Journal of the Korean Statistical Society 45 (2016) 358–370

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Interval-valued data regression using nonparametric additive models

Changwon Lim [*]

Department of Applied Statistics, Chung-Ang University, Seoul 156-756, Republic of Korea

ARTICLE INFO

Article history:
Received 10 June 2015
Accepted 27 December 2015
Available online 28 January 2016

AMS 2000 subject classifications:
primary 62G08
secondary 62J99

Keywords:
Interval-valued data
Nonparametric additive model
Symbolic data
Penalized regression spline
Generalized cross validation

ABSTRACT

Interval-valued data are observed as ranges instead of single values and frequently appear with advanced technologies in current data collection processes. Regression analysis of interval-valued data has been studied in the literature, but mostly focused on parametric linear regression models. In this paper, we study interval-valued data regression based on nonparametric additive models. By employing one of the current methods based on linear regression, we propose a nonparametric additive approach to properly analyze interval-valued data with a possibly nonlinear pattern. We demonstrate the proposed approach using a simulation study and a real data example, and also compare its performance with those of existing methods.

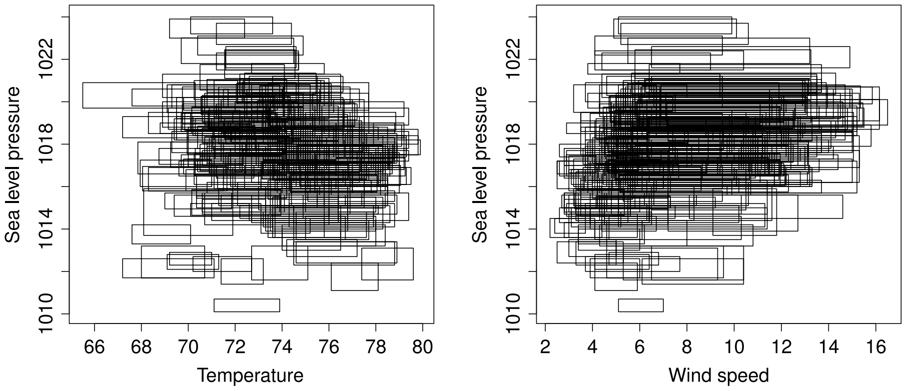
1. Introduction

Humans have been interested in weather forecasting through the ages. Reoccurring astronomical and meteorological events were used to record seasonal changes in the weather during early times. After developing instruments to measure the properties of the atmosphere, such as temperature, pressure, and humidity, efforts were made to understand the atmosphere using the measurements of the properties. Knowledge of the atmosphere has been considered a key factor for weather forecasting (Lutgens & Tarbuck, 2007). Statistical weather forecasting is a method of weather prediction using statistical models to describe relations among such meteorological variables. It was first studied during the mid-twentieth century by Wadsworth (1951) and Wadsworth, Bryan, and Gordon (1948). After that, statistical prediction of variability in weather or meteorological variables such as surface temperature and sea level pressure has been an important problem and studied by many researchers for decades. For example, see Barnett (1985), Davis (1976), Gillett, Zwiers, Weaver, and Stott (2003), Kutzbach (1967), and Min, Legutke, Hense, and Kwon (2005), among others. Stations were constructed world wide to observe weather and meteorological variables, and with development of science and technology the number of stations and the amount of data generated from them have increased exponentially.

Although computing power is highly advanced recently, sometimes it is not practical to analyze massive sized data sets. Consequently, such huge data sets are aggregated to intervals with lower and upper bounds or to histograms. Researchers sometimes encounter interval-valued data, which are either inherently observed as or processed to be intervals. Examples are blood pressure (Billard & Diday, 2000) and income level in survey data (Xu, 2010), among others. Interval-valued data belong to a broader category of data forms called symbolic data (Diday, 1987). It is difficult to analyze these types of data

[*] Tel.: +82 2 820 5547.
E-mail address: clim@cau.ac.kr.

http://dx.doi.org/10.1016/j.jkss.2015.12.003
1226-3192/© 2016 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.



(a) Sea level pressure vs. temperature.

(b) Sea level pressure vs. wind speed.

Fig. 1. Plots of interval-valued data for Hawaiian climate data.

with classical methods. To illustrate this point, we consider a real interval-valued Hawaiian climate data set. There are three random variables: X_1 = the daily temperature in Hawaii, X_2 = the daily wind speed in Hawaii, and Y = the daily sea level pressure in Hawaii. The interval-valued data used in this illustration were converted from a total of 5408 single-valued observations which are collected in 2012 from 16 stations. The lower bound and the upper bound of the intervals are the Q1 and Q3 of the 16 stations, respectively. The sample size of the data is thus 366. The original data are publicly available from the National Climate Data Center at http://www.ncdc.noaa.gov/.

Fig. 1 shows the plots of the interval-valued data for Hawaiian climate data. We observe a decreasing pattern for the relationship between the sea level pressure and the temperature and an increasing pattern between the sea level pressure and the wind speed. However, the main difficulty is to take into account internal variation or structure within an observation, that is, an interval.

Researchers have studied adaptation of classical methods to the symbolic data extensively. For example, see Diday (1995), Diday and Emilion (1996, 1998), and Diday, Emilion, and Hillali (1996), among others. After the establishment of the adaptation, researchers have considered regression approaches to interval-valued data actively. Billard and Diday (2000) introduced a regression approach first, which fits a linear regression model on the center point of the intervals and applies the fitted model to the lower and the upper bounds of the predictor variables to obtain a prediction. Lima Neto, de Carvalho, and Tenorio (2004) extended this approach to the range of the intervals and proposed a regression method. This method fits two separate linear regression models on the center and the range of the intervals. Later, Billard and Diday (2007) employed this idea and proposed a bivariate approach which fits two regression models on both of the center and the range of the intervals simultaneously as the predictors. Recently, Lima Neto, Cordeiro, and de Carvalho (2011); Lima Neto, Cordeiro, Carvalho, Anjos, and Costa (2009) considered the bivariate generalized linear model by Iwasaki and Tsubaki (2005) to analyze interval-valued data, and Lima Neto and de Carvalho (2010) introduced a method for fitting a constrained linear regression model to interval-valued data. The proposed method fits a constrained linear regression model on the center point and range of the interval values. Xu (2010) proposed a symbolic covariance method based on the symbolic sample covariance introduced by Billard (2007, 2008). Research for analyzing interval-valued data has been a very active area and considered using various approaches. See, for example, Ahn, Peng, Park, and Jeon (2012), Blanco-Fernandez, Corral, and Gonzalez-Rodriguez (2011), Silva, Lima Neto, and Anjos (2011), and Yang, Jeng, Chuang, and Tao (2011) among others, and Blanco-Fernandez, Colubi, and Gonzalez-Rodriguez (2013) for a recent review.

Regression approaches for interval-valued data in the literature have been mostly developed based on linear regression models as described above. However, there might be cases where interval-valued data are not generated from a linear regression model, but some nonlinear regression model. One could check scatter plots of data to see if there are nonlinear patterns between the response variable and some of the explanatory variables. For such cases it may not be appropriate to use the existing regression approaches for interval-valued data.

In this paper we consider regression analysis of interval-valued data based on nonparametric additive models in order to provide a better prediction for interval-valued data with nonlinear patterns. In many practical applications, using linear regression models is too restrictive and may have a problem of misspecification. To avoid such limitations, researchers often prefer to use nonparametric regression models such as kernel regression, local polynomial, k-nearest neighbors, and so on. The reason is that nonparametric regression models make few assumptions about the regression function. However, because of the same reason they are very difficult to interpret compared to the classic linear models. Not only that, completely unstructured nonparametric regressions would not work well due to the curse of dimensionality (Friedman & Stuetzle, 1981). Nonparametric additive models (Buja, Hastie, & Tibshirani, 1989; Hastie & Tibshirani, 1990; Stone, 1985) provide a useful compromise between the restrictive linear model and the fully unstructured nonparametric model.

The additive model is a special case of the projection pursuit regression model proposed by Friedman and Stuetzle (1981). The alternating least squares model (van der Burg & de Leeuw, 1983) and the alternating conditional expectation model

(Breiman & Friedman, 1985) also contain the additive model as a special case. The procedure called the backfitting algorithm (Buja et al., 1989; Friedman & Stuetzle, 1981) can be used to fit the additive model.

