistical Methodology)* 60(2):271–293.

Kalbfleich, J. D., Prentice, R. L. (1980). Estimation of the average hazard ratio. *Biometrika* 68(1):105–112.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282):457–481.

Ke, C., Wang, Y. (2001). Semiparametric nonlinear mixed effects models and their applications (with discussions). *Journal of the American Statistical Association* 96:1272–1208.

Kim, H. T., Truong, Y. K. (1998). Nonparametric regression estimates with censored data; local linear smoothers and their applications. *Biometrics* 54(4):1434–1444.

Koul, H., Susarla, V., Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics* 9(6):1276–1285.

Lai, T. L., Ying, Z. (1992). Asymptotically efficient estimation in censored and truncated regression models. *Statistica Sinica* 2:17–46.

Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* 74:301–309.

Peterson, A. V. Jr. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association* 72:854–858.

Qin, G., Jing, B.-Y. (2001). Empirical likelihood for censored linear regression. *Scandinavian Journal of Statistics* 28(4):661–673.

Qin, G., Jing, B.-Y. (2000). Censored partial linear models and empirical likelihood. *Journal of Multivariate Analysis* 78(1):37–61.

Reis, P. T., Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of Royal Statistical Society. Series B (Statistical Methodology)* 71:505–523.

Ruppert, D., Wand, M. P., Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* 11:735–757.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6:461–464.

Wahba, G. (1990). *Spline Model for Observational Data*. Philadelphia: SIAM.

Wang, F.-Y. (1996). Sharp explicit lower bounds of heat kernels. *The Annals of Probability* 25(4):1995–2006.

Winter, S. (2013). Smoothing spline regression estimates for randomly right censored data. Dissertation, University of Stuttgart.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(1):3–36.

Yang, L. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61(4):793–815.

Yu, Y., Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 97:1042–1054.

Zheng, Z. K. (1984). Regression analysis with censored data. PhD. Dissertation, Univ. of Colombia.

Zhou, M. (1992). Asymptotic normality of the synthetic data regression estimation for censored survival data. *The Annals of Statistics* 20(2):1002–1021.