Empirical Economics
https://doi.org/10.1007/s00181-020-01944-x

Semiparametric modeling of the right-censored time-series based on different censorship solution techniques

Dursun Aydın[1] · **Ersin Yılmaz**[1]

Received: 28 February 2020 / Accepted: 16 September 2020

Abstract
In this paper, we employ the penalized spline method to estimate the components of a right-censored semiparametric time-series regression model with autoregressive errors. Because of the censoring, the parameters of such a model cannot be directly computed by ordinary statistical methods, and therefore, a transformation is required. In the context of this paper, we propose three different data transformation techniques, called Gaussian imputation (*GI*), *k* nearest neighbors (*k*NN) and Kaplan–Meier weights (KMW). Note that these data transformation methods, which are modified extensions of ordinary *GI*, *k*NN and KMW approximations, are used to adjust the censoring response variable in the setting of a time-series. In this sense, detailed Monte Carlo experiments and a real time-series data example are carried out to indicate the performances of the proposed approaches and to analyze the effects of different censoring levels and sample sizes. The obtained results reveal that the censored semiparametric time-series models based on *k*NN imputation often work better than those estimated by *GI* or KMW.

Keywords Right-censored time-series · Gaussian imputation · *k*NN imputation · Kaplan–Meier weights · Penalized splines · Semiparametric regression

1 Introduction

In econometrics and statistics literature, the term right-censored data is employed for observations that cannot be observed beyond a cutoff value. Generally, time-series measurements are often observed with data irregularities, such as observations due to a detection limit. Namely, some response observations exceeding the detection limit

✉ Ersin Yılmaz
yilmazersin13@hotmail.com

Dursun Aydın
duaydin@hotmail.com

[1] Department of Statistics, Faculty of Science, Mugla Sitki Kocman University, 48000 Mugla, Turkey

Published online: 18 October 2020

will not be known, and these incomplete observations will be recorded as the value of the detection limit. Depending on this issue, the known-classical semiparametric time-series regression analysis cannot be directly applied to the right-censored data. Note that in the case of uncensored response observations, classical time-series regression models with autoregressive errors are analyzed by parametric methods. For instance, see Box and Jenkins (1970), Brockwell and Davis (1991) for more detailed discussions. In the presence of censoring, the estimates obtained from parametric methods are highly biased and unreliable. A way to handle this problem is to replace censored data points with reasonable values from observations of a data set via imputation methods. Note that imputation refers to the process of replacing the censored data with substituted values. Another way to cope with censorship data is to consider the weighted Kaplan–Meier estimator of the observed response variable distribution that can replace the empirical distribution. Note also that Kaplan–Meier gives suitable weights to the censored observations (see, Miller 1976; Stute 1993).

Several authors studied the imputation methods in dealing with censored data. For example, Park et al. (2007) considered the *GI* method to analyze censored time-series with autoregressive moving average models. Batista and Monard (2002) analyzed the use of the *k*NN method as an imputation to solve missing data problem in machine learning algorithms. The *k*NN method computes the imputed value from the mean of measured *k* uncensored values in the data set. Some examples of studies about *k*NN imputation include Malarvizhi and Thanamani (2012) and Chen and Shao (2000). The main idea of the *GI* method, on the other hand, is that the censored values are replaced by estimating observations with the help of the conditional truncated normal distribution. There are some important studies related to *GI* in the literature. See, for example, the studies of Park et al. (2009), Faubel et al. (2009) and Silva and Deutsch (2017). In addition, see Lee et al. (2018) to see a different perspective on imputation technique.

Note that the aforementioned studies are essentially designed for parametric methods. But, in the real-world, time-series we work with often do not have a parametric linear structure and thus they cannot always be handled by parametric methods. Therefore, in practice, many authors suggested the use of nonparametric techniques for analyzing time-series data. See, for example, Hardle et al. (1997), Morton et al. (2009) and Aneiros-Perez et al. (2011).

It should be emphasized that nonparametric estimators, unlike parametric approaches, are very flexible but their statistical accuracy decreases greatly if we add several explanatory variables in the regression model. Such a case is always possible in a regression problem and is known as the curse of dimensionality. To overcome the curse of dimensionality problem, we used, in this paper, a semiparametric regression model that combines the features of parametric and nonparametric models. In such models, the parametric part can be interpreted as a linear model, while the nonparametric part flexes the model from the rigid structural assumptions. Further advantages of these models can be stressed as the inclusion of categorical variables in a parametric way, an easy interpretation of the outcomes and a part specification of a semiparametric regression model. Therefore, in the last two decades, many authors have shown interest in semiparametric regression techniques to model time-series with nonlinearity. Examples of such work include Truong and Stone (1994), Gao (1995), Yu and

Chen (2007), Gao (2007), Kato and Shiohama (2009), Gao and Philips (2010) and Linton et al. (2009).

The main theme of this paper is the use of the semiparametric techniques to fit and make inferences concerning a semiparametric regression model with censored time-series data. The key problem here is that the data are censored from the right, as in many environmental and econometric time-series applications. One common routine in such a case is then to adjust for the censoring effect by transforming the observations of the response variable. Based on this consideration, we propose three different data transformation techniques, which are based on generalization of the ordinary *GI*, *k*NN and KMW methods in case of the uncensored data. These methods, which are modified extensions of ordinary statistical approximations, are employed to determine missing response observations. Note that mentioned data transformation techniques provide useful censoring response observations with the help of efficient algorithms described newly in this article. Hence, the transformed response variable can be treated as uncensored variable and standard semiparametric regression methods can be applied, as in classical regression analysis. After the transformation of data, we apply the semiparametric technique which is partially linear model based on the penalized spline method. See, Aydin and Yilmaz (2018) for more details on the partially linear model using a penalized spline. It should be also noted that we compare the performances of the suggested *GI*, *k*NN imputations and KMW method. Their effects on the semiparametric regression estimates are also measured. To the best of our knowledge, such a study has not yet been discussed.

The rest of this paper is organized as follows. Fundamental ideas on the right-censored time-series and semiparametric model are expressed in Sect. 2. Section 3 involves the solution methods that are Gaussian imputation, *k*NN imputation, and Kaplan–Meier weights. Performance measurements are expressed in Sect. 4. To see methods' behaviors in practice, simulation and real data studies are carried out in Sects. 5 and 6, respectively. Finally, conclusions are given in Sect. 7.

2 Materials and methods

In the classical time-series processes, we assumed that the value of each sample unit is completely observed or known. In many applications, however, all of the units in the sample may not be followed (or observed). These types of data are commonly called censored time-series data. Some techniques in this context are developed. The usual approach is to fill in (impute) the unobserved values in some way. There are also various ways to deal with the censoring data:

Throwing or ignoring a censored observation. Analyzing data using only uncensored ones. Although this method is preferable for its simplicity, the results will be biased if censored observations did not fit the assumption that data points are censoring at random. Also, it causes a loss of information when the censoring level is getting higher. It is a primitive method to handle censored data.

Forcing data to fit into a particular distribution (i.e., Weibull, Normal, Exponential, etc.*).* Here, the probabilities of the observations and the censorship effect are added

to the estimation process. If the distribution of the data is clear, this technique will be beneficial but in general, the distribution of time-series data is unspecified.

Data transformation or using Kaplan–Meier weights. If the data do not follow any distribution, then the synthetic data transformation (Koul et al. 1981) and Kaplan–Meier weights (Miller 1976) based on Kaplan and Meier (1958) estimator can be used to overcome the censorship.

Imputation methods for handling censorship. Commonly used imputations techniques include the mean imputation, Gaussian imputation (Park et al. 2007), *k*NN imputation (Batista and Monard 2002), singular value decomposition (SVD)-based imputation, Hot-deck imputation, regression imputation and so on. In this paper, the *k*NN and Gaussian imputation techniques are considered as representatives of the many important imputation methods. They are also chosen for an important difference between them: Gaussian depends on the normal distribution, but *k*NN is free from all distributions.

One of the major concerns of this study is to detect the behaviors of three censorship solution methods on modeling time-series in the semiparametric setting. In this context, consider the uncensored semiparametric time-series model

$$Y_t = \mathbf{x}_t\boldsymbol{\beta} + g(z_t) + \varepsilon_t, t = 1, .., n \tag{2.1}$$

where Y_t's are the uncensored values of stationary time-series, $\mathbf{x}_t = \left(x_{1t}, \ldots, x_{pt}\right)$ is a $(n \times p)$ dimensional matrix of parametric covariates for time t, $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)'$ is a $(p \times 1)$ vector of regression coefficients, $g(.)$ is an unknown smooth function to be estimated based on values of nonparametric variable z_t's, and finally, ε_t's are the stationary autoregressive error terms, given by

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \tag{2.2}$$

where ρ is an autocorrelation parameter and u_t's are independent and identically distributed random error terms with $u_t \sim N\left(0, \sigma_{u_t}^2\right)$ and $|\rho| < 1$. It should be noted that when $\rho = 0$, this model reduces to an ordinary semiparametric regression model.

According to the concept of this study, Y_t's are censored from the right by a constant detection limit C_t. Therefore, instead of observing the values of Y_t, we now observe the data set defined as

$$S_t = \min(Y_t, C_t), \delta_t = I(Y_t \leq C_t) \tag{2.3}$$

where S_t's are the updated response values, δ_t includes the information on whether an observation is censored or uncensored and $I(.)$ is an indicator function. One thing to point out here is that if an observation is censored, we take $S_t = C_t$ and $\delta_t = 0$; otherwise, we choose $S_t = Y_t$ and $\delta_t = 1$. Thus, we obtain a new data sets and model (2.1) turns into a right-censored semiparametric time-series model

$$S_t = \mathbf{x}_t\boldsymbol{\beta} + g(z_t) + \varepsilon_t, t = 1, .., n, \tag{2.4}$$

As indicated before, the key idea of this paper is to estimate the components of the semiparametric model stated in (2.4) using penalized spline method. In this sense, we modified the GI, kNN methods and KMW for dealing with the censored observations of response variable S_t in a semiparametric regression setting. Also, we want to say that none of these methods is used in a semiparametric regression model setting under the right-censored time-series data. This is the most important innovation of this paper. In the next section, the penalized spline method is first expressed, and then, the imputation methods and KMW are introduced.

2.1 Penalized splines

In this section, the penalized spline method is introduced to estimate the parametric and nonparametric part of a semiparametric model with right-censored time-series data. Note that although some semiparametric approximations could be employed, we prefer to use penalized spline technique. One of the most important reasons is that this technique is highly resistant to censorship, as proved in the study of Aydin and Yilmaz (2018).

The penalized splines method is first adapted to estimate an unknown function in a nonparametric regression model by Eilers and Marx (1996) and then improved to a partially linear (or semiparametric) model by Liang (2006). Penalized spline method provides the estimates by using piecewise polynomial functions with nonzero derivatives at special knot points to be selected. Such polynomial functions (i.e., fixed-knot splines) are also known as regression splines. In general, this method works only for the required knot points, so the method runs faster and is not affected by outliers. This property is very critical when one of the main considerations is to model censored data appropriately.

The key idea in the penalized spline is to estimate the components of model (2.4) so that sum of squares of the differences between the censored response observations S_t and $\left(\mathbf{x}_t\hat{\boldsymbol{\beta}} + \hat{g}(z_t)\right)$ is a minimum. In here, the unknown smooth function $\hat{g}(z_t)$ is approximated by a qth degree regression spline with a truncated power basis

$$g(z_t) = b_0 + b_1 z_{t1} + \cdots + b_q z_{tq}^q + \sum_{k=1}^{K} b_{q+k}(z_t - \kappa_k)_+^q + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{2.5}$$

where $\mathbf{b} = \left(b_0, b_1, \ldots, b_q, b_{q+1}, \ldots, b_{q+K}\right)'$ is a vector of unknown regression coefficients, $q \geq 1$ indicates the degree of regression spline, $(z_t - \kappa_k)_+ = (z_t - \kappa_k)$ when $(z_t - \kappa_k) > 0$ and $(z_t - \kappa_k)_+ = 0$ otherwise. Also, $\kappa_1, \kappa_2 \ldots, \kappa_K$ denote the selected knot points provided $\{\min(z_t) \leq \kappa_1 < \cdots < \kappa_K \leq \max(z_t)\}$.