The nonparametric additive model has been extended to the generalized additive model (GAM) introduced by Hastie and Tibshirani (1984). Like the generalized linear model it assumes that the response variable comes from an exponential family. The GAM has been studied extensively by many researchers (e.g., Hastie & Tibshirani, 1990; Horowitz & Mammen, 2011; Linton & Härdle, 1996 and Yang, Sperlich, & Härdle, 2003). This model is estimated by penalized maximum likelihood estimation where the penalized likelihood is maximized by penalized iteratively reweighted least squares (P-IRLS). Using the backfitting algorithm in P-IRLS has been considered, but estimating the smoothing parameters was difficult to integrate into this approach (Wood, 2004). Then, Wood (2000, 2004) proposed a generalized cross validation (GCV) method which can be used for the GAM with penalized regression splines. The (generalized) nonparametric additive model is still an active research area. See Carroll, Maity, Mammen, and Yu (2009), Curtis, Banerjee, and Ghosal (2014), McLean, Hooker, Staicu, Scheipl, and Ruppert (2014), Wong, Yao, and Lee (2014) and Yu, Park, and Mammen (2008) for examples; Horowitz (2014) for a recent review.

A new regression approach for interval-valued data based on a nonparametric additive model is proposed in Section 2, after describing some of the current methods for analyzing interval-valued data based on the linear regression model. The performances of the proposed method are compared with those of the existing methods on the basis of simulation studies in Section 3. In Section 4, the proposed method is illustrated using a real data set which is publicly available. We conclude this article in Section 5 by providing discussion and future research topics.

2. Methodology

Let $X_1, \ldots, X_p$ be p interval-valued explanatory variables, and Y be the interval-valued response variable. That is, we observe their realizations in intervals: $X_{ij} = [X_{Lij}, X_{Uij}] \subset \mathbb{R}$, with $X_{Lij} \le X_{Uij},\ X_{Lij}, X_{Uij} \in \mathbb{R}$, and $Y_i = [Y_{Li}, Y_{Ui}] \subset \mathbb{R}$ with $Y_{Li} \le Y_{Ui},\ Y_{Li}, Y_{Ui} \in \mathbb{R}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

2.1. Current methods

Although we consider the nonparametric additive model to analyze interval-valued data in this paper, most of the current methods have been studied for the linear regression model. Thus, we first state two of them here based on the following linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T,\ \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T,\ \mathbf{X}_i = (1, X_{i1}, \ldots, X_{ip})^T$ for $i = 1, \ldots, n,\ \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T,\ \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ and ϵ_i are independently distributed from $N(0, \sigma^2)$.

Billard and Diday (2000) proposed the center method (CM) that fits a linear regression model to the midpoint of the interval values. Let $X_1^c, \ldots, X_p^c, Y^c$ be the center points of the intervals $X_1, \ldots, X_p, Y$, respectively. Then, the CM changes the original model (1) into a standard linear regression model by

$$\mathbf{Y}^c = \mathbf{X}^c\boldsymbol{\beta}^c + \boldsymbol{\epsilon}^c, \tag{2}$$

where $\mathbf{Y}^c = (Y_1^c, \ldots, Y_n^c)^T,\ \mathbf{X}^c = (\mathbf{X}_1^c, \ldots, \mathbf{X}_n^c)^T,\ \mathbf{X}_i^c = (1, X_{i1}^c, \ldots, X_{ip}^c)^T$ for $i = 1, \ldots, n,\ \boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, \ldots, \beta_p^c)^T,\ \boldsymbol{\epsilon}^c = (\epsilon_1^c, \ldots, \epsilon_n^c)^T$. An estimator of $\boldsymbol{\beta}^c$, $\widehat{\boldsymbol{\beta}}^{\mathbf{c}}$ is obtained by the least squares estimation. The CM applies the fitted model to the lower and upper bounds of the interval of explanatory variables to predict the lower and upper bounds of the interval of the response variables, respectively.

Lima Neto et al. (2004) proposed the center and range method (CRM). In addition to the same center model given in (2), the CRM fits another linear regression model to the range of the intervals. Let $X_1^r, \ldots, X_p^r, Y^r$ be the ranges of the intervals of $X_1, \ldots, X_p, Y$, respectively. Also, let the observed values of X_j^r and Y^r be $X_{ij}^r = X_{Uij} - X_{Lij}$ and $Y_i^r = Y_{Ui} - Y_{Li}$, respectively, where $i = 1, \ldots, n$, and $j = 1, \ldots, p$. Then, the range model is given by

$$\mathbf{Y}^r = \mathbf{X}^r\boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r, \tag{3}$$

where $\mathbf{Y}^r = (Y_1^r, \ldots, Y_n^r)^T,\ \mathbf{X}^r = (\mathbf{X}_1^r, \ldots, \mathbf{X}_n^r)^T,\ \mathbf{X}_i^r = (1, X_{i1}^r, \ldots, X_{ip}^r)^T$ for $i = 1, \ldots, n,\ \boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \ldots, \beta_p^r)^T$ and $\boldsymbol{\epsilon}^r = (\epsilon_1^r, \ldots, \epsilon_n^r)^T$. Again, an estimator, $\widehat{\boldsymbol{\beta}}^{\mathbf{r}}$ is obtained by the least squares estimation. The center and the range of the response variable are predicted separately, and the predicted interval $\widehat{Y} = [\widehat{Y}_L, \widehat{Y}_U]$ is obtained by

$$\widehat{Y}_L = \widehat{Y}^c - \widehat{Y}^r/2, \qquad \widehat{Y}_U = \widehat{Y}^c + \widehat{Y}^r/2, \tag{4}$$

where $\widehat{Y}^c$ and $\widehat{Y}^r$ are the predicted values under the models (2) and (3), respectively.

Recently, Xu (2010) proposed the symbolic covariance method (SCM) that uses the symbolic covariance matrix proposed by Billard (2007, 2008). He considered the following model with centered variables,

$$\mathbf{Y} - \bar{\mathbf{Y}} = (\mathbf{X} - \bar{\mathbf{X}})\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

where $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are the symbolic sample mean matrix of the explanatory variables and the vector of the symbolic sample means of the response variable, respectively. The symbolic sample mean of interval-valued variable X_j $(j = 1, \ldots, p)$ is defined as (Bertrand & Goupil, 2000)

$$\bar{X}_j = \frac{1}{2n}\sum_{i=1}^{n}(X_{Lij} + X_{Uij}).$$

The least squares estimator $\widehat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \{(\mathbf{X}-\bar{\mathbf{X}})^T(\mathbf{X}-\bar{\mathbf{X}})\}^{-1}(\mathbf{X}-\bar{\mathbf{X}})^T(\mathbf{Y}-\bar{\mathbf{Y}}) \\ &= \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY},\end{aligned} \tag{6}$$

where $\mathbf{S}_{XX}$ is the symbolic sample variance–covariance matrix of the explanatory variables and $\mathbf{S}_{XY}$ is the vector of the symbolic sample covariances between the explanatory variables and the response variable. The symbolic sample covariance between X_j and X_k $(j, k = 1, \ldots, p)$ is defined as follows (Billard, 2007, 2008):

$$\begin{aligned}\mathrm{Cov}(X_j, X_k) = \frac{1}{6n}\sum_{i=1}^{n}\Big[&2(X_{Lij}-\bar{X}_j)(X_{Lik}-\bar{X}_k) + (X_{Lij}-\bar{X}_j)(X_{Uik}-\bar{X}_k) \\ &+ (X_{Uij}-\bar{X}_j)(X_{Lik}-\bar{X}_k) + 2(X_{Uij}-\bar{X}_j)(X_{Uik}-\bar{X}_k)\Big].\end{aligned} \tag{7}$$

2.2. Proposed method

Suppose $(Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$ are n independent samples from the following nonparametric additive model:

$$Y = \mu + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{8}$$

where μ is an intercept term, f_j's are unknown smooth functions, and ϵ is an unobserved random variable with mean zero and finite variance σ^2. It should be assumed that $E[f_j(X_j)] = 0$ for $j = 1, \ldots, p$ in order to prevent identifiability problems. This model includes the linear regression model as a special case, where $f_j(X_j) = \beta_j X_j$, but it is more general. This is nonparametric because f_j's can be any arbitrary nonlinear functions without assuming any parameters. However, the idea is still the same as the linear model that each explanatory variable makes a separate contribution to the response variable and they just add up.

