

Communications in Statistics - Simulation and Computation

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

Functional linear models for interval-valued data

Ufuk Beyaztas, Han Lin Shang & Abdel-Salam G. Abdel-Salam

To cite this article: Ufuk Beyaztas, Han Lin Shang & Abdel-Salam G. Abdel-Salam (2020): Functional linear models for interval-valued data, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2020.1714662

To link to this article: https://doi.org/10.1080/03610918.2020.1714662

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC

Published online: 25 Jan 2020.

Submit your article to this journal

Article views: 602

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=lssp20

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2020.1714662

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Functional linear models for interval-valued data

Ufuk Beyaztas[a], Han Lin Shang[b], and Abdel-Salam G. Abdel-Salam[c]

[a]Department of Statistics, Bartin University, Bartın, Turkey; [b]Department of Econometrics and Business Statistics, Research School of Finance Actuarial Studies and Statistics Australian National University, Canberra, Australia; [c]Department of Mathematics, Statistics and Physics, Qatar University, Doha, Qatar

ABSTRACT

Aggregation of large databases in a specific format is a frequently used process to make the data easily manageable. Interval-valued data is one of the data types that is generated by such an aggregation process. Using traditional methods to analyze interval-valued data results in loss of information, and thus, several interval-valued data models have been proposed to gather reliable information from such data types. On the other hand, recent technological developments have led to high dimensional and complex data in many application areas, which may not be analyzed by traditional techniques. Functional data analysis is one of the most commonly used techniques to analyze such complex datasets. While the functional extensions of much traditional statistical techniques are available, the functional form of the interval-valued data has not been studied well. This article introduces the functional forms of some well-known regression models that take interval-valued data. The proposed methods are based on the function-on-function regression model, where both the response and predictor/s are functional. Through several Monte Carlo simulations and empirical data analysis, the finite sample performance of the proposed methods is evaluated and compared with the state-of-the-art.

ARTICLE HISTORY
Received 22 July 2019
Accepted 7 January 2020

KEYWORDS
Functional data; interval-valued data; maximum likelihood; regression

1. Introduction

Due to recent technological advances, the process of collecting data has become complicated, causing high dimensional and complex data structures. Symbolic data analysis is one of the commonly used methods in modeling such complex and large datasets, see Billard (2011) and Noirhomme-Fraiture and Brito (2011) for recent developments in symbolic data analysis. Contrary to single-valued observations in p-dimensional space where classical statistical methods work on, symbolic data may be in the form of hypercubes in p-dimensional space. There are many symbolic data types, for example, list, histogram, modal-valued, and interval-valued data. In this research, we restrict our attention to the interval-valued data only. The data expressed in an interval format (minimum and maximum values of the data) is called the interval-valued data. Such

CONTACT Abdel-Salam G. Abdel-Salam Abdo@qu.edu.qa Department of Mathematics, Statistics and Physics, Qatar University, Doha, Qatar.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

datasets are frequently encountered in daily life, for example, air and/or surface temperature, wind speed, energy production, blood pressure, and exchange rates. The main problem encountered during the modeling of the interval-valued data with classical statistical techniques is "how the variability of observations within the range is involved in modeling?". Traditional methods analyze interval-valued data using its summary (i.e., mid-points), and this approach results in loss of information. Therefore, interval-valued data analysis techniques are needed to obtain more accurate information.