In the light of the information given above, semiparametric regression model with right-censored time-series data can be written as follows

$$S_t = x_{t1}\beta_1 + \cdots + x_{tp}\beta_p + b_0 + b_1 z_{t1} + \cdots + b_q z_{tq}^q + \sum_{k=1}^{K} b_{q+k}(z_t - \kappa_k)_+^q + \varepsilon_t \tag{2.6}$$

where $(z_t - \kappa_k)_+ = max(0, (z_t - \kappa_k))$. Equation (2.6) in matrix and vector form is rewritten as

$$\mathbf{S} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{U}\mathbf{b} + \boldsymbol{\varepsilon} \tag{2.7}$$

where $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p, b_0, \ldots, b_q\right)'$ denotes the coefficients of the parametric linear component, while $\mathbf{b} = \left(b_{q+1}, \ldots, b_{q+K}\right)'$ denotes the coefficients of the nonparametric component, $\mathbf{X}$ and $\mathbf{U}$ are the design matrices that can be defined by

$$\mathbf{X} = \begin{bmatrix} 1 & x_{t1} & \ldots & x_{tp} & z_t & \ldots & z_t^q \\ \vdots & \vdots & & \vdots & & & \\ 1 & x_{n1} & \ldots & x_{np} & z_n & \ldots & z_n^q \end{bmatrix}, \mathbf{U} = \begin{bmatrix} (z_t - \kappa_1)_+^q & \ldots & (z_t - \kappa_K)_+^q \\ \vdots & \ddots & \vdots \\ (z_n - \kappa_1)_+^q & \ldots & (z_n - \kappa_K)_+^q \end{bmatrix}, t = 1, \ldots, n \tag{2.8}$$

and $\boldsymbol{\varepsilon}_t = (\varepsilon_1, \ldots, \varepsilon_n)'$ is a vector of the stationary autoregressive error terms, as defined in (2.2). Note that we assume that $\boldsymbol{\varepsilon}_t \sim \boldsymbol{N}_n(0, \boldsymbol{A})$, where the covariance matrix $\boldsymbol{A}$ is a symmetric and positive definite matrix and its entries are determined by

$$\mathbf{A} = \frac{\sigma_u^2}{1 - \rho^2}\boldsymbol{R}, R_{i,j} = \rho^{|i-j|}, i, j = 1, 2, \ldots, n. \tag{2.9}$$

For convenience, we assume that $\mathbf{A}$ is known. Then, for any symmetric positive semidefinite matrix $\mathbf{D}$ and scalar $\lambda > 0$, the penalized spline estimators $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{b}_0, \hat{b}_1, \ldots, \hat{b}_q\right)'$ and $\hat{\boldsymbol{g}} = \left(\hat{b}_{q+1}, \ldots, \hat{b}_{q+K}\right)'$ of $\boldsymbol{\beta}$ and $\mathbf{b}$ in (2.7) can be obtained by minimizing the penalized residual sum of squares (*PRSS*) criterion

$$\begin{aligned} PRSS(\boldsymbol{\beta}, \mathbf{b}; \lambda) &= \sum_{t=1}^{n} A_t (S_t - x_t\boldsymbol{\beta} - g(z_t))^2 + \lambda \sum_{k=1}^{K} \mathbf{b}_{p+k}^2 \\ &= (\mathbf{S} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\mathbf{b})' \mathbf{A} (\mathbf{S} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\mathbf{b}) + \lambda \mathbf{b}' \mathbf{D} \mathbf{b} \end{aligned} \tag{2.10}$$

where $\lambda \sum_{k=1}^{K} \mathbf{b}_{p+k}^2$ denotes the penalty term depends on the knot points and λ is a smoothing parameter that controls the amount of the penalty. $\mathbf{D} = \text{diag}(0_{r+1}, 1_K)$ is a diagonal penalty matrix with $(r + 1)$ (where $r = p + q$) diagonal entries of zeros for $\boldsymbol{\beta}$ and K diagonal elements of ones for $\mathbf{b}$, as shown in (2.6) or (2.7).

By simple algebraic operations, it follows that Eq. (2.10) is minimized when $\boldsymbol{\beta}$ and $\mathbf{b}$ satisfy the system of equations

$$\begin{pmatrix} \mathbf{X}'\mathbf{A}\mathbf{X} & \mathbf{X}'\mathbf{A}\mathbf{U} \\ \mathbf{U}'\mathbf{A}\mathbf{X} & (\mathbf{U}'\mathbf{A}\mathbf{U} + \lambda\mathbf{D}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}' \\ \mathbf{U}' \end{pmatrix} \mathbf{S} \tag{2.11}$$

After some algebraic manipulations in the block matrix in (2.11), the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$, respectively, of the parameters $\boldsymbol{\beta}$ and $\mathbf{b}$ can be easily obtained by

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^{\mathrm{T}}\mathbf{A}\left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{T}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{T}\mathbf{A}^{T}\right)\mathbf{X}\right]^{-1}\mathbf{X}^{T}\mathbf{A}\left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{T}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{T}\mathbf{A}^{T}\right)\mathbf{S} \tag{2.12a}$$

$$\hat{\mathbf{b}} = \left(\mathbf{U}^{T}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{T}\mathbf{A}^{T}\left(\mathbf{S} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \tag{2.12b}$$

From (2.12a–b) fitted values can be described as

$$\hat{\boldsymbol{\mu}} = \left(\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{U}\hat{\boldsymbol{b}}\right) = (\mathbf{H}_{\lambda}\mathbf{S}) = \hat{\mathbf{S}} = E[Y|x, z] \tag{2.12c}$$

where $\mathbf{H}_{\lambda}$ is a smoothing matrix, which is also known as hat matrix given by

$$\mathbf{H}_{\lambda} = \mathbf{U}\left(\mathbf{U}^{T}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{T}\mathbf{A}^{T} + \mathbf{C}\mathbf{X}\left(\mathbf{X}^{T}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{C} \tag{2.12d}$$

with $\mathbf{C} = \left(\mathbf{I} - \mathbf{U}(\mathbf{U}^{T}\mathbf{A}\mathbf{U} + \lambda\mathbf{D})^{-1}\mathbf{U}^{T}\mathbf{A}^{T}\right)$.

Derivations of Eqs. (2.12a–d) are given in "Appendix."

In practice, the estimators given in the (2.12a–b) cannot be used directly unless the values of response variable **S** are observed completely. To solve this problem, we propose three data transformations techniques such as *GI*, *k*NN and KMW discussed in the next section.

3 Solution methods for censorship

There are mainly two approaches in the literature to overcome censorship. One is to eliminate censored observations and continue to analyze with uncensored ones, and the other one is to use censored data points as observed. However, Park et al. (2007) show that both methods give highly biased and inefficient estimates. According to Helsel (1990), these two approaches may only be useful for data sets with low censorship rates. Of course, such ideas are not permanent solutions in many applications. This study aims to complete censored time-series data correctly and provide useful methods for time-series analysis in a semiparametric regression setting. Therefore, in the case of censored observations, three different approaches with different advantages and disadvantages are introduced in the next sections.

3.1 Gaussian imputation

Assume that $Y_t = (Y_1, Y_2, \ldots, Y_n)^T$ is a realization from a stationary time-series defined in model (2.1) with correlated errors. Note also that the error terms follow a multivariate normal distribution with mean zero and covariance matrix **A**: $\varepsilon_t \sim N_n(0, \mathbf{A})$, where $\mathbf{A} = \sigma^2\boldsymbol{R}(\rho)$ is an nxn matrix, given by

$$\boldsymbol{A} = \sigma^2 \boldsymbol{R}(\rho) = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho^1 & \dots & \rho^{n-1} \\ \rho^1 & 1 & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{bmatrix} = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{bmatrix}$$

where $\boldsymbol{R}$ is a (nxn) autocorrelations matrix with elements $R_{i,j} = \rho^{|i-j|}$, $i, j = 1, 2, \dots, n$, as defined in (2.9), and $1, \rho^1, \dots, \rho^{n-1}$ are theoretical autocorrelations of the autoregressive process.

From the ideas given above, it is understood that $Y_t \sim N_n(\mu, \boldsymbol{A})$ for complete data. When we consider the response observations with a censoring mechanism, $Y_t \sim TN_n(\mu, \boldsymbol{A}; R_c)$, where $TN_n(.; R_c)$ denotes the truncated normal distribution on the interval R_c (see Vaida and Liu 2009). Note that the interval R_c depends on whether data point is censored. Essentially, the interval R_c is $(0, C_t)$ if $\delta_t = 1$ and R_c is $[C_t, \infty)$ otherwise. To calculate the components in the censored regression model with autoregressive error, the first task is to consider separately the observed and censored data points of the response variable at the beginning of the estimation procedure. In this context, by using permutation matrix P, which maps $(1, .., l)'$ into the permutation vector $\boldsymbol{p} = (p_1, \dots, p_l)'$, the order of the data can be rearranged as

$$PY_t = \begin{bmatrix} P_o \\ P_c \end{bmatrix} Y_t = \begin{bmatrix} \mathbf{Y}_o \\ \mathbf{Y}_c \end{bmatrix} \tag{3.1}$$

where $\mathbf{Y}_o$ represents the vector of observed response values, whereas $\mathbf{Y}_c$ denotes the vector of the unobserved response values.

Using a similar procedure to (3.1), the new observed response variable S_t calculated according to the censoring mechanism in (2.3) can be portioned into the sub-vectors, as follows

$$PS_t = \begin{bmatrix} P_o \\ P_c \end{bmatrix} S_t = \begin{bmatrix} \mathbf{S}_o \\ \mathbf{S}_c \end{bmatrix} \tag{3.2}$$

As stated before, we want to find suitable values, instead of unobserving $\mathbf{S}_c$ given in the (3.2). In this sense, the conditional truncated normal distribution is frequently used in practical implementations (see, Lee and Carlin 2010; Yuan 2009). The key idea is to replace the values of the right-censored vector $\mathbf{S}_c$ by sampling values obtained from the conditional distribution of the censored response vector $\mathbf{Y}_c$ given $\mathbf{S}_o$ and $\mathbf{S}_c$. This procedure is equivalent to applying the truncated normal distribution:

$$(\mathbf{Y}_c | \mathbf{S}_o, \mathbf{S}_c \in R_c) \sim TN_{n_c}(\mathbf{M}, \mathbf{V}, R_c) \tag{3.3}$$

where n_c denotes the number of censored observations, TN_{n_c} shows a truncated multivariate normal distribution with n_c-dimension and R_c determines the region associated with the censoring of the response observations, as defined previously. The symbols

and expressed in (3.3) denote the conditional mean and covariance of uncensored part. Note also that the probability density function of the truncated normal distribution is

$$f(S_t) = g(S_t)I(\delta_t = 0)/[1 - F(C_t)] \tag{3.4}$$

where (.) denotes the indicator function, $g(.)$ and (.) are the probability density function of the standard normal distribution and its cumulative distribution function, respectively. It should be emphasized that $f(.)$ is used for observations in the interval $[, \infty]$ to obtain the distribution of the right-censored part of the data.