The functions f_j can be estimated using the backfitting algorithm (Buja et al., 1989; Friedman & Stuetzle, 1981), where one can use an arbitrary smoother to estimate the functions. The algorithm is as follows:

(1) Initialize. Set $\hat{\mu} = \bar{Y}$ and set $f_j^{(0)} = 0$ for $j = 1, \ldots, p$

(2) Cycle. For $k = 1, \ldots, p$ set

$$f_k^{(l+1)} = S\left[\left\{Y_i - \hat{\mu} - \sum_{j\neq k} f_j^{(l)}\right\}_1^n\right].$$

(3) Until. The individual functions converges.

Here, S is a smoothing operator which is usually selected to be a cubic spline smoother, but can be any other appropriate fitting smoothers.

There is another approach for representing the smooth functions, which uses penalized regression splines. In penalized regression splines the model's smoothness can be controlled by adding a penalty function to the fitting objective. For a cubic spline, let the knot locations and the basis functions be denoted by $\{X_{jk}^* : k = 1, \ldots, q_j{-}2\}$ and $R(X_j, X_{jk}^*)$ for $k = 1, \ldots, q_j{-}2$, respectively, for $j = 1, \ldots, p$. Using this cubic spline basis for f_j the additive model (8) becomes a linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the ith row of the model matrix is $\mathbf{X}_i = (\mathbf{v}_1^T, \ldots, \mathbf{v}_p^T)$, $\mathbf{v}_j = (1, X_j, R(X_j, X_{j1}^*), \ldots, R(X_j, X_{j,(q_j-2)}^*))^T$ for $j = 1, \ldots, p$, and $\boldsymbol{\beta}$ is the vector of the unknown parameters. Then, the penalized regression splines minimizes

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p}\theta_j\int\left[f_j''(x)\right]^2 dx,$$

where θ_j's are the smoothing parameters estimated by cross validation. See Wood (2000, 2004, 2011) for details.

When there are several smooth functions in an additive model, some of them might not be important to explain the variation in the response variable. In order to address such problem, one should consider performing a hypothesis testing for some subset, $\boldsymbol{\beta}_j$, of $\boldsymbol{\beta}$. Let $\mathbf{V}_{\hat{\beta}_j}$ be the covariance matrix of $\hat{\boldsymbol{\beta}}_j$. We have that under the null hypothesis $\boldsymbol{\beta}_j = \mathbf{0}$ $\hat{\boldsymbol{\beta}}_j$ is approximately normally distributed with mean vector $\mathbf{0}$ and the covariance matrix $\mathbf{V}_{\hat{\beta}_j}$. Therefore, the p-value for the test that $\boldsymbol{\beta}_j = \mathbf{0}$ is calculated based on the approximate result,

$$\frac{\hat{\boldsymbol{\beta}}_j^T \hat{\mathbf{V}}_{\hat{\beta}_j}^{r-} \hat{\boldsymbol{\beta}}_j / r}{\hat{\phi}/(n - \text{edf})} \sim F_{r,\,\text{edf}},$$

where $r = \text{rank}(\mathbf{V}_{\hat{\beta}_j})$, $\mathbf{V}_{\hat{\beta}_j}^{r-}$ is the rank r pseudoinverse of the covariance matrix, ϕ is an unknown scale parameter, and 'edf' denotes the estimated degrees of freedom for the model. See Wood (2006) for details.

We introduce the CRM for interval-data linear regression to the above nonparametric additive model (8) and propose the center and range additive model (CRAM). The CRAM fits two separate nonparametric additive models to the center point and the range of the intervals, respectively, as follows:

$$Y_i^c = \mu^c + \sum_{j=1}^{p} f_j^c(X_{ij}^c) + \epsilon_i^c, \tag{9}$$

$$Y_i^r = \mu^r + \sum_{j=1}^{p} f_j^r(X_{ij}^r) + \epsilon_i^r, \tag{10}$$

where μ^c and μ^r are intercept terms, f_j^c's and f_j^r's are unknown smooth functions, and ϵ_i^c and ϵ_i^r are unobserved random variables with mean zeros and finite variances σ_c^2 and σ_r^2, respectively. Then, the predicted interval $\widehat{Y} = [\widehat{Y}_L, \widehat{Y}_U]$ is obtained by the same manner as in the CRM (4).

3. Simulation studies

In this simulation study we compared the proposed CRAM with the two existing methods, the CRM and the SCM. Note that the CRAM is for nonparametric additive regression models while the CRM and the SCM are for parametric linear regression models. Since there are no current methods for analyzing interval-valued data in additive regression, it is meaningful to compare the proposed method with those.

3.1. Study design

Let $X_1, X_2, \ldots, X_{10}$ be ten interval-valued explanatory variables, and Y be the interval-valued response variable, where $X_{ij} = [X_{Lij}, X_{Uij}]$ for $i = 1, \ldots, n$, $j = 1, \ldots, 10$, and $Y_i = [Y_{Li}, Y_{Ui}]$ for $i = 1, \ldots, n$, as defined in Section 2. Let X_{ij}^c and Y_i^c be the centers of jth explanatory variable and the response variable of the ith observation, respectively. Also, let X_{ij}^r and Y_i^r be the ranges of jth explanatory variable and the response variable of the ith observation, respectively. Then, we considered three simulation settings as follows:

1. Setting I (A linear regression model based on CRM):
   (i) For ith observation, we first generated X_{i1}^c uniformly from $\{11, \ldots, 20\}$, X_{i2}^c from $\{21, \ldots, 30\}$, X_{i3}^c from $\{31, \ldots, 50\}$, X_{i4}^c from $\{51, \ldots, 70\}$, X_{i5}^c from $\{71, \ldots, 90\}$, X_{i6}^c from $\{91, \ldots, 110\}$, X_{i7}^c from $\{111, \ldots, 130\}$, X_{i8}^c from $\{131, \ldots, 150\}$, X_{i9}^c from $\{151, \ldots, 170\}$, and $X_{i,10}^c$ from $\{171, \ldots, 190\}$ for $i = 1, \ldots, n$.
   (ii) Then we generated Y_i^c using the following linear regression model:
   $$Y_i^c = \beta_0 + \beta_1 X_{i1}^c + \cdots + \beta_{10} X_{i,10}^c + \epsilon_i^c. \tag{11}$$
   The values of regression coefficients were set to be $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_{10}) = (0, 0.3, -0.6, 0.5, -0.9, 0, 0, 0, 0, 0, 0)$. The errors ϵ_i^c are normally distributed with mean 0 and variance σ^2.
   (iii) For the ith observation, we randomly generated X_{i1}^r from $U(1, 2)$, X_{i2}^r from $U(2, 3)$, X_{i3}^r from $U(3, 4)$, X_{i4}^r from $U(4, 5)$, X_{i5}^r from $U(5, 6)$, X_{i6}^r from $U(6, 7)$, X_{i7}^r from $U(7, 8)$, X_{i8}^r from $U(8, 9)$, X_{i9}^r from $U(9, 10)$, $X_{i,10}^r$ from $U(10, 11)$, and Y_i^r from $U(1, 2)$ for $i = 1, \ldots, n$.
2. Setting II (A linear regression model based on SCM):
   (i) For ith observation, we first generated X_{Lij} and X_{Uij} from $N(0, 1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, 10$.
   (ii) Then we generated Y_{Li} and Y_{Ui} using the following linear regression model:
   $$\begin{aligned} Y_{Li} &= \beta_0 + \beta_1 X_{Li1} + \cdots + \beta_{10} X_{Li,10} + \epsilon_{Li}; \\ Y_{Ui} &= \beta_0 + \beta_1 X_{Ui1} + \cdots + \beta_{10} X_{Ui,10} + \epsilon_{Ui}. \end{aligned} \tag{12}$$
   The values of regression coefficients were set to be $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_{10}) = (0, 0.3, -0.6, 0.5, -0.9, 0, 0, 0, 0, 0, 0)$, which is the same as in Setting I. The errors ϵ_{Li} and ϵ_{Ui} are normally distributed with mean 0 and variance σ^2.
3. Setting III (A nonparametric additive model):
   (i) For ith observation, we first generated X_{ij}^c from $U(0, 1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, 10$.