To be able to carry out the ideas of the Gaussian imputation method, in the first stage, the parameters of the distributions outlined above must be estimated by iteratively applying an appropriate algorithm defined in Table 1.

From output of the algorithm, we see that $\mathbf{S}^{GI}$ is a vector of response values from kth iteration of the imputation. In this case, we replace the right-censored response vector $\mathbf{S}$ in (2.10) with the vector $\mathbf{S}^{GI}$ imputed by GI method to estimate the regression coefficients. Hence, the estimators given in Eqs. (2.12a–b) are defined, respectively, as

$$\hat{\boldsymbol{\beta}}_{GI} = \left[\mathbf{X}^{\mathrm{T}}\mathbf{A}\left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{A}^{\boldsymbol{T}}\right)\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A}\left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{A}^{\boldsymbol{T}}\right)\mathbf{S}^{GI}, \tag{3.5a}$$

$$\hat{\mathbf{b}}_{GI} = \left(\mathbf{U}^{T}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{T}\mathbf{A}^{T}\left(\mathbf{S}^{GI} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GI}\right) \tag{3.5b}$$

where $\hat{\boldsymbol{\beta}}_{GI}$ and $\hat{\mathbf{b}}_{GI}$ represent the estimators based on Gaussian imputation for parametric and nonparametric parts of model (2.1), respectively. The fitted values are also given as follows

$$\hat{\mathbf{S}}^{GI} = \left(\mathbf{X}\hat{\boldsymbol{\beta}}_{GI} + \mathbf{U}\hat{\mathbf{b}}_{GI}\right) = \left(\mathbf{H}_{GI}\mathbf{S}^{GI}\right) = \hat{\mathbf{Y}}_{GI} = E[Y|x, z] \tag{3.5c}$$

where $\mathbf{H}_{GI} = \mathbf{X}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C} + \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\boldsymbol{T}}\mathbf{A}^{\boldsymbol{T}}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C}\right)$ shows the smoothing matrix with $\mathbf{C}$ described in (2.12d) for a parameter λ obtained with the help of observations vector $\mathbf{S}^{GI}$.

3.2 *k*NN imputation

*k*NN imputation is a common method to overcome the missing data in the literature but in this part of the study, it is modified for imputation of the censored observations. The main purpose of using *k*NN is that censored data points can be imputed and replaced by using *k*NN method. Note that a censored value is imputed by either a value measured as the average of measured values for multiple (k) neighbors. Some of the advantages of this technique can be ordered as follows

i. The method is free from distribution assumptions which provide an important superiority in the analysis of right-censored data that does not fit any distribution.

Table 1 Algorithm for the Gaussian imputation method

Input: Right-censored data set S

Censoring indicator δ_i associated with S

1:Generate the truncated normal distribution and compute the density of the censored partition of the data from equation (3.4)

2: Obtain the initial parameter estimates $\hat{\mu}^0, \hat{\rho}^0$ *and* $\hat{\sigma}^0$ as given below

$$\hat{\mu}^0 = n^{-1}\sum_{i=1}^{n} S_t^0$$

$$\hat{\rho}^0 = \left[\sum_{t=2}^{n}(S_{t-1}^0 - S_{-n}^0)^2\right]^{-1}\left[\sum_{t=2}^{n}(S_t^0 - \bar{S}_{-1}^0)(S_{t-1}^0 - \bar{S}_{-n}^0)\right]$$

$$(\sigma^0)^2 = (n-3)^{-1}\sum_{t=2}^{n}\left(S_t^0 - \hat{\mu}^0 - \rho^0(Z_{t-1}^0 - \hat{\mu}^0)\right)$$

where $\hat{\mu}^0$ denotes the estimated mean for zero iteration, $\bar{S}_{-n}^0 = (n-1)^{-1}\sum_{t=1}^{n-1} S_t$ and similarly $\bar{S}_{-1}^0 = (n-1)^{-1}\sum_{t=2}^{n} S_t$.

3: Compute the conditional mean and variance $\widehat{\mathbf{M}}^0$ and $\widehat{\mathbf{V}}^0$ based on the censored part of the data and create $\Phi = (\mathbf{M}, \mathbf{V})$ as follows

$$\widehat{\mathbf{M}}^0 = \hat{\mu}_c^0 + \hat{\Sigma}_{co}^0\left(\hat{\Sigma}_{oo}^0\right)^{-1}(S_o - \hat{\mu}^0),$$

$$\widehat{\mathbf{V}}^0 = \Sigma_{cc}^0 - \Sigma_{co}^0(\Sigma_{oo}^0)^{-1}\Sigma_{oc}^0$$

4: Generate the vector $\mathbf{S}_c^1$ for the right-censored data points from truncated normal distribution, which is denoted by $TN\left(\widehat{\mathbf{M}}^0, \widehat{\mathbf{V}}^0, R_c\right)$.

5: Construct the augmented data from the observed part and estimated vector $\mathbf{S}_c^1$ defined in previous step:

$$\mathbf{S}^1 = \mathbf{P}^{-1}\begin{bmatrix}\mathbf{S}_o \\ \mathbf{S}_c^1\end{bmatrix}$$

6: Re-compute the estimates obtained in Step 1 according to $\mathbf{S}^1$ and update the mean and variance, which is shown in Step 3.

7: Repeat the Steps 2-6 while $\psi > 0.005$ which is a threshold for this study and it is calculated as follows

$$\psi = \frac{\left[\left(\hat{\theta}^{k+1} - \hat{\theta}^k\right)^T\left(\hat{\theta}^{k+1} - \hat{\theta}^k\right)\right]}{\left(\hat{\theta}^k\right)^T\hat{\theta}^k}$$

where $\theta = (\mu, \sigma, \rho)^T$ and k denotes the kth iteration.

8. Obtain imputed dataset $S^{GI(k)}$ calculated for k^{th} iteration.

Output: Imputed dataset $\boldsymbol{S}^{GI}$

ii. kNN method replaces censored observations with their actual estimates, not synthetic values and also it does not manipulate all data points different from Kaplan–Meier weights.

iii. Separate from synthetic data transformation and K–M weights, the kNN method can use predictor variables to obtain additional information for completing censored data points. That is a very beneficial property, especially in the time-series analysis because it takes into account the effect of time in the imputation process.

Table 2 Algorithm for *k*NN imputation method

Input: Right − censored data set S
Censoring indicator δ_i associated with S
Number of nearest neigbours k
Values of predictor variable z_i related with S
Output: Imputed dataset $\mathbf{Y}^{knn}$
begin
for $(i = 1$ to $n)$ **do**
  if$(\delta_i = 0)$ **do** (if data point is censored)
    for$(j = 1$ to $n)$ **do**
      Find the distances between z_j *and* $z_i$for each censored data point with (3.6)
  Sort the distances from small to large
  for $(j = 1$ to $k)$ **do**
    Take the first *uncensored* $k-$ values of s_i associated to sorted distances
    Calculate the ith imputed value $\left(y_i^{kNN}\right)$ with average of nearest $k-$ records of s_i
Replace the imputed values $\left(y_i^{kNN}\right)$ with censored data points $(r_i, \delta_i = 0)$ in censored data set $\mathbf{r} = (r_1, \ldots, r_n)$
Return $\mathbf{Y}^{kNN} = (y_1^{kNN}, \ldots, y_n^{kNN})^T$
end

iv. It should be indicated that *k*NN imputation is a fully nonparametric method and it does not require any restrictions about the relationship between observation pairs (x_t, z_t, Y_t) or (x_t, z_t, S_t), $t = 1, \ldots, n$.

*k*NN method uses the average value of k closest neighbors for continuous attributes. In this study, Euclidean norm which is a very common distance measurement is used to evaluate the similarity between the corresponding data point and neighbors. Euclidean distance can be calculated by using Minkowski distance when $p = 2$ which is expressed in Eq. (3.6).

$$d_M(X, Y) = \left(\sum_{i=1}^{n} |x_i - Y_i|^p\right)^{\frac{1}{p}} \tag{3.6}$$

In this paper, an algorithm is developed for *k*NN imputation for simplifying the calculations and making procedure more understandable which is given in Table 2.

when $\mathbf{S}$ given in Eqs. (2.12a–b) is by $\mathbf{Y}^{kNN}$ defined in Table 2, we can obtain the penalized spline estimators $\hat{\boldsymbol{\beta}}_{kNN}$ and $\hat{\mathbf{g}}_{kNN}$, based on *k*NN imputation method, respectively, as

$$\hat{\boldsymbol{\beta}}_{kNN} = \left[\mathbf{X}^{\mathrm{T}}\mathbf{A}\left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\mathbf{T}}\mathbf{A}^{\boldsymbol{T}}\right)^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\mathbf{T}}\mathbf{A}^{\boldsymbol{T}}\right)^{-1}\mathbf{Y}^{kNN}, \tag{3.7a}$$

$$\hat{\mathbf{b}}_{kNN} = \left(\mathbf{U}^T\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}^T\left(\mathbf{Y}^{kNN} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{kNN}\right) \tag{3.7b}$$

and the fitted values are

$$\hat{\mathbf{Y}}^{kNN} = \left(\mathbf{X}\hat{\boldsymbol{\beta}}_{kNN} + \mathbf{U}\hat{\mathbf{b}}_{kNN}\right) = \left(\mathbf{H}_{kNN}\mathbf{Y}^{kNN}\right) = E[Y|X, Z] \tag{3.7c}$$

where $\mathbf{H}_{kNN} = \mathbf{X}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C} + \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\boldsymbol{T}}\mathbf{A}^{\boldsymbol{T}}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{A}\mathbf{C}\right)$ is a smoother matrix with $\mathbf{C}$ given in (2.12d) for a smoothing parameter λ described by means of observations vector $\mathbf{Y}^{kNN}$.

3.3 Kaplan–Meier weights

In this section, we begin by adapting the penalized spline based on censored response observations. To handle censored observations, we use Kaplan–Meier (K–M) weights discussed in the study of Stute (1993). In the context of penalized spline, the squared term in the penalized least criterion (2.10) is multiplied by a weight matrix $\boldsymbol{W}$. Then, the penalized least squares (2.10) transform to

$$PRSS_{KMW}(\boldsymbol{\beta}; \mathbf{b}) = (\mathbf{S} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\mathbf{b})^T\mathbf{A}\mathbf{W}(\mathbf{S} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\mathbf{b}) + \lambda\mathbf{b}^T\mathbf{D}\mathbf{b} \tag{3.8}$$

where $\mathbf{A}$ is a covariance matrix, as defined in Sect. 3.1 and $\mathbf{W}$ is a $n \times n$ diagonal matrix that denotes the K–M weights associated with $\left\{S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(n)}\right\}$. The diagonal elements of this matrix are computed by

$$w_{(i)} = \hat{F}_{KM}\left(S_{(i)}\right) - \hat{F}_{KM}\left(S_{(i-1)}\right) = \frac{\delta_{(i)}}{n - i + 1}\prod_{j=1}^{i-1}\left(\frac{n - j}{n - j + 1}\right)^{\delta_{(i)}} \tag{3.9}$$

where $\delta_{(i)}$ denotes the value of censoring indicator associated with ordered values $S_{(i)}$'s. It should be emphasized that the K–M weights defined in (3.9) can also be computed as the contribution of the K–M estimator $\hat{F}$ of the distribution function F of response observations Y_i's at each ordered value $S_{(i)}$.