(ii) Then we generated Y_i^c using the following nonparametric additive model:

$$Y_i^c = \mu^c + \sum_{j=1}^{10} f_j^c(X_{ij}^c) + \epsilon_i^c. \tag{13}$$

Here, the intercept term μ^c and the smooth functions $f_j^c(x)$, $j = 1, \ldots, 10$, were given by

$$\begin{aligned} \mu^c = 0, \quad & f_1^c(x) = -6x, \\ f_2^c(x) = 20\sin(0.5\pi x), \quad & f_3^c(x) = -3\cos\{\pi(6x-5)/3\} \\ f_4^c(x) = (2-3x)^3/2, \quad & f_5^c(x) = \cdots = f_{10}^c(x) \equiv 0. \end{aligned}$$

The errors ϵ_i^c are normally distributed with mean 0 and variance σ^2.

(iii) For the ith observation, we randomly generated X_{ij}^r from $U(0, 1)$ for $j = 1, \ldots, 10$ and Y_i^r from $U(0, 1)$ for $i = 1, \ldots, n$.

In Setting I we generated the centers from uniform distributions with different intervals, respectively, such that the centers of the intervals get larger from 15 to 180. Thus, one can expect the generated centers are getting larger as well. Thus, we also generated ranges from uniform distribution with different intervals, respectively, so that the lengths of the ranges get larger from 1 to 11. On the other hand, in Setting III all ranges were generated from the same distribution $U(0, 1)$ because all centers were generated from the same distribution as well.

Nonparametric additive models can be used to analyze data generated from a linear regression model because f_js are arbitrary smooth functions. As mentioned in the previous section, the linear regression model is a special case of the nonparametric additive model. On the other hand, when data are generated from a nonparametric additive model such as (13), one should not use linear regression model (11) or (12) to analyze the data because the model cannot capture the nonlinearity of the data well. However, we can try to use a cubic polynomial regression model to fit to such data and hence, we used CRM with the following cubic polynomial center model, denoted CRM3, and compared it with other methods for Setting III:

$$Y_i^c = \beta_0 + \sum_{j=1}^{10}\{\beta_j X_{ij}^c + \beta_{j+10}(X_{ij}^c)^2 + \beta_{j+20}(X_{ij}^c)^3\} + \epsilon_i^c. \tag{14}$$

Similarly, we used SCM with higher order terms (including quadratic and cubic) in the design matrix, denoted SCM3, for Setting III.

We tested three sample sizes ($n = 100, 200$ and 500) and two error variances ($\sigma^2 = 9$ and 25) for each simulation setting. Using 100 simulation runs, we compared the performance of the three methods in terms of three criteria found in the literature: (i) the lower bound root mean squared error (RMSE$_L$), (ii) the upper bound root mean squared error (RMSE$_U$), and (iii) the symbolic correlation coefficient (r). Lima Neto and de Carvalho (2008) proposed the RMSE$_L$ and the RMSE$_U$ which measure the difference between the predicted intervals and the observed intervals. They are defined as follows:

$$\mathrm{RMSE}_L = \sqrt{\frac{\sum_{i=1}^{n}(Y_{Li} - \widehat{Y}_{Li})^2}{n}} \quad \text{and} \quad \mathrm{RMSE}_U = \sqrt{\frac{\sum_{i=1}^{n}(Y_{Ui} - \widehat{Y}_{Ui})^2}{n}},$$

where $[Y_{Li}, Y_{Ui}]$ and $[\widehat{Y}_{Li}, \widehat{Y}_{Ui}]$ are the observed and the predicted intervals of ith observation, respectively. Billard (2007, 2008) proposed the symbolic correlation coefficient which measures the correlation between the predicted intervals and the observed intervals. Then, Xu (2010) applied it to the regression problem. The symbolic correlation coefficient, r, is defined by

$$r(Y, \widehat{Y}) = \frac{\mathrm{Cov}(Y, \widehat{Y})}{\sqrt{S_Y^2 S_{\widehat{Y}}^2}},$$

where $\mathrm{Cov}(Y, \widehat{Y})$ is the symbolic sample covariance between Y and $\widehat{Y}$ in (7), and the S_Y^2 and $S_{\widehat{Y}}^2$ are the symbolic sample variances of Y and $\widehat{Y}$, respectively. The symbolic sample variance of Y, S_Y^2, is defined by

$$S_Y^2 = \frac{1}{3n}\sum_{i=1}^{n}(Y_{Li}^2 + Y_{Li}Y_{Ui} + Y_{Ui}^2) - \left[\frac{1}{2n}\sum_{i=1}^{n}(Y_{Li} + Y_{Ui})\right]^2.$$

In order to compute the above three criteria we generated testing sets with the same sample size ($n = 100, 200$ and 500) independently.

Table 1
RMSE$_L$, RMSE$_U$, the symbolic correlation coefficient (r) and their standard errors (in parentheses) from the simulation setting I (a linear regression model based on CRM) with $\sigma = 3, 5$ and $n = 100, 200, 500$.

σ_c	n		CRM		SCM		CRAM	
3	100	RMSE$_L$	1.0410	(0.0248)	1.2474	(0.0280)	1.5039	(0.0456)
		RMSE$_U$	1.0223	(0.0250)	1.2625	(0.0306)	1.5230	(0.0466)
		r	0.9900	(0.0004)	0.9835	(0.0006)	0.9735	(0.0016)
	200	RMSE$_L$	0.7072	(0.0138)	1.1187	(0.0188)	0.9822	(0.0234)
		RMSE$_U$	0.7018	(0.0153)	1.1262	(0.0159)	0.9833	(0.0232)
		r	0.9951	(0.0002)	0.9871	(0.0004)	0.9893	(0.0005)
	500	RMSE$_L$	0.4546	(0.0097)	0.9250	(0.0096)	0.5958	(0.0145)
		RMSE$_U$	0.4537	(0.0100)	0.9243	(0.0128)	0.5972	(0.0149)
		r	0.9980	(0.0001)	0.9909	(0.0002)	0.9959	(0.0002)
5	100	RMSE$_L$	1.7694	(0.0392)	1.9296	(0.0373)	2.5054	(0.0505)
		RMSE$_U$	1.7750	(0.0382)	1.9296	(0.0419)	2.5385	(0.0526)
		r	0.9691	(0.0015)	0.9626	(0.0015)	0.9296	(0.0033)
	200	RMSE$_L$	1.1815	(0.0278)	1.4146	(0.0278)	1.6104	(0.0447)
		RMSE$_U$	1.1837	(0.0297)	1.3926	(0.0290)	1.6095	(0.0458)
		r	0.9855	(0.0008)	0.9807	(0.0009)	0.9719	(0.0019)
	500	RMSE$_L$	0.7395	(0.0142)	1.0842	(0.0176)	0.9321	(0.0245)
		RMSE$_U$	0.7385	(0.0152)	1.1013	(0.0187)	0.9385	(0.0248)
		r	0.9943	(0.0002)	0.9871	(0.0004)	0.9896	(0.0006)

Table 2
RMSE$_L$, RMSE$_U$, the symbolic correlation coefficient (r) and their standard errors (in parentheses) from the simulation setting II (a linear regression model based on SCM) with $\sigma = 3, 5$ and $n = 100, 200, 500$.