Performing a little bit of algebra reveals that the solutions for $\boldsymbol{\beta}$ and $\mathbf{b}$ in (3.8) can be defined, respectively, as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{KM} &= \left[\mathbf{X}^T\mathbf{A}\mathbf{W}\boldsymbol{C}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{A}\mathbf{W}\boldsymbol{C}^{-1}\mathbf{S} \text{ with } \mathbf{C} \\ &= \left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}^{\boldsymbol{T}}\mathbf{W}^{\boldsymbol{T}}\right)^{-1}\end{aligned} \tag{3.10a}$$

$$\hat{\mathbf{b}}_{KM} = \left(\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}^T\mathbf{W}^T\left(\mathbf{S} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{KM}\right) \tag{3.10b}$$

and the fitted values are

$$\hat{\mathbf{Y}}^{KM} = \left(\mathbf{X}\hat{\boldsymbol{\beta}}_{KM} + \mathbf{U}\hat{\mathbf{b}}_{KM}\right) = \left(\mathbf{H}_{KM}\mathbf{Y}^{KM}\right) = E[Y|x,z] \tag{3.10c}$$

where $\mathbf{H}_{KM} = \mathbf{X}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{AWCX}\right)^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{AWC} + \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{AWU} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\boldsymbol{T}}\mathbf{A}^{\boldsymbol{T}}\mathbf{W}^{T}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^{\boldsymbol{T}}\mathbf{AWCX}\right)^{-1}\mathbf{X}^{\boldsymbol{T}}\mathbf{AWC}\right)$ denotes the smoother matrix for a parameter λ found by using observations vector $\mathbf{Y}^{KM}$

Derivations of Eqs. (3.10a–b) are given in "Appendix."

It is important to emphasize that smoothing parameter λ discussed in the above formulas has a crucial role in the estimation process. In order to obtain accurate estimates of $\boldsymbol{\beta}$ and $\mathbf{b}$ for three methods, one needs to select an optimum value of parameter λ. From the study of Aydin and Yilmaz (2018), it follows that the improved version of the Akaike information criterion (AIC_c) has a good performance on selection of a smoothing parameter. Calculation of the AIC_c score defined as (Hurvich et al. 1998)

$$AIC_c(\lambda) = 1 + \log\left[(\|\mathbf{H}_\lambda - \mathbf{I})\mathbf{S}\|^2/n\right] + [\{2tr(\mathbf{H}_\lambda) + 1\}/n - tr(\mathbf{H}_\lambda) - 2] \tag{3.7}$$

where $\mathbf{H}_\lambda$ is a smoother matrix depends on a parameter λ. Note also that the matrix $\mathbf{H}_\lambda$ is replaced by the $\mathbf{H}_{GI}$ defined in (3.5c) to select a parameter λ for the estimators based on Gaussian imputation. Similar procedures are performed for the other methods. Hence, a value of λ that minimizes the AIC_c expressed in (3.7) is chosen as an optimum smoothing parameter for each method.

As already noted, the amount of penalty in Eq. (2.10) depends on the set of knots and a smoothing parameter λ. The idea is to choose enough knots and an optimum smoothing parameter to resolve the essential structure in the underlying semiparametric regression model with censored time-series data. In this sense, we see in study of Aydin and Yilmaz (2018) that using improved Akaike information criterion (AIC_c) to choose the parameter λ and the full search algorithm (*FSA*) to select a set of knot points is generally an effective strategy. It should be emphasized the *FSA* searches the whole sequence of trial values and employs the one that minimizes the criterion AICc. See Ruppert et al. (2003) for more detailed discussions about the *FSA*.

4 Assessing the quality of estimators

We now consider several measures for evaluating the quality of estimators, which are obtained in a semiparametric regression setting. Some of these measures denote the quality of estimators with small samples, while other measures represent the quality of estimators with large samples. Note that the quality of an estimator relates to its estimation capability (or its performance) on data. Evaluation of such a performance is extremely important in application areas, since it guides the selection of a model, and provides us a measure of the quality of the ultimately selected model.

To evaluate estimates of the semiparametric time-series model based on censored data, one needs to consider the abilities of methods in terms of parametric component

($\hat{\boldsymbol{\beta}}$), nonparametric component ($\hat{\mathbf{g}}$) and the fitted values ($\hat{\mathbf{S}}$). These different parts of the semiparametric model (2.1) are inspected separately in the next sections.

4.1 Assessment of parametric part

We use the terms *bias* and *variance* to determine the performance of the semiparametric model based on censored time-series data. Note that one can easily decompose the errors of the semiparametric model into two parts such as bias and variance. Such a decomposition helps us understand considering estimators, as these concepts are related to overfitting and under-fitting.

To see the computations of each estimator, we first expand the parametric coefficients estimator $\hat{\boldsymbol{\beta}}_{GI}$ in (3.5a) with the matrix and vector form of (2.4) being replaced by $\mathbf{S}^{GI}$ to find

$$\hat{\boldsymbol{\beta}}_{GI} = \left[\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{S} = \boldsymbol{\beta} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\boldsymbol{g} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\text{"} \tag{4.1}$$

where $\mathbf{C} = \left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{U} + \boldsymbol{\lambda}\mathbf{D}\right)^{-1}\mathbf{U}^{T}\boldsymbol{A}^{T}\right)$, as defined in Eq. (2.12d).

Hence, the bias and variance–covariance matrix of this estimator are obtained, respectively, as follows

$$Bias\left(\hat{\boldsymbol{\beta}}_{GI}\right) = E\left(\hat{\boldsymbol{\beta}}_{GI}\right) - \boldsymbol{\beta} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\boldsymbol{g} \tag{4.2a}$$

$$Var\left(\hat{\boldsymbol{\beta}}_{GI}\right) = \sigma^2\left(\mathbf{X}^{T}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{A}\mathbf{C}\mathbf{X}\left(\mathbf{X}^{T}\mathbf{A}\mathbf{C}\mathbf{X}\right)^{-1} \tag{4.2b}$$

Similarly, we expand $\hat{\boldsymbol{\beta}}_{kNN}$ in (3.7a) with (2.4), which is replaced by $\mathbf{Y}^{\boldsymbol{kNN}}$, to define $Bias\left(\hat{\boldsymbol{\beta}}_{kNN}\right) = E\left(\hat{\boldsymbol{\beta}}_{kNN}\right) - \boldsymbol{\beta}$ and $Var\left(\hat{\boldsymbol{\beta}}_{kNN}\right)$. Note also that since the bias and variance matrix from the kNN method have the same form as those in (4.2a–b), they are not given here.

Finally, as in the above statements, expanded form of the $\hat{\boldsymbol{\beta}}_{KM}$ in (3.10a) can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{KMW}} &= \left[\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{C}\mathbf{W}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{W}\mathbf{S} \\ &= \boldsymbol{\beta} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{W}\boldsymbol{g} + \left(\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{W}\text{"}\end{aligned} \tag{4.3}$$

where $\mathbf{C} = \left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\boldsymbol{T}}\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^{\mathbf{T}}\mathbf{A}^{\boldsymbol{T}}\mathbf{W}^{\boldsymbol{T}}\right)^{-1}$ as describe in Eq. (3.10a).

Thus, the bias and variance–covariance matrix of estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{KMW}}$ are obtained, respectively, as

$$Bias\left(\hat{\boldsymbol{\beta}}_{KMW}\right) = E\left(\hat{\boldsymbol{\beta}}_{KM}\right) - \boldsymbol{\beta} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{C}\mathbf{W}\boldsymbol{g} \tag{4.3a}$$

$$Var\left(\hat{\boldsymbol{\beta}}_{KMW}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{ACWX}\right)^{-1}\mathbf{X}^T\mathbf{ACWX}\left(\mathbf{X}^T\mathbf{ACWX}\right)^{-1} \tag{4.3b}$$

From Eqs. (4.2b) and (4.3b), we can see that the variance matrices are not practical since they depend on the unknown σ^2. In this context, an estimate of σ^2 is required to obtain the aforementioned variance–covariance matrices. In this sense, the natural option is to consider the squared differences between observed responses and its fitted values.

Noting that these squared differences are also known as squared residuals from the semiparametric regression model and the vector form of squared residuals can be written as follows

$$RSS = \mathbf{e}^T\mathbf{e} = \left(\mathbf{Y} - \hat{\mathbf{Y}}\right)^T\left(\mathbf{Y} - \hat{\mathbf{Y}}\right) = [(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{Y}]^T[(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{Y}] = \|(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{Y}\|^2 \tag{4.4}$$

Using (4.4), typically one estimates the variance σ^2 by

$$\hat{\sigma}^2 = \frac{RSS}{tr(I - \boldsymbol{H}_\lambda)^2} = \frac{\|((\mathbf{I} - \mathbf{H}_\lambda)\mathbf{Y})\|^2}{tr\left[(\mathbf{I} - \mathbf{H}_\lambda)^T(\mathbf{I} - \mathbf{H}_\lambda)\right]} \tag{4.5}$$

where $tr(.)$ denotes the trace of a matrix, $tr(\mathbf{I} - \mathbf{H}_\lambda) = n - 2tr(\mathbf{H}_\lambda) + tr\left(\mathbf{H}_\lambda^T\mathbf{H}_\lambda\right)$ is a degrees of freedom depends on smoothing parameter λ. Note that $tr(\mathbf{H}_\lambda)$ need $O(n)$ algebraic operations. It should be noted that the $\mathbf{H}_\lambda$ given in (4.5) is replaced by $\mathbf{H}_{GI}$ in (3.5c), and hence, $\hat{\sigma}^2$ is defined for *GI* method. In a similar fashion, when the smoother matrix $\mathbf{H}_\lambda$ expressed in (4.5) is replaced by $\mathbf{H}_{kNN}$ in (3.7c) and $\mathbf{H}_{KM}$ in (3.10c), the estimates of variance are obtained for the *k*NN and KMW methods.

Note also that $\hat{\sigma}^2$ in (4.5) has an asymptotically negligible bias. If data have a normal distribution, Gaussian imputation finds every censored data point accurately (see Park et al. 2007). However, the same idea cannot be said for *k*NN imputation, due to a machine learning method. Of course, *k*NN imputation has the advantage of being a fully nonparametric method between the other two solution techniques. Hence, it is highly useful for chaotic, unstable and time-series data.

4.2 Assessment of nonparametric part

As denoted in Sect. 2.1, the penalized spline estimate $\hat{\boldsymbol{g}} = \left(\hat{b}_{q+1}, \ldots, \hat{b}_{q+K}\right)^T$ of $\mathbf{b}$ in (2.7) is the corresponding estimation of the nonparametric component $g(z_t)$ in the model (2.1). Viewed from this perspective, we compare the performances of proposed data transformations techniques for evaluating the model in terms of nonparametric parts.

First, we evaluate the performances of the proposed methods by average squared errors, which is also known as mean square error (MSE), given by

$$MSE(\boldsymbol{g}, \hat{\boldsymbol{g}}) = \frac{1}{n}\sum_{j=1}^{n}\left[g(z_j) - \hat{g}(z_j)\right]^2 = n^{-1}(\boldsymbol{g} - \hat{\boldsymbol{g}})^T(\boldsymbol{g} - \hat{\boldsymbol{g}}) \tag{4.6}$$

where $\hat{g}(z_j)$ denotes the value estimated at the jth time point by one of the three methods considered here, such as, GI, kNN and KMW.

Then we assess the relative efficiency of an estimator $\hat{\boldsymbol{g}}_{M1}$ compared to another estimator $\hat{\boldsymbol{g}}_{M2}$. The aforementioned efficiency can be defined as the ratio of MSE ($RoMSE$) values, given by

$$RoMSE(\hat{\boldsymbol{g}}_{M1}, \hat{\boldsymbol{g}}_{M2}) = MSE(\boldsymbol{g}, \hat{\boldsymbol{g}}_{M1})/MSE(\boldsymbol{g}, \hat{\boldsymbol{g}}_{M2}) \tag{4.7}$$

If $RoMSE(\hat{\boldsymbol{g}}_{M1}, \hat{\boldsymbol{g}}_{M2}) > 1$, then it can be said that $\hat{\boldsymbol{g}}_{M2}$ is more efficient than $\hat{\boldsymbol{g}}_{M1}$ and vice versa. The results obtained from (4.6–4.7) are shown in both simulation and real data studies.

4.3 Overall performance of model

In this section, to evaluate the fitted values from the semiparametric regression model with censored time-series data using three techniques we first use the performance measures such as mean absolute relative error ($MARE$), generalized mean square error ($GMSE$) defined by Li and Liang (2008) and mean absolute percentage error ($MAPE$). Then, we assess the relative efficiencies of the methods by the ratio of generalized mean square error ($RGMSE$). These measures are formulated in the following way.

$$MARE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|/|Y_t|,\ GMSE = \left(\hat{\mathbf{Y}} - \mathbf{Y}\right)^T E\left(\mathbf{Y}\mathbf{Y}^T\right)\left(\hat{\mathbf{Y}} - \mathbf{Y}\right),$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|/Y_t,\ \text{and}\ RGMSE = GMSE\left(\hat{\mathbf{Y}}_{M1}\right)/GMSE\left(\hat{\mathbf{Y}}_{M2}\right)$$

Note also that similar to that used for $RoMSE$, it can be said that the fitted values ($\hat{\mathbf{Y}}_{M2}$) obtained from an estimator are more efficient than fitted values ($\hat{\mathbf{Y}}_{M1}$) defined by another estimator, when $RGMSE\left(\hat{\mathbf{Y}}_{M1}, \hat{\mathbf{Y}}_{M2}\right) > 1$.

5 Simulation design and results

In this section, a Monte Carlo simulation study is performed to compare the estimation performances of the modified data transformation techniques such as GI, kNN and KMW, defined in Sect. 3. In this context, simulated data sets are generated from the following model

$$Y_t = x_{1t}\beta_1 + x_{2t}\beta_2 + g(z_t) + \varepsilon_t, t = 1, 2, \ldots, n \tag{5.1}$$

as defined in (2.1).

In Eq. (5.1), $\boldsymbol{\beta} = (\beta_1, \beta_2) = (-1, 0.5)^{'}$, $x_{1t}$and x_{2t} are constructed by the uniform distribution $U[0, 1]$; the regression function $g(z_t) = 8z_t \sin(z_t)^2$ and $z_t =$

$(t - 0.5)/n$; the error terms ε_t are generated using a first-order autoregressive process (that is, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$) with $\rho = 0.5$ and $u_t \sim NIID\big(0, \sigma_u^2 = 1\big)$.

To introduce right censoring, we generate the censoring indicator δ from the Bernoulli distribution with specific censoring levels (C.L.) at 2%, 20% and 40%. Using these C.L. ($\omega = 2\%, 20\%, 40\%$), a cutoff value c is determined by (Park et al. 2007)

$$c = \mu_Y + \sigma \frac{F^{-1}(1-\omega)}{\sqrt{1-\rho^2}} \sqrt{1-\rho^{2(n+1)}}$$

where ω is the censoring probability stated as $\omega = P(Y_t > c)$, and μ_Y is the mean of response variable Y_t, $F(.)$ represents the standard normal distribution function, ρ is the autocorrelation parameter, as defined in (2.2), and $\sqrt{1-\rho^{2(n+1)}}$ is the correction term for the finite sample sizes.

After deciding the cutoff value c, censored time-series C_t can be produced as

$$C_t = Y_t(1 - I(Y_t > c)) + c.I(Y_t > c), \quad t = 1, \ldots, n$$

Thus, the new incompletely observed response measurements S_t are constructed by Eq. (2.3). However, because of the censoring, ordinary methods cannot be applied to these measurements directly. To overcome this problem, we use the observed response variables obtained by three data transformation techniques, denoted as *GI*, *k*NN and KMW, given in Sect. 3. Note also that for each simulation configuration, we generate 1000 random samples of size $n = 50$, 200 and 300 based on censoring levels.

Figure 1 shows the uncensored observations generated from model (5.1) together with right-censored values for a single simulated data set based on various sample sizes and censoring levels. Note also that in this simulation experiment, different configurations are established to provide perspective of the adequacy of the data transformations techniques stated in main text. Because there are many different simulation configurations, it is not possible to present all of them. Therefore, the results from the simulation study are summarized in the following tables and figures. But the codes of simulation experiments will be provided in https://github.com/yilmazersin13.

5.1 Outcomes from the parametric component

When tables are inspected roughly, some expected outputs can be seen such as estimates getting worse versus increasing censoring level and better results for larger samples. It should be noted that these common inferences are not valid for *k*NN imputation which is a machine learning method. Although in most of the cases *k*NN seems ensured the mentioned expected results, it is not an obligation for it. It is already shown in Table 3; *k*NN-based estimates for 20% and 40% censoring levels are better than 2%. It is also counted as an advantage of *k*NN because it may be useful for any censoring level. It cannot be generalized for the *GI* and *KMW* methods because of their theoretical properties.

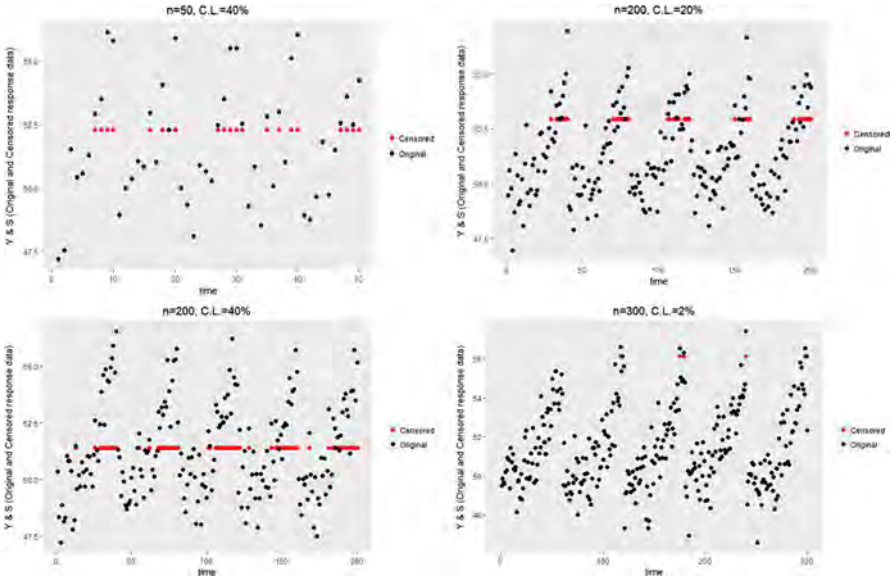

Fig. 1 Scatterplot of the uncensored and right-censored observations versus time for different sample sizes and censoring levels: Red points denote the censored observations, while black points show the uncensored observations. (Color figure online)

Table 3 Outcomes from parametric components of the model (5.1) with right-censored data for $n = 50$

C.L.	Method	$\left[\hat{\beta}_1,\ \hat{\beta}_2\right]$	$\left[B(\hat{\beta}_1),\ B(\hat{\beta}_2)\right]$	$\left[Var(\hat{\beta}_1),\ Var(\hat{\beta}_2)\right]$
2%	GI	[− **0.97**; 0.52]	[**0.021**; 0.008]	[0.384; **0.297**]
	*k*NN	[− 0.88; 0.42]	[0.027; **0.001**]	[0.404; 0.331]
	KMW	[− 0.85; **0.50**]	[0.027; 0.085]	[**0.382**; 0.344]
20%	GI	[− 0.90; 0.58]	[0.090; 0.081]	[0.447; 0.448]
	*k*NN	[− **1.00**; **0.55**]	[**0.009**; **0.058**]	[0.466; 0.478]
	KMW	[− 0.41; 0.60]	[0.580; 0.108]	[**0.425**; **0.422**]
40%	GI	[− 0.58; 0.23]	[0.416; 0.269]	[0.494; 0.470]
	*k*NN	[− **0.93**; **0.41**]	[**0.060**; **0.086**]	[0.446; 0.433]
	KMW	[− 0.39; 0.06]	[0.601; 0.431]	[0.397; 0.384]

In Tables 3, 4 and 5, best scores for each estimation are marked with bold color. Details of tables show that the estimates based on *k*NN imputation are better than the other two methods in terms of regression coefficients $\left(\hat{\beta}_1,\ \hat{\beta}_2\right)$ and their biases $\left\{B\left(\hat{\beta}_1\right),\ B\left(\hat{\beta}_2\right)\right\}$. In the case of variance, the estimates based on *GI* appear more satisfying than others.

To see the performance of the imputation methods for estimating the parametric component of the model, the box plots of the estimated regression coefficients in 1000 replications are presented in Fig. 2. For different combinations, the biases of

Table 4 Similar to Tables 3 but for $n = 200$

1. C.L.	Method	$\left[\hat{\beta}_1,\ \hat{\beta}_2\right]$	$\left[B(\hat{\beta}_1),\ B(\hat{\beta}_2)\right]$	$\left[Var(\hat{\beta}_1),\ Var(\hat{\beta}_2)\right]$
2%	GI	[− 1.04; 0.52]	[0.040; 0.021]	[0.036; 0.030]
	kNN	[− **1.02**; **0.51**]	[**0.026**; **0.015**]	[0.035; 0.029]
	KMW	[− 0.96; 048]	[0.032; **0.015**]	[**0.007**; **0.008**]
20%	GI	[− **0.90**; **0.46**]	[**0.091**; **0.037**]	[**0.033**; **0.036**]
	kNN	[− 1.15; 0.58]	[0.156; 0.085]	[0.044; 0.044]
	KMW	[− 0.72; 0.36]	[0.274; 0.133]	[0.119; 0.133]
40%	GI	[− 0.71; 0.36]	[0.284; 0.131]	[0.049; 0.073]
	kNN	[− **1.25**; **0.62**]	[**0.252**; **0.125**]	[**0.048**; 0.073]
	KMW	[− 0.54; 0.26]	[0.459; 0.230]	[0.082; **0.067**]

Table 5 Similar to Tables 3 and 4 but for $n = 300$

C.L.	Method	$\left[\hat{\beta}_1,\ \hat{\beta}_2\right]$	$\left[B(\hat{\beta}_1),\ B(\hat{\beta}_2)\right]$	$\left[Var(\hat{\beta}_1),\ Var(\hat{\beta}_2)\right]$
2%	GI	[− 1.03; **0.51**]	[0.030; 0.018]	[0.020; 0.014]
	kNN	[− **1.01**; **0.51**]	[**0.019**; **0.013**]	[0.021; **0.001**]
	KMW	[− 0.97; 0.48]	[0.029; 0.021]	[**0.005**; 0.005]
20%	GI	[− **0.93**; **0.47**]	[**0.064**; **0.027**]	[0.013; **0.004**]
	kNN	[− 1.16; 0.58]	[0.169; 0.088]	[0.034; 0.012]
	KMW	[− 0.74; 0.37]	[0.252; 0.120]	[**0.009**; 0.009]
40%	GI	[− 0.84; 0.46]	[0.157; 0.039]	[**0.019**; **0.018**]
	kNN	[− **1.03**; **0.51**]	[**0.038**; **0.018**]	[0.030; 0.042]
	KMW	[− 0.56; 0.36]	[0.435; 0.138]	[0.027; 0.029]

the predictions are also indicated by line graphs in the panel (d) of the same figure according to censoring levels and sample sizes, respectively.