σ_c	n		CRM		SCM		CRAM	
3	100	RMSE$_L$	1.5462	(0.0245)	1.3932	(0.0262)	1.8239	(0.0308)
		RMSE$_U$	1.5394	(0.0260)	1.3369	(0.0196)	1.8349	(0.0264)
		r	0.7263	(0.0091)	0.5968	(0.0095)	0.5714	(0.0139)
	200	RMSE$_L$	1.4414	(0.0156)	1.2501	(0.0140)	1.5848	(0.0188)
		RMSE$_U$	1.4062	(0.0152)	1.2140	(0.0134)	1.5454	(0.0179)
		r	0.7929	(0.0044)	0.6414	(0.0059)	0.7054	(0.0061)
	500	RMSE$_L$	1.3551	(0.0109)	1.1685	(0.0121)	1.3979	(0.0084)
		RMSE$_U$	1.3485	(0.0094)	1.1823	(0.0075)	1.3988	(0.0083)
		r	0.8140	(0.0033)	0.6643	(0.0031)	0.7825	(0.0035)
5	100	RMSE$_L$	2.7905	(0.0433)	2.1945	(0.0375)	3.1908	(0.0540)
		RMSE$_U$	2.7679	(0.0375)	2.1676	(0.0352)	3.2544	(0.0492)
		r	0.5708	(0.0145)	0.4500	(0.0148)	0.4451	(0.0145)
	200	RMSE$_L$	2.5589	(0.0345)	1.9806	(0.0245)	2.7414	(0.0382)
		RMSE$_U$	2.5918	(0.0310)	1.9933	(0.0305)	2.7933	(0.0335)
		r	0.6303	(0.0076)	0.4995	(0.0101)	0.5267	(0.0100)
	500	RMSE$_L$	2.4117	(0.0237)	1.7930	(0.0189)	2.4430	(0.0213)
		RMSE$_U$	2.4674	(0.0165)	1.8412	(0.0210)	2.5402	(0.0181)
		r	0.6763	(0.0050)	0.5464	(0.0056)	0.6228	(0.0065)

3.2. Results

We summarize the simulation results in Tables 1–3. Table 1 shows the values of the three criteria and their standard errors from the simulation of Setting I (a linear regression model based on CRM) with $\sigma = 3, 5$ and $n = 100, 200, 500$. When data were generated from a linear regression model, as expected, RMSE$_L$ and RMSE$_U$ based on CRM were smaller (and r was larger) than those based on SCM and CRAM. However, as n increased, the values of RMSE$_L$ and RMSE$_U$ for CRAM decreased rapidly while those for SCM decreased slowly. As a result, RMSE$_L$ and RMSE$_U$ for CRAM became smaller than those for SCM. For example, when $\sigma = 3$ and $n = 200$, the RMSE$_U$ for CRAM was 0.9833 while that for SCM was 1.1262 (14.5% larger). Overall, although CRM performed the best, CRAM outperformed SCM for some cases, especially for large samples with small standard deviation in this simulation setting.

Table 2 shows the values of the three criteria and their standard errors from the simulation of Setting II (a linear regression model based on SCM) with $\sigma = 3, 5$ and $n = 100, 200, 500$. The results are similar with those from Setting I. RMSE$_L$ and RMSE$_U$ based on CRM and SCM were smaller than those based on CRAM. Especially, we can see from the table that RMSE$_L$ and RMSE$_U$ based on SCM are consistently smaller than those based on the other methods in all cases. However, the difference got smaller as the sample size n increased. Interestingly, the symbolic correlation coefficient, r, for SCM was not higher than that for the other methods except when $n = 100$. CRAM outperformed SCM in terms of r in most cases.

Table 3 shows the three criteria and their standard errors from the simulation of Setting III (a nonparametric additive model) with $\sigma = 3, 5$ and $n = 100, 200, 500$. As expected, CRM3 and SCM did not perform well in terms of RMSE$_L$, RMSE$_U$ and the symbolic correlation coefficient (r) for data generated from a nonparametric additive model, while CRAM performed

Table 3
RMSE$_L$, RMSE$_U$, the symbolic correlation coefficient (r) and their standard errors (in parentheses) from the simulation setting III (a nonparametric additive model) with $\sigma = 3, 5$ and $n = 100, 200, 500$.

σ_c	n		CRM3		SCM		CRAM	
		RMSE$_L$	2.0391	(0.0272)	11.3007	(0.1140)	1.4977	(0.0274)
	100	RMSE$_U$	1.9901	(0.0279)	9.9525	(0.0671)	1.5497	(0.0301)
		r	0.9622	(0.0011)	−0.2142	(0.0085)	0.9786	(0.0010)
		RMSE$_L$	1.3694	(0.0186)	11.4418	(0.0925)	1.0544	(0.0226)
3	200	RMSE$_U$	1.3510	(0.0162)	9.9366	(0.0451)	1.0458	(0.0231)
		r	0.9823	(0.0005)	−0.2014	(0.0056)	0.9900	(0.0005)
		RMSE$_L$	0.8934	(0.0107)	11.4104	(0.0593)	0.6316	(0.0089)
	500	RMSE$_U$	0.8953	(0.0106)	9.8209	(0.0333)	0.6363	(0.0089)
		r	0.9925	(0.0002)	−0.2016	(0.0035)	0.9966	(0.0001)
		RMSE$_L$	3.3200	(0.0560)	11.8915	(0.1848)	2.2484	(0.0514)
	100	RMSE$_U$	3.3294	(0.0598)	10.1318	(0.0886)	2.2951	(0.0532)
		r	0.9021	(0.0035)	−0.2001	(0.0092)	0.9542	(0.0023)
		RMSE$_L$	2.1980	(0.0315)	11.5124	(0.0962)	1.7179	(0.0296)
5	200	RMSE$_U$	2.1833	(0.0330)	9.9949	(0.0521)	1.7094	(0.0297)
		r	0.9546	(0.0017)	−0.1974	(0.0057)	0.9726	(0.0011)
		RMSE$_L$	1.3163	(0.0163)	11.5607	(0.0743)	1.0659	(0.0265)
	500	RMSE$_U$	1.3205	(0.0146)	9.8279	(0.0373)	1.0722	(0.0260)
		r	0.9836	(0.0004)	−0.1997	(0.0035)	0.9899	(0.0005)

Table 4
The estimated regression coefficients of Hawaiian climate data using CRM, CRM3, SCM and CRAM2.

	CRM		CRM3		SCM	CRAM2
	$\widehat{\boldsymbol{\beta}}^c$	$\widehat{\boldsymbol{\beta}}^r$	$\widehat{\boldsymbol{\beta}}^c$	$\widehat{\boldsymbol{\beta}}^r$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\beta}}^c$
$\widehat{\beta}_0$	1030.94	0.7212	−7720.14	0.5609	1033.19	1013.64
$\widehat{\beta}_1$	−0.2331		350.777		−0.2531	
$\widehat{\beta}_2$	0.5224	0.0289	1.2134	0.1716	0.4216	0.5183
$\widehat{\beta}_3$			4.6906			
$\widehat{\beta}_4$			−0.0474	−0.0336		
$\widehat{\beta}_5$			0.0208			
$\widehat{\beta}_6$				0.0024		

better. Especially, SCM performed very poorly. In most cases, the values of RMSE$_L$ and RMSE$_U$ for SCM were very large compared to CRAM and the symbolic correlation coefficient for SCM was closed to zero while that for CRAM was almost one.

Our simulation study suggests that one should use the proposed method when data have a nonlinear pattern. It performed much better than the two existing methods for data generated from a nonparametric additive model. The gains in terms of RMSE$_L$ and RMSE$_U$ were substantial compared to SCM.