All the graphs plotted for the parametric components of the model, also shown in Fig. 2, confirm all simulation results given in Tables 3, 4 and 5. The thick red lines in panels (a), (b) and (c) of Fig. 2 show the real values of the regression coefficients. In this context, when the panels are examined in detail, it can be clearly seen that as the censoring level increases, the box graphs start to deviate from the red line and more outliers appear. However, in order to better understand the success of the methods of estimating the parametric component, it is also possible to see some interesting results in the line graphs showing the biases given in panel (d). Due to their theoretical properties, it can be seen that GI and KMW methods give better results with increasing sample sizes and worse results with increasing censoring levels. Moreover, it can be said that the *k*NN method is not affected by censoring levels and sample sizes. For example, when $n = 300$, the measured bias values for the 40% censoring level are almost identical to the results obtained at the 2% censoring level.

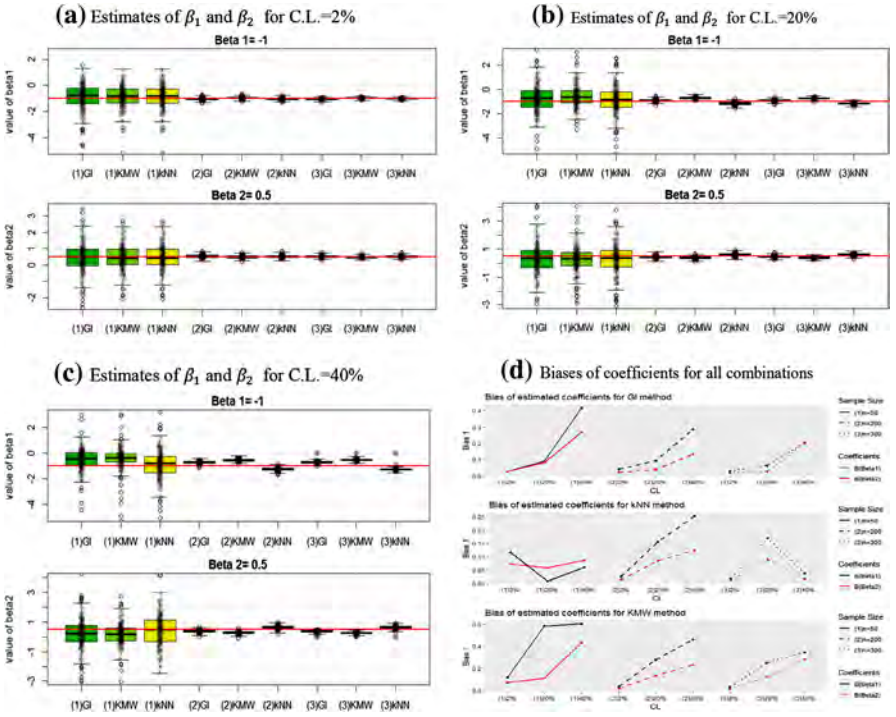

Fig. 2 **a**–**c** Represent the boxplots of estimated regression coefficients and **d** is formed to see biases of the estimations for all simulation combinations and all methods

As a result, when the y-axes in panel (d) of Fig. 2 are examined, it can be said that the *k*NN method generally has lower bias values than the other two imputation methods, but the *GI* and KMW methods give more stable results than the kNN method. The idea here is that *GI* and KMW can give better results at low censoring levels.

5.2 Outcomes from the nonparametric component

Table 6 presents the results from the estimation of nonparametric components of the model (5.1) based on each imputation methods { that is, $\hat{\boldsymbol{g}}_{GI}$, $\hat{\boldsymbol{g}}_{kNN}$, and$\hat{\boldsymbol{g}}_{KMW}$ }. As mentioned earlier, MSE and $RoMSE$ criteria are used to evaluate the performance of the nonparametric part. Note that the $RoMSE$ scores are given in Fig. 3 to facilitate understanding of the results. Furthermore, the curves fitted by each method are shown in Fig. 4.

The MSE scores in Table 6 show that *k*NN and *GI* methods are generally satisfactory. Compared to other two methods, the KMW method performs the worst in most cases and particularly when the censoring levels increase. However, when the results in Table 6 are examined in detail, the KMW gives a very good second score for low censoring level (i.e., C.L. = 2%). The *GI* method can be quite unstable in some cases. The *k*NN method is the most stable better than the other existing methods. Moreover,

Table 6 MSE values from the estimates based on imputation techniques

C.L.	$n = 50$			$n = 200$			$n = 300$		
	GI	*k*NN	KMW	GI	*k*NN	KMW	GI	*k*NN	KMW
2%	0.0377	**0.0148**	0.0188	0.0122	**0.0068**	0.0501	0.0069	**0.0015**	0.0016
20%	**0,1334**	0.5730	0.9537	**0.0410**	0.4738	0.3451	**0.0318**	0.7298	0.3068
40%	1,9510	**0.7339**	3,1284	**0.4820**	1,6799	2,1844	0.4474	**0,0674**	1,4555

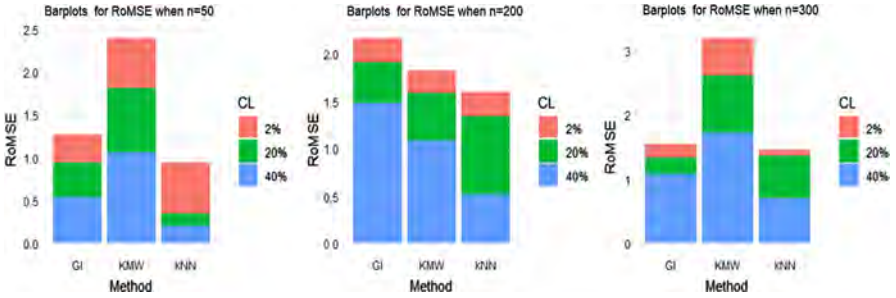

Fig. 3 Bar plots represent the $RoMSE$ scores for all sample sizes and censoring levels

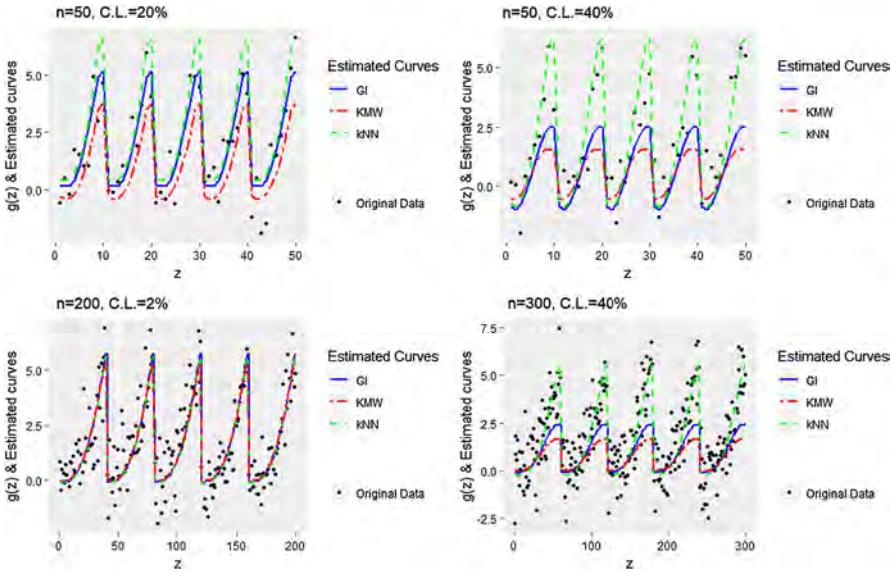

Fig. 4 Real observations and their estimated curves corresponding to the nonparametric part from *GI*, KMW and *k*NN, respectively, for different sample sizes and censoring levels

it can be said that *GI* and *k*NN have lower MSE values, especially for heavy censoring levels.

Figure 3, which displays the bar graphs of $RoMSE$ values, also supports the results given in Table 6. As can be seen from this graph, the *kNN* has the best performance,

Table 7 Outcomes of performance criteria for fitted values

C.L.		$n = 50$			$n = 200$			$n = 300$		
		MARE	GMSE	MAPE	MARE	GMSE	MAPE	MARE	GMSE	MAPE
2%	GI	0.0245	3.0930	0.0160	0.0173	0.8342	0.0234	0.0180	0.8279	**0.0219**
	kNN	**0.0217**	**1.6946**	**0.0158**	**0.0169**	**0.7329**	**0.0233**	**0.0173**	**0.6379**	**0.0219**
	KMW	0.0250	4.5642	0.0161	0.0177	0.8848	0.0234	0.0186	0.8705	**0.0219**
20%	GI	**0.0259**	**1.9820**	0.0159	0.0182	**0.9417**	0.0237	0.0189	1.1233	0.0220
	kNN	0.0277	2.3161	**0.0144**	**0.0111**	1.3550	**0.0229**	**0.0111**	**1.0893**	**0.0218**
	KMW	0.0358	3.2021	0.0165	0.0203	1.0987	0.0251	0.0207	1.2251	0.0240
40%	GI	0.0258	**4.1755**	**0.0180**	0.0212	**1.1567**	0.0252	**0.0219**	1.3122	0.1869
	kNN	**0.0220**	4.7934	0.0255	**0.0206**	2.8378	**0.0204**	0.0266	**1.2750**	**0.1689**
	KMW	0.0291	3.0430	0.0218	0.0257	2.3560	0.0281	0.0258	2.2614	0.1938

while the KMW has the worst performance. Figure 3 also shows the relative performance of the methods relative to each other. The basic idea here is that the comparison of solution techniques is to make clearer.

Figure 4 is designed for the estimates of the nonparametric component obtained from imputation techniques. In this sense, many different simulation configurations are analyzed here. But, it is not possible to show the details of each configurations due to occupying more space. Therefore, only a few of them are displayed in Fig. 4 for all censoring levels and sample sizes. In this context, the two top panels of Fig. 4 are obtained for $n = 50$ and censoring levels C.L. $= 20\%$ and 40%. The estimated curves for the low censoring level (i.e., C.L. $= 2\%$) are also given in the bottom left panel of the same figure. Here, the effect of the censoring rate can easily be seen in the bottom-right panel. Moreover, in each of the panels, estimated curves of the KMW seem worse than the others. By looking the top panels of Fig. 4, one can easily notice the improvement in the estimation from *kNN* when the censoring rate is getting larger.

5.3 Assessing the fitted values from semiparametric model

Finally, we evaluate the overall performance of the model with right-censored data. In this sense, Table 7 displays the results for the semiparametric time-series regression model with autoregressive errors defined in Sect. 2. Besides the fitting such a semiparametric model, it is also important to able to accurately estimate the parametric and nonparametric components of the model. For these purposes, the fitted values from the model for right-censored data based on *GI*, *k*NN and KMW techniques are assessed in terms of MARE, GMSE and MAPE criteria.

The outcomes in Table 7 denote that the KMW method designed for censored data performs poorly, whereas *k*NN method performs better in almost all simulation configurations. Furthermore, from Table 7, we observe that the *GI* method is understood to be the second best performing method after *k*NN. To see the results in more detail, bar graphs of the $RGMSE$ values for all simulation combinations are given in Fig. 5.