4. Application to real data

4.1. Interval-valued data

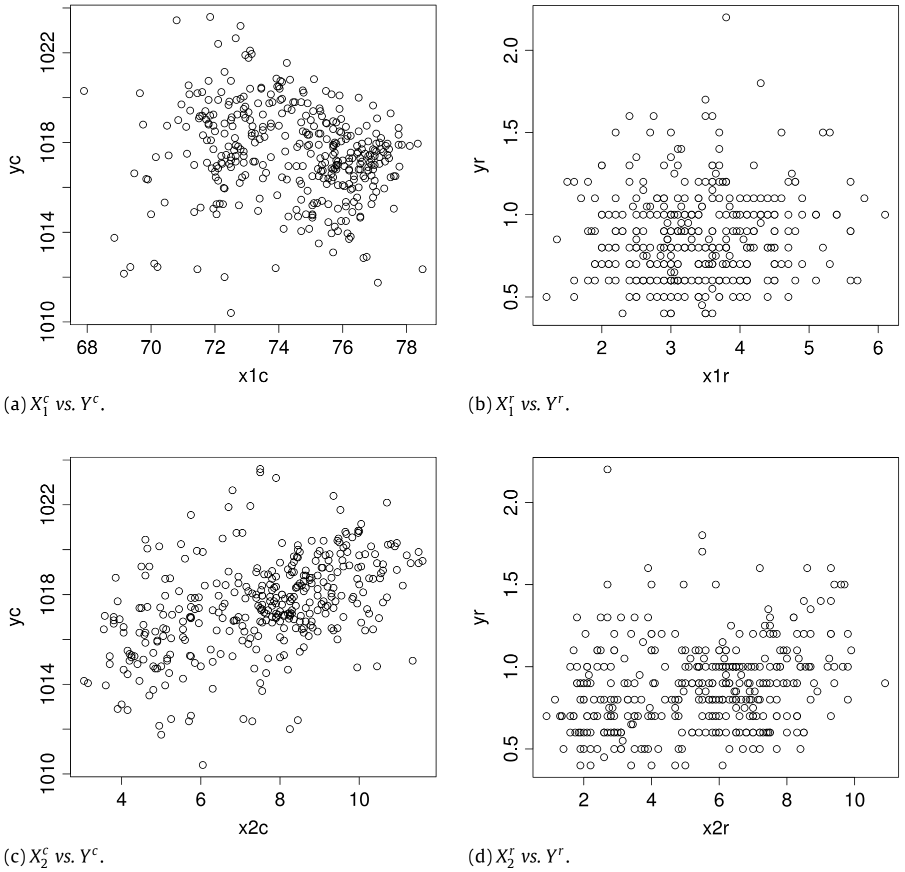
In this section we apply the proposed CRAM to a real interval-valued data set. We use CRM(3), SCM and CRAM to analyze the data and compare the results each other. To compare the performances of the proposed and the current methods, we calculate the three measures, RMSE$_L$, RMSE$_U$ and r using the leave-one-out cross-validation (CV). We consider the Hawaiian climate data set described in the introduction. We use models (1), (5) and (8) to describe the relationship between the daily sea level and other variables.

From Fig. 1 we can see nonlinear relations in both plots although it may be weak because of the large variability of the data. Fig. 2 shows the scatter plots of the center and the range of the response variable against those of the explanatory variables. Here we can also see weak nonlinear patterns between X_1^c and Y^c; X_2^r and Y^r. Thus, the linear models (1) or (5) may not be appropriate to describe the relationships.

When applying the proposed method to this data set, since the relationship between X_2^c and Y^c looks linear from Fig. 2(c), we consider two additive models for the center as follows:

(i) CRAM1: $Y_i^c = \mu^c + f_1^c(X_{i1}^c) + f_2^c(X_{i2}^c) + \epsilon_i^c$.
(ii) CRAM2: $Y_i^c = \mu^c + f_1^c(X_{i1}^c) + \beta_2^c X_{i2}^c + \epsilon_i^c$.

Table 4 shows the results of data analysis using the four methods. Note that CRAM1 is established for fully nonparametric additive models, and hence there is no result of the estimated regression coefficients for CRAM1 in Table 4. Instead, one can



(a) X_1^c *vs.* Y^c.

(b) X_1^r *vs.* Y^r.

(c) X_2^c *vs.* Y^c.

(d) X_2^r *vs.* Y^r.

Fig. 2. Scatter plots of the center and the range of the response variable against those of the explanatory variables for Hawaiian climate data.

Table 5
RMSE$_L$, RMSE$_U$ and r for CRM, CRM3, SCM, CRAM1 and CRAM2 of Hawaiian climate data.

	CRM	CRM3	SCM	CRAM1	CRAM2
RMSE$_L$	1.8223	1.8041	1.9631	1.7676	1.7712
RMSE$_U$	1.7700	1.7604	1.7890	1.7164	1.7204
r	0.5189	0.5337	0.4910	0.5594	0.5525

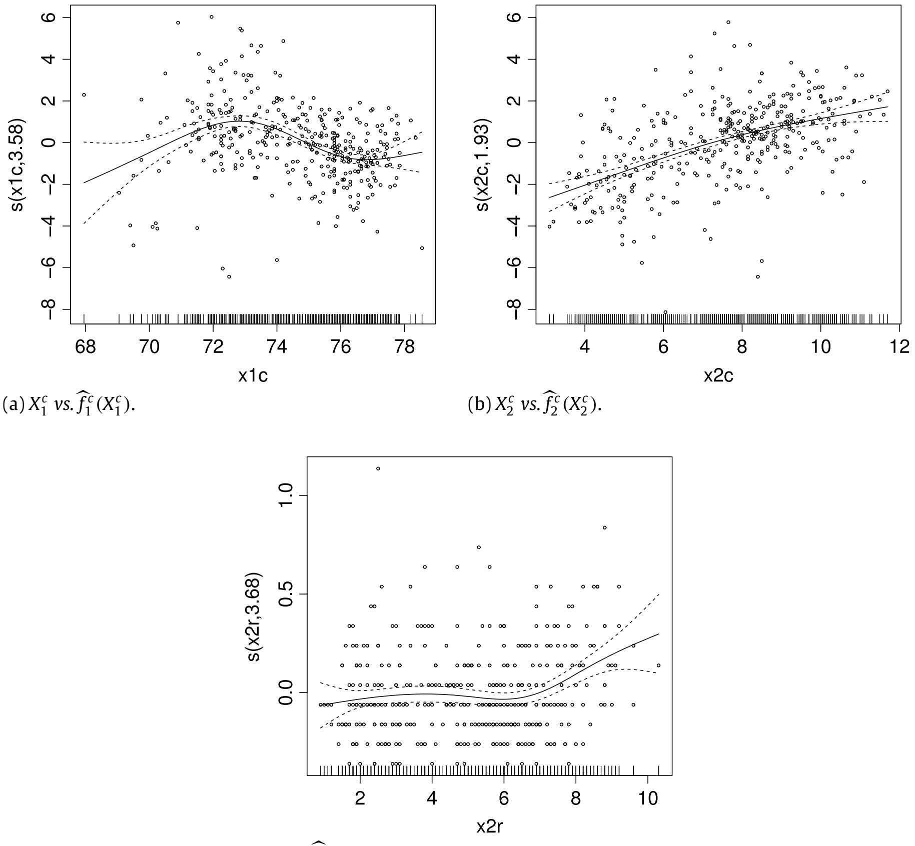
see plots of the fitted component smooth functions against the center and the range of the explanatory variables in Fig. 3. The values of the estimated parameters using CRM (the center model) and SCM are similar to each other. The center model for CRAM2 contains a regression coefficient β_2^c whose estimated value is also similar with those using CRM and SCM. The estimated intercept parameters $\widehat{\beta_0^c}$ (CRM), $\widehat{\beta_0}$ (SCM) and $\widehat{\mu}^c$ (CRAM2) have similar values as well. For the range model of the CRM, the estimated coefficient of X_1^r was not significant and the corresponding term was removed from the model. For CRM3, since we use cubic polynomial regression models for the center and the range models, there are 7 parameters in the model. However, insignificant parameters were removed from the models.

From Fig. 3 we can see that there may be nonlinear relationships between the center of the response variable, Y^c and X_1^c, X_2^c, although the curvature is somewhat weak for X_2^c. For the range model, smooth terms for X_1^r were not significant, while we can see a strong curvature in the smooth term for X_2^r.

Table 5 shows the three measures, RMSE$_L$, RMSE$_U$ and r, for the five methods. From the table, we can see that CRAM1 performs the best while SCM performs the worst. Since CRM3 uses cubic polynomial regression models for the center and the range models, it performs slightly better. The center model for CRAM2 includes a parametric term for X_2^c since its relationship with Y^c looks linear. It performs better than other methods and similarly well compared with CRAM1.

4.2. Comparison with original data

In this section we consider the original data and compare the results of analyzing the original data with those from fitting the interval-valued data. We first analyze the original data using the standard linear regression model and the nonparametric

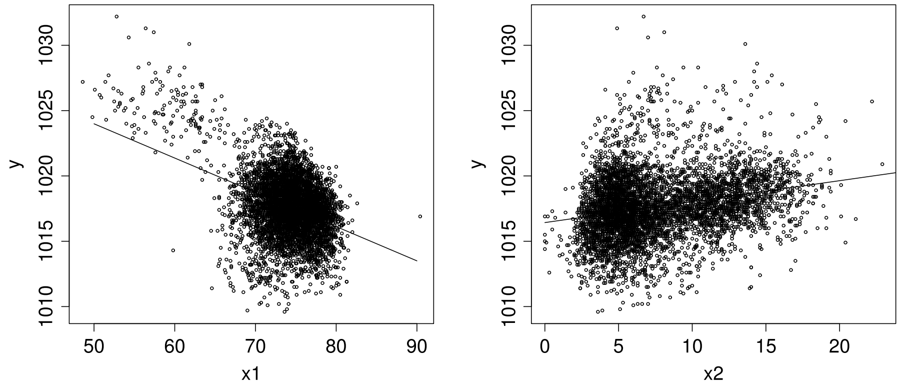


(a) X_1^c vs. $\widehat{f_1^c}(X_1^c)$.

(b) X_2^c vs. $\widehat{f_2^c}(X_2^c)$.

(c) X_2^r vs. $\widehat{f_2^c}(X_2^r)$.

Fig. 3. Plots of the fitted component smooth functions (solid) against the center and the range of the explanatory variables with their corresponding confidence bands (dashed) and partial residuals (circle) for Hawaiian climate data.



(a) X_1 vs. Y.

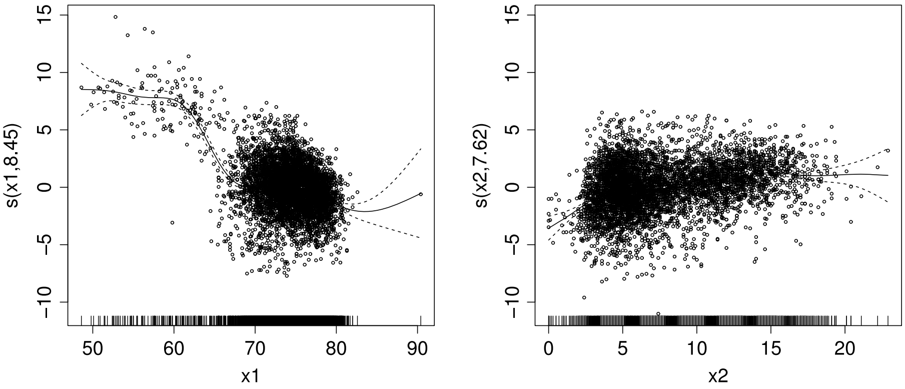
(b) X_2 vs. Y.

Fig. 4. Scatter plots of the original Hawaiian climate data and the corresponding fitted lines based on simple linear regression.

additive model and compare the results each other. We calculate RMSE and r using the leave-one-out CV to compare the performances.

From Fig. 4 we can see again weak nonlinear relations in both plots. We superimposed the fitted lines on the scatter plots based on simple linear regression, from which we can see using the linear model may not be appropriate since it underestimated the data in the lower part of X_1 and overestimated the data in the lower part of X_2 because of the large amount of variability of the data as well as the denseness in a certain region of the data.

Table 6 shows the results of data analysis using the linear model. The values of the estimated parameters using the linear model are overall similar with those using CRM (the center model) and SCM based on the interval-valued data. However, the estimate of the regression coefficient for X_2 using the original data is much smaller than those using the interval-valued



(a) X_1 *vs.* $\widehat{f}_1(X_1)$.

(b) X_2 *vs.* $\widehat{f}_2(X_2)$.

Fig. 5. Plots of the fitted component smooth functions (solid) against the explanatory variables with their corresponding confidence bands (dashed) and partial residuals (circle) for the original Hawaiian climate data.

Table 6
The estimated regression coefficients of the original Hawaiian climate data using linear regression.

$\widehat{\beta}_0$	1036.43
$\widehat{\beta}_1$	−0.2712
$\widehat{\beta}_2$	0.1743

Table 7
RMSE and r for the linear model and the additive model of the original Hawaiian climate data.

	Linear model	Additive model
RMSE	2.1567	2.0508
r	0.4864	0.5565

data. Fig. 5 shows plots of the fitted component smooth functions against the explanatory variables using the additive model. We can see from the figure that there may be nonlinear relationships between Y and X_1, X_2. Table 7 shows RMSE and r for the linear model and the additive model. From the table we can see that the additive model performs better than the linear model.

The results from fitting the original Hawaiian climate data are different from those from fitting the interval-valued data. Since the interval-valued Hawaiian climate data were converted from the original single-valued observations, comparing the results based on the two different types of data may help us to understand where the difference comes from. As mentioned earlier, the main difference is from the estimated value of the regression coefficient for X_2. The estimate of the regression coefficient for X_2 using the original data is smaller than those using the interval-valued data while the estimated coefficients for X_1 are similar to each other. One possible reason for this difference may be that the data are distributed in the larger region of X_2 than X_1. Because of the reason the data are denser in the range of X_1 except the lower part than X_2. Consequently, the intervals for X_1 of the interval-valued data may contain the information of the data more uniformly than those for X_2, and hence the estimated coefficients for X_1 are similar to each other while those for X_2 are not. This difference may imply that analyzing interval-valued data has a limitation because we cannot correctly restructure the variability of the original data using the interval-valued data.

In order to compare the performance of the methods we computed RMSE_L, RMSE_U and the symbolic correlation coefficient, r, for the interval-valued data while RMSE and the correlation coefficient, r, between the predicted value and the observed value for the original data. Although their definitions are not the same, we may still use those corresponding criteria to compare the results from fitting the interval-valued data with those from the original data in the sense that a point is a special case of intervals where the lower and upper bounds are equal. From Tables 5 and 7 we can see that RMSE_L and RMSE_U for CRAM1 are smaller than RMSE for the additive model while the values of r are similar to each other. The difference between the results may come from the fact that the size and the variability of the original data are larger than those of the interval-valued data.

However, the smoothing parameter selection using the GCV may play a role in the discrepancy between the results. There are some issues with CV for grouped/discretized data in kernel density estimation. Jang and Loh (2010) discussed them and proposed a method using a combined cross-validation to select the optimal bandwidth in kernel density estimation with grouped data. In our analysis the value of GCV is 2.9293 for the center model of the interval-valued data while 4.203 for the original data. The selected smoothing parameters based on those values of GCV are 0.00294 and 0.07036 for X_1^c and X_2^c, respectively for the center model of the interval-valued data while 0.00039 and 0.00091 for X_1 and X_2, respectively for the original data. To check whether the smoothing parameter selection plays any role we replaced the smoothing parameters for the interval-valued data by those for the original data and fitted the CRAM again. The resulting values of RMSE_L, RMSE_U and r are 1.7674, 1.7180 and 0.5556, respectively, which are very similar to the original values. Therefore, we can see that the difference between the results from fitting the original and the interval-valued data may not come from the smoothing parameter selection using the GCV for the nonparametric additive models.

5. Concluding remarks

In this paper, we proposed the center and range additive method (CRAM) to analyze interval-valued data. When scatter plots of data show nonlinear patterns, current methods such as the center and range method (CRM) may not be appropriate because they are designed to fit a linear regression model on interval-valued data. The proposed method introduces a nonparametric additive regression model, and fits two separate nonparametric additive models to the center point and the range of the intervals, respectively. We demonstrated its utility using simulation studies and a real data example from the literature.

There is an issue to be addressed for the proposed method. It does not ensure that the estimated ranges are positive. Although one can take the absolute value of an estimated range to establish predicted intervals, it would be more appropriate if there is a method to obtain a positive estimator of the range. We would address this issue using constrained estimation methods in the near future.

The proposed method depends on the model used for describing the relationship between the variables in data. Depending on underlying data structure, one can easily extend the proposed method to semiparametric additive models. Of course, one could consider other models such as nonlinear regression models, nonparametric or semiparametric additive models.

One major issue of interval-valued data is that there is no or little information about internal variation of the data. Due to this lack of information, one cannot carry out statistical inference on regression coefficients such as confidence interval and hypothesis testing. Although we did hypothesis testing and determined the best model in the real data examples, it was done separately based on the center and range models and hence it might not be a very reliable decision. In order to address this issue, Ahn et al. (2012) proposed a resampling approach for interval-valued data regression. One could employ their proposed method to ours and carry out a proper statistical inference.