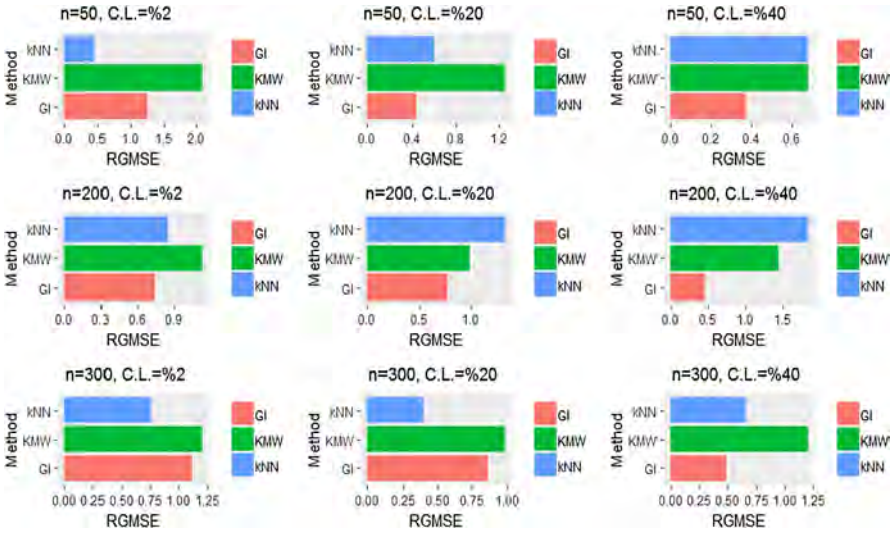

Fig. 5 Bar plots show the $RGMSE$ scores from kNN, GI and KMW for all sample sizes and censoring levels

In this context, a remarkable aspect of Fig. 5 is that it provides an alternative way to compare the data transformation techniques. It is interesting to note that even though KMW and GI are badly affected by censorship, they seem to have a more stable structure than kNN.

As noted earlier, the sample size or censorship level is not binding for the kNN method. This can be considered both an advantage and a disadvantage for kNN, because this method can give very good results under high censorship, as well as poor performance for low censoring levels. In Fig. 5, the results from the kNN method for $n = 300$ and $C.L. = 2\%$ can be shown as an example for this case.

6 Real data work

In this section, real-world data are considered to see the performances of the data transformation techniques designed for right-censored data. To achieve this goal, the data set showing the duration of unemployment is used. The data set includes the monthly unemployment period rates between 2004–2019 and is taken from the https://ec.europa.eu/eurostat/data/databas for Turkey. In this data set, none of 2004 and the last three months of 2019 are correctly obtained. Since these data points cannot take negative values, they can be censored from right to zero as a detection limit. Thus, the proposed analysis can be performed using this data set. In these sense, semiparametric time-series model can be written as follows

$$Unemp_t = \beta_1 Unemp_{(t-1)} + g(se_t) + \varepsilon_t, t = 1, \ldots, 186 \tag{6.1}$$

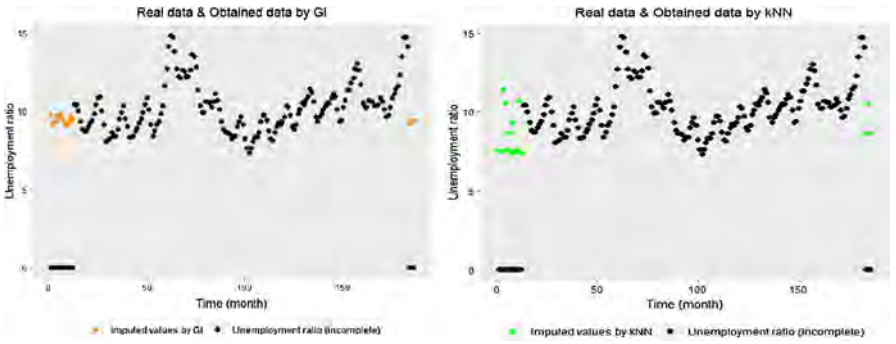

Fig. 6 Obtained new response variable and censored original data for two imputation methods

Table 8 Estimations from the parametric part of the model (6.1) with censored unemployment ratios

	GI	kNN	KMW
$\hat{\beta}_1$	0.2590	0.3792	1.0000
$Var\left(\hat{\beta}_1\right)$	0.0019	**0.0012**	0.0180

Table 9 Overall performance scores for fits from semiparametric model using *GI*, *k*NN and KMW

	MARE	GMSE	MAPE
GI	0.0602	0.0088	0.0270
*k*NN	**0.0108**	**0.0051**	**0.0209**
KMW	0.0655	0.2106	0.0552

where $Unemp_t$'s are the values of unemployment duration ratio depend on time, $Unemp_{(t-1)}$ denotes the first lag of the response variable, $se_t = (1, \ldots, n)^T$ is constructed to represent seasonality, and ε_t's are the random error terms with zero mean and constant variance.

As denoted before, to deal with censoring data problem the *k*NN and *GI* methods replace the censored observations with the imputed observations, while the KMW method uses the Kaplan–Meier weights. In this context, both the real and observations imputed with *k*NN and *GI* are shown in Fig. 6. Thus, by defining response observations (i.e., unemployment ratios) for three methods, we fit semiparametric model (6.1) with right-censored data. Tables 8 and 9 report the results for the parametric component of this model, whereas Fig. 7 displays the nonparametric component of the same model.

Table 8 displays the evaluation measures for parametric component. According to these result, it is clearly seen that the *k*NN imputation produces better estimates than other two methods. As in simulation experiments, it can also be said that KMW method does not give a good estimate.

Regarding nonparametric component, three different estimations of the unknown regression function are graphically illustrated in Fig. 7. The MSE values for these fittings designed by *GI*, *k*NN and KMW methods are also calculated as 0.7567, 0.8384 and 1.5258, respectively. Fitted the curves denoted by $\hat{g}_{GI}(se_t)$, $\hat{g}_{kNN}(se_t)$and $\hat{g}_{KMW}(se_t)$ are shown in Fig. 7.

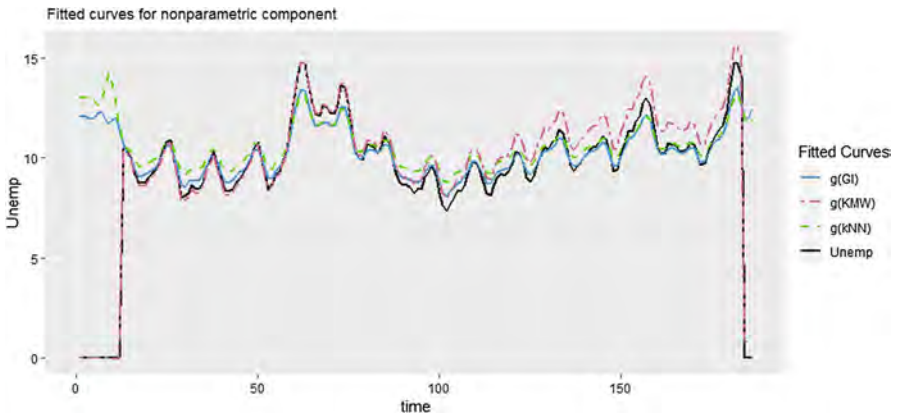

Fig. 7 Fitted curves for nonparametric component of the model

As can be seen in Fig. 7, although the estimated curve for KMW captures the actual data line, it tries to overcome censorship with increasing magnitudes of some data points that can be clearly detected after t = 100. The *GI* and *k*NN methods confirm the above MSE values and the values given in Table 8 and estimates from these methods follow each other closely. In addition, it is important that the fitted curves appear more understandable in terms of unemployment duration rates, since *k*NN and *GI* use imputed values. Overall performances of these methods are illustrated in Table 9.

From Table 9, we see that the estimate based on *k*NN has the best scores in terms of performance criteria. The KMW method has not shown a good performance especially for GMSE criterion but in general, similar to simulation study, scores are close to each other. Low censoring level (8%) can be presented as a reason for this case.

7 Concluding remarks

In this paper, we use penalized spline to fit a semiparametric regression model with right-censored time-series data. Since the censored data cannot be used directly with an ordinary statistical method such as penalized splines, a data transformation is generally required to solve this problem. For these purposes, we consider three different techniques GI, *k*NN and KMW. Note that Aydin and Yilmaz (2018) modified the ordinary penalized splines method to estimate a semiparametric regression model in which censored observations are replaced with synthetic data points. In this paper, we propose three different data transformation methods to solve the censored data problem. It should be noted that proposed methods are based on a generalization of the ordinary *GI*, *k*NN and KMW methods in case of the uncensored data. To achieve these ideas, Monte Carlo simulation experiments and a real data example are carried out. Accordingly, although three solution methods give the satisfying results, the *k*NN method works much better than the other two almost for all of the simulation combinations and the real data example. In addition, the findings obtained in this paper

show that the semiparametric regression model captures the changes of variability in the data and provides a reasonable fit to censored time-series.

The empirical results of both the real data and simulation studies confirmed that for all the methods, as expected the variances and bias values of the estimated coefficients start to decrease as the sample size n gets larger. Note also that for small sized sample, the bias values of coefficients increase as the censoring levels increase. One of the important ideas of this paper is that the kNN imputation method provides the satisfactory results in most cases.

In summary, the results of the simulation study show that although the GI and KMW methods give good results for low censoring level (2%), as the censoring levels increase, the kNN method improves and provides much better performance in estimating the parametric component of the right-censored semiparametric time-series model.

In terms of the nonparametric component, the kNN and GI methods give similar MSE scores. However, KMW does not give a satisfactory nonparametric function estimate. In addition, the performance of the three estimated models are evaluated by MARE, GMSE, MAPE and RGMSE and it is seen that kNN has had the best estimates.

In the real data study, unemployment rates are modeled with three introduced estimators and similar to the simulation study, kNN and GI methods provide better results than KMW with a high difference. The failure of KMW can be explained by the fact that the censored data points are far from uncensored due to Kaplan–Meier weights. Details are given in Sect. 3.3

Finally, the kNN method performs better than the other two methods in terms of performance criteria and the variance of estimates considered here, for all sample sizes and censoring levels.

MSE A possible extension of the proposed estimators can be obtained using different imputation techniques such as regression imputation, multiple imputation, SVD-based imputation, and so on. It can also be designed for different smoothing techniques such as kernel smoothing or smoothing spline for future research. Thus, significant contributions can be provided for improving of this study. In addition, new approaches can be developed for not only right-censored data, but also for time-series, left-censored and interval-censored data points.

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

Appendix A1: Derivation of Eqs. (2.12a–b)

To see derivation of Eqs. (2.12a–b), we first consider the penalized criterion defined in (2.10). According to this equation, the matrix and vector form

$$\begin{aligned} PRSS(\boldsymbol{\beta}, \mathbf{b}; \lambda) &= \sum_{t=1}^{n} A_t (S_t - x_t\boldsymbol{\beta} - g(z_t))^2 + \lambda \sum_{k=1}^{K} \mathbf{b}_{p+k}^2 \\ &= (\mathbf{S} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\mathbf{b})^T \mathbf{A} (\mathbf{S} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}\mathbf{b}) + \lambda \mathbf{b}^T \mathbf{D}\mathbf{b} \end{aligned}$$

Some simple algebraically show that

$$\begin{aligned} PRSS(\boldsymbol{\beta},\mathbf{b};\lambda) &= \left(\mathbf{S}^T\mathbf{A}-\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}-\mathbf{b}^T\mathbf{U}^T\mathbf{A}\right)(\mathbf{S}-\mathbf{X}\boldsymbol{\beta}-\mathbf{U}\mathbf{b})+\lambda\mathbf{b}^T\mathbf{D}\mathbf{b} \\ &= \mathbf{S}^T\mathbf{A}\mathbf{S}-\mathbf{S}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta}-\mathbf{S}^T\mathbf{A}\mathbf{U}\mathbf{b}-\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{S}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta} \\ &\quad +\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{U}\mathbf{b}-\mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{S}+\mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta}+\mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{U}\mathbf{b}+\lambda\mathbf{b}^T\mathbf{D}\mathbf{b} \\ &= \mathbf{S}^T\mathbf{A}\mathbf{S}-2\mathbf{S}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta}-2\mathbf{S}^T\mathbf{A}\mathbf{U}\mathbf{b}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta}+2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{U}\mathbf{b} \\ &\quad +\mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{U}\mathbf{b}+\lambda\mathbf{b}^T\mathbf{D}\mathbf{b} \end{aligned} \tag{A1.1}$$

In order to find the minimizers of (A1.1), we set the partial derivatives of this expression to zero. From (A1.1), it follows that the partial derivate of (A1.1) with respect to $\mathbf{b}$ is

$$\frac{\partial PRSS}{\partial \mathbf{b}} = -2\mathbf{S}^T\mathbf{A}\mathbf{U}+2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{U}+2\mathbf{U}^T\mathbf{A}\mathbf{U}\mathbf{b}+2\lambda\mathbf{D}\mathbf{b}=0 \tag{A1.2}$$

Replacing $\mathbf{b}$ by $\hat{\mathbf{b}}$, and after some algebra we find that

$$\hat{\mathbf{b}} = \left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}(\mathbf{S}-\mathbf{X}\boldsymbol{\beta}) \tag{A1.3}$$

as claimed in the main text.