Acknowledgment

This research was supported by the Chung-Ang University Research Grants in 2014. We thank the editor, the associate editor, and the two anonymous referees for several helpful comments that led to substantial improvement in the content and presentation of the article.

References

Ahn, J., Peng, M., Park, C., & Jeon, Y. (2012). A resampling approach for interval-valued data regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*, 336–348.

Barnett, T. P. (1985). Variations in near-global sea level pressure. *Journal of the Atmospheric Sciences*, *42*, 478–501.

Bertrand, P., & Goupil, F. (2000). Descriptive statistics for symbolic data. In H.-H. Bock, & E. Diday (Eds.), *Analysis of symbolic data* (pp. 103–124). Berlin: Springer-Verlag.

Billard, L. (2007). Dependencies and variation components of symbolic interval-valued data. In P. Brito, G. Cucumel, P. Bertrand, & F. de Carvalho (Eds.), *Selected contributions in data analysis and classification* (pp. 3–13). Berlin: Springer-Verlag.

Billard, L. (2008). Sample covariance functions for complex quantitative data. In *World congress*. Yokohama, Japan: International Association of Computational Statistics.

Billard, L., & Diday, E. (2000). Regression analysis for interval-valued data. In H. A. L. Kiers, J.-P. Rassoon, P. J. F. Groenen, & M. Schader (Eds.), *Data analysis, classification, and related methods* (pp. 369–374). Berlin: Springer-Verlag.

Billard, L., & Diday, E. (2007). *Symbolic data analysis: conceptual statistics and data mining*. Chichester: Wiley.

Blanco-Fernandez, A., Colubi, A., & Gonzalez-Rodriguez, G. (2013). Linear regression analysis for interval-valued data based on set arithmetic: A review. In C. Borgelt, M. A. Gil, J. M. C. Sousa, & M. Verleysen (Eds.), *Studies in fuzziness and soft computing*: *Vol. 285. Towards advanced data analysis by combining soft computing and statistics* (pp. 19–31). Berlin: Springer-Verlag.

Blanco-Fernandez, A., Corral, N., & Gonzalez-Rodriguez, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics & Data Analysis*, *55*, 2568–2578.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, *80*, 580–619.

Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *Annals of Statistics*, *17*, 453–555.

Carroll, R. J., Maity, A., Mammen, E., & Yu, K. (2009). Nonparametric additive regression for repeatedly measured data. *Biometrika*, *96*, 383–398.

Curtis, S. M., Banerjee, S., & Ghosal, S. (2014). Fast Bayesian model assessment for nonparametric additive regression. *Computational Statistics & Data Analysis*, *71*, 347–358.

Davis, R. E. (1976). Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *Journal of Physical Oceanography*, *6*, 249–266.

Diday, E. (1987). The symbolic approach in clustering and related methods of data analysis. In H.-H. Bock (Ed.), *Classification and related methods of data analysis*. Amsterdam: North-Holland.

Diday, E. (1995). Probabilist, possibilist and belief object for knowledge analysis. *Annals of Operations Research*, *55*, 227–276.

Diday, E., & Emilion, R. (1996). Lattices and capacities in analysis of probabilist object. In E. Diday, Y. Lechevallier, & O. Opilz (Eds.), *Studies in classification* (pp. 13–30).

Diday, E., & Emilion, R. (1998). Capacities and credibilities in analysis of probabilistic objects by histograms and lattices. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, & Y. Baba (Eds.), *Data science, classification, and related methods* (pp. 353–357).

Diday, E., Emilion, R., & Hillali, Y. (1996). Symbolic data analysis of probabilist objects by capacities and credibilities. XXXVIII Societa Italiana Di Statistica. Rimini, Italy.

Friedman, J. H., & Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association*, *76*, 817–823.

Gillett, N. P., Zwiers, F. W., Weaver, A. J., & Stott, P. A. (2003). Detection of human influence on sea-level pressure. *Nature*, *422*, 292–294.

Hastie, T. J., & Tibshirani, R. J. (1984). *Generalized additive models. Technical report*. Division of Biostatistics, Stanford University.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.

Horowitz, J. L. (2014). Nonparametric additive models. In *The oxford handbook of applied nonparametric and semiparametric econometrics and statistics* (pp. 129–148).

Horowitz, J. L., & Mammen, E. (2011). Oracle-efficient nonparametric estimation of an additive model with an unknown link function. *Econometric Theory*, *27*, 582–608.

Iwasaki, M., & Tsubaki, H. (2005). A bivariate generalized linear model with an application to meteorological data analysis. *Statistical Methodology*, *2*, 175–190.

Jang, W., & Loh, J. M. (2010). Density estimation for grouped data with application to line transect sampling. *The Annals of Applied Statistics*, *4*, 893–915.

Kutzbach, J. E. (1967). Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America. *Journal of Applied Meteorology*, *6*, 791–802.

Lima Neto, E. A., Cordeiro, G., & de Carvalho, F. (2011). Bivariate symbolic regression models for interval-valued variables. *Journal of Statistical Computation and Simulation*, *81*, 1727–1744.

Lima Neto, E.A., Cordeiro, G.M., Carvalho, F.A.T., Anjos, U., & Costa, A. (2009). Bivariate generalized linear model for interval-valued variables. In *Proceedings 2009 IEEE international joint conference on neural networks, Vol. 1* (pp. 2226–2229). Atlanta, USA.

Lima Neto, E. A., & de Carvalho, F. A. T. (2008). Center and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, *52*, 1500–1515.

Lima Neto, E. A., & de Carvalho, F. A. T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, *54*, 333–347.

Lima Neto, E. A., de Carvalho, F. A. T., & Tenorio, C. P. (2004). Univariate and multivariate linear regression methods to predict interval-valued features. In *Lecture notes in computer science*, *AI 2004 advances in artificial intelligence* (pp. 526–537). Berlin: Springer-Verlag.

Linton, O. B., & Härdle, W. (1996). Estimating additive regression models with known links. *Biometrika*, *83*, 529–540.

Lutgens, F. K., & Tarbuck, E. J. (2007). *The atmosphere: an introduction to meteorology*. New Jersey: Prentice Hall.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., & Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, *23*, 249–269.

Min, S. K., Legutke, S., Hense, A., & Kwon, W. T. (2005). Internal variability in a 1000-yr control simulation with the coupled climate model ECHO-G—I. Near-surface temperature, precipitation and mean sea level pressure. *Tellus A*, *57*, 605–621.

Silva, A., Lima Neto, E.A., & Anjos, U. (2011). A regression model to interval-valued variables based on copula approach. In *Proceedings of the 58th world statistics congress of the international statistical institute*. Dublin, Ireland.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, *13*, 689–705.

van der Burg, E., & de Leeuw, J. (1983). Non-linear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, *36*, 54–80.

Wadsworth, G. P. (1951). Application of statistical methods to weather forecasting. In T. F. Malone (Ed.), *Compendium of meteorology* (pp. 849–855). Boston: American Meteorological Society.

Wadsworth, G. P., Bryan, J. G., & Gordon, C. H. (1948). *Short range and extended forecasting by statistical methods. Air. Wea. Serv. Tech. Rep. (105-37)*. (p. 202).

Wong, R. K. W., Yao, F., & Lee, T. C. M. (2014). Robust estimation for generalized additive models. *Journal of Computational and Graphical Statistics*, *23*, 270–289.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society: Series B*, *62*, 413–428.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*, 673–686.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. Boca Raton: CRC Press.

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

Xu, W. (2010). *Symbolic data analysis: interval-valued data regression*. (PhD thesis), University of Georgia.

Yang, C.-Y., Jeng, J.-T., Chuang, C.-C., & Tao, C. (2011). Constructing the linear regression models for the symbolic interval-values data using PSO algorithm. In *2011 international conference on system science and engineering (ICSSE)* (pp. 177–181). IEEE.

Yang, L., Sperlich, S., & Härdle, W. (2003). Derivative estimation and testing in generalized additive models. *Journal of Statistical Planning and Inference*, *115*, 521–542.

Yu, K., Park, B. U., & Mammen, E. (2008). Smooth backfitting in generalized additive models. *Annals of Statistics*, *36*, 228–260.