Similarly, the partial derivate of (A1.1) with regard to $\boldsymbol{\beta}$ is

$$\frac{\partial PRSS_m}{\partial \boldsymbol{\beta}} = -2\mathbf{S}^T\mathbf{A}\mathbf{X}+2\mathbf{X}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta}+2\mathbf{X}^T\mathbf{A}\mathbf{U}\mathbf{b}=0 \tag{A1.4}$$

Simple algebra shows that

$$\begin{aligned} \mathbf{X}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta} &= \mathbf{S}^T\mathbf{A}\mathbf{X}-\mathbf{X}^T\mathbf{A}\mathbf{U}\mathbf{b} \\ \mathbf{X}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{A}(\mathbf{S}-\mathbf{U}\mathbf{b}) \end{aligned} \tag{A1.5}$$

Substituting Equation (A1.3) into Equation (A1.5), we get

$$\mathbf{X}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{A}\left[\mathbf{S}-\mathbf{U}\mathbf{X}^T\left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}(\mathbf{S}-\mathbf{X}\boldsymbol{\beta})\right]$$

$$\mathbf{X}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{A}\mathbf{S}-\mathbf{X}^T\mathbf{A}\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{S}+\mathbf{X}^T\mathbf{A}\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{X}\boldsymbol{\beta}$$

$$\left[\mathbf{X}^T\mathbf{A}\mathbf{X}-\mathbf{X}^T\mathbf{A}\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{X}\right]\boldsymbol{\beta} = \mathbf{X}^T\mathbf{A}\mathbf{S}-\mathbf{X}^T\mathbf{A}\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{S}$$

as stated in (A1.3), replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and simple algebra shows that

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}^T\mathbf{A}\mathbf{X}-\mathbf{X}^T\mathbf{A}\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{X}\right]^{-1}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{U}+\lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\right)\mathbf{X}^T\mathbf{A}\mathbf{S} \tag{A1.6}$$

as described in the main text.

Appendix A2: Derivation of Eqs. (3.10a–b)

In the context of KMW, the penalized least-squares estimates are the values of $\hat{\boldsymbol{\beta}}_{KM}$ and $\hat{\mathbf{b}}_{KM}$ that minimize the criterion (3.8), given by $PRSS_{KM}(\boldsymbol{\beta};\mathbf{b}) = (\mathbf{S}-\mathbf{X}\boldsymbol{\beta}-\mathbf{U}\mathbf{b})^T\mathbf{A}\mathbf{W}(\mathbf{S}-\mathbf{X}\boldsymbol{\beta}-\mathbf{U}\mathbf{b}) + \lambda\mathbf{b}^T\mathbf{D}\mathbf{b}$

This expression could be written as

$$\begin{aligned} PRSS_{KM}(\boldsymbol{\beta};\mathbf{b}) &= \left(\mathbf{S}^T\mathbf{A}\mathbf{W} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W} - \mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{W}\right)(\mathbf{S}-\mathbf{X}\boldsymbol{\beta}-\mathbf{U}\mathbf{b}) + \lambda\mathbf{b}^T\mathbf{D}\mathbf{b} \\ &= \mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{S} - \mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} - \mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{S} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} \\ &+ \quad \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} - \mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{S} + \mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} + \lambda\mathbf{b}^T\mathbf{D}\mathbf{b} \\ &= \mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{S} - 2\mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} \\ &+ 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} + \mathbf{b}^T\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} + \lambda\mathbf{b}^T\mathbf{D}\mathbf{b} \end{aligned} \tag{A2.1}$$

Similar to the procedures that used in equation (A1.1), the partial derivate of (A2.1) with respect to $\mathbf{b}$ is

$$\frac{\partial PRSS_m}{\partial \mathbf{b}} = -2\mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{U} + 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{U} + 2\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} + 2\lambda\mathbf{D}\mathbf{b} = 0 \tag{A2.2}$$

Equation (A2.1) could be written as follows

$$\begin{aligned} \mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} + \lambda\mathbf{D}\mathbf{b} &= \mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{U} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{U} \\ \left(\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)\mathbf{b} &= \mathbf{U}^T\mathbf{A}\mathbf{W}(\mathbf{S}-\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

After some algebra, we find that the estimator $\hat{\mathbf{b}}_{KM}$ of $\mathbf{b}$ is

$$\hat{\mathbf{b}}_{KM} = \left(\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{W}(\mathbf{S}-\mathbf{X}\boldsymbol{\beta}) \tag{A2.3}$$

as determined in Sect. 3.3.

Similarly, the partial derivate of (A2.1) with regard to $\boldsymbol{\beta}$ is

$$\frac{\partial PRSS_m}{\partial \boldsymbol{\beta}} = -2\mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{X} + 2\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} + 2\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} = 0 \tag{A2.4}$$

From (A2.3), it follows that

$$\begin{aligned} \mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} &= \mathbf{S}^T\mathbf{A}\mathbf{W}\mathbf{X} - \mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{U}\mathbf{b} \\ \mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{A}\mathbf{W}(\mathbf{S}-\mathbf{U}\mathbf{b}) \end{aligned} \tag{A2.5}$$

Substituting Equation (A2.3) into Equation (A2.5), we obtain

$$\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{A}\mathbf{W}\left[\mathbf{S} - \mathbf{U}\mathbf{X}^T\left(\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{W}(\mathbf{S}-\mathbf{X}\boldsymbol{\beta})\right]$$

$$\mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{S} - \mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{S} + \mathbf{X}^T\mathbf{A}\mathbf{W}\mathbf{X}\left(\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{U} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{A}\mathbf{W}\mathbf{X}\boldsymbol{\beta}$$

$$\left[\mathbf{X}^T\mathbf{AWX} - \mathbf{X}^T\mathbf{AWX}\left(\mathbf{U}^T\mathbf{AWU} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{AWX}\right]\boldsymbol{\beta} = \mathbf{X}^T\mathbf{AWS} - \mathbf{X}^T\mathbf{AWX}\left(\mathbf{U}^T\mathbf{AWU} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{AWS}$$

Consequently, after a bit of algebra we find that the estimator $\hat{\boldsymbol{\beta}}_{KM}$ of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{KM} = \left[\mathbf{X}^T\mathbf{AWX} - \mathbf{X}^T\mathbf{AWX}\left(\mathbf{U}^T\mathbf{AWU} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{AWX}\right]^{-1}\left(\mathbf{I} - \mathbf{X}\left(\mathbf{U}^T\mathbf{AWU} + \lambda\mathbf{D}\right)^{-1}\mathbf{U}^T\mathbf{AW}\right)\mathbf{X}^T\mathbf{AWS} \tag{A2.6}$$

as described in Sect. 3.3.

References

Aneiros-Perez G, Cao R, Vilar-Fernandez JM (2011) Functional methods for time-series prediction: a nonparametric approach. J Forecast 30(4):377–392

Aydın D, Yılmaz E (2018) Modified estimators in semiparametric regression models with right-censored data. J Stat Comput Simul 88(8):1470–1498

Batista G, Monard M (2002) An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell 17(5–6):519–533

Box G, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco

Brockwell PJ, Davis RA (1991) Time series: theory and methods, 2nd edn. Springer, New-York. https://doi.org/10.1007/978-1-4419-0320-4

Chen J, Shao J (2000) Nearest neighbor imputation for survey data. J Off Stat 16(2):113–131

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11(2):89–121

Faubel F, Mcdonough J, Klakow D (2009) Bounded conditional mean imputation with Gaussian mixture models: a reconstruction approach to partly occluded features. In: 2009 IEEE international conference on acoustics, speech and signal processing

Gao J (1995) Asymptotic properties of some estimators for partly linear stationary autoregressive models. Commun Stat Theory Methods 24(8):2011–2026

Gao J (2007) Nonlinear time-series: semiparametric and nonparametric methods. Monogr Stat Appl Probab 108:1–237

Gao J, Philips PCB (2010) Semiparametric estimation in simultaneous equations of time-series models, School of Economics Working Papers, 26, The University of Adelaide School of Economics

Hardle W, Lütkepohl H, Chen R (1997) A review of nonparametric time-series analysis. Int Stat Rev 65(1):49–72

Helsel DR (1990) Less than obvious: statistical treatment of data below the detection limit. Environ Sci Technol 24:1766–1774

Hurvich CM, Simonoff JS, Tsai C-L (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Stat Soc Stat Methodol Ser B 60(2):271–293

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53(282):457–481

Kato R, Shiohama T (2009) Model and variable selection procedures for semiparametric time-series regression. J Probab Stat. https://doi.org/10.1155/2009/487194

Koul H, Susarla V, Van Ryzin J (1981) Regression analysis with randomly right-censored data. Ann Stat 9:1276–1285

Lee KJ, Carlin JB (2010) Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol 171(5):624–632

Lee YK, Mammen E, Nielsen JP, Park BU (2018) In-sample forecasting: a brief review and new algorithms. ALEA Lat Am J Probab Math Stat 15(2):875–895

Li R, Liang H (2008) Variable selection in semiparametric regression modelling. Ann Stat 36(1):261–286

Liang H (2006) Estimation in partially linear models and numerical comparisons. Comput Stat Data Anal 50(3):675–687

Linton O, Nielsen SF, Nielsen JP (2009) Nonparametric regression with a latent time-series. Econom J 12(2):187–207

Malarvizhi R, Thanamani AS (2012) K-nearest neighbor in missing data imputation. Int J Eng Res Dev 5(1):5–7

Miller RG (1976) Least squares regression with censored data. Biometrika 63:449–464

Morton R, Kang EL, Henderson BL (2009) Smoothing splines trend estimation and prediction in time-series. Environmetrics 20(3):249–259

Park JW, Genton MG, Ghosh SK (2007) Censored time series analysis with autoregressive moving average models. Can J Stat 35(1):151–168

Park JW, Genton MG, Ghosh SK (2009) Nonparametric autocovariance estimation from censored time-series by Gaussian imputation. J Nonparametr Stat 21(2):241–259

Ruppert D, Wand MP, Carrol R (2003) Semiparametric regression. Cambridge University Press, Cambridge

Silva DSF, Deutsch CV (2017) Multiple imputation framework for data assignment in truncated pluri-Gaussian simulation. Stoch Environ Res Risk Assess 31:2251–2263

Stute W (1993) Consistent estimation under random censorship when covariables are present. J Multivar Anal 45:89–103

Truong YK, Stone CJ (1994) Semiparametric time-series regression. J Time-Ser Anal 15(4):405–428

Vaida F, Liu L (2009) Fast implementation for normal mixed effects models with censored response. J Comput Graph Stat 18(4):797–817

Yu J, Chen G (2007) Semiparametric generalized least squares estimation in partially linear regression models with correlated errors. J Stat Plan Inference 137(1):117–132

Yuan K-H (2009) Normal distribution based pseudo ML for missing data: with applications to mean and covariance structure analysis. J Multivar Anal 100(9):1900–1918