The early studies about the interval-valued data regression were conducted by Billard and Diday (2000), who extended traditional statistical techniques to the interval-valued data. Billard and Diday (2002) extended several classical regression models to interval-valued data and they proposed a regression equation for fitting histogram-valued symbolic data; Neto, de Carvalho, and Tenorio (2004) proposed two interval-valued regression models using the mid-points and ranges of the interval values; Alfonso, Billard, and Diday (2004) suggested a regression model for fitting taxonomic variables; Billard and Diday (2007) suggested several interval-valued regression models; Billard (2007) proposed a covariance function for the interval-valued data; Maia and Carvalho (2008) introduced an interval-valued regression model based on least absolute deviation; Neto and de Carvalho (2008) suggested an interval-valued data regression model inspired by the work of Neto, de Carvalho, and Tenorio (2004) and Neto et al. (2009) proposed a bivariate generalized linear model for interval-valued symbolic data; Neto and de Carvalho (2010) presented a constrained interval-valued linear regression model; Neto, Cordeiro, and de Carvalho (2011) introduced a bivariate symbolic regression model that considers the interval-valued variables as bivariate random vectors; Ahn et al. (2012) proposed a resampling based interval-valued regression model; Lim (2016) suggested a nonparametric additive approach for analyzing interval-valued data that allows a nonlinear pattern; and Neto and de Carvalho (2017) introduced a nonlinear regression model for interval-valued data and estimated the model parameters by several optimization algorithms.

The datasets repeatedly measured over discrete time points provide more information than those obtained from a single time point, where most of the interval-valued data regression studies have focused on in the pertinent literature. Available regression models may not be able to model such data type due to some common regression problems, such as high dimensionality, multicollinearity, and the high correlation between the sequential observations. On the other hand, functional data, which considers the data in the form of curves, can characterize this kind of data that is sampled over continuum measures. Functional Data Analysis (FDA) techniques reduce the problem of high dimensionality and focus on temporal dependence between the curves. It provides several advantages over the classical methods: for example, it is possible to look at the data as a whole; FDA is not affected by the missing data and the high correlation problem between the repeated measurements and minimizes the data noise by smoothing techniques. Consult Ramsay and Silverman (2002, 2005), Ferraty and Vieu (2006), and Kokoszka and Reimherr (2017) for more information about the FDA.

In this paper, we extend some well-known interval-valued data regression models to the functional data context. Functional regression models have become important analytical tools to explore the relationship between the response and predictor variables. In

this context, several regression models have been proposed depending on whether response and/or predictor/s are scalar or functional; (i) functional response-scalar predictors; (ii) scalar response-functional predictors; and (iii) functional response-functional predictors. For the cases (i) and (ii), examples include Cardot, Ferraty, and Sarda (1999, 2003), James (2002), Hu, Wang, and Carroll (2004), Müller (2005), Amato, Antoniadis, and De Feis (2006), Hall and Horowitz (2007), Ferraty and Vieu (2009), Cook, Forzani, and Yao (2010), Chen, Hall, and Müller (2011), Dou, Pollard, and Zhou (2012), Febrero-Bande and Gonzalez-Manteiga (2013), and Goia and Vieu (2015). For the case (iii), see Ramsay and Dalzell (1991), Fan and Zhang (1999), Şentürk and Müller (2005, 2008), Yao, Müller, and Wang (2005), Harezlak et al. (2007), Matsui, Kawano, and Konishi (2009), Valderrama et al. (2010), Müller and Stadtmuller (2005), Jiang and Wang (2011), Ivanescu et al. (2015), Chiou, Yang, and Chen (2016) and Zhang et al. (2018). Also, Müller and Stadtmuller (2005), Horvath and Kokoszka (2012), and Cuevas (2014) present an excellent overview of research on functional regression models and their applications. In this study, we consider the case (iii), which is called function-on-function regression. As a summary, the interval-valued functional regression models proposed in this study work as follows. First, the discretely observed interval-valued data are converted to functional form using a B-spline basis, and the function-on-function regression model is used to investigate the relationships between the intervals of the response and predictors. The parameter surfaces are estimated using the maximum likelihood (ML) method, and finally, the lower and upper limit functions of the response variable are obtained using the smoothing step. Throughout this study, the primary attention is paid for the prediction performance of the proposed methods. The finite sample performance of the proposed methods is evaluated numerically via Monte Carlo simulations and an empirical data example.

The remainder of this article is organized as follows. Section 2 reviews the classical interval-valued data regression models considered in this study and discusses their functional data extensions. In Sec. 3.1, Monte Carlo simulations are conducted to examine the finite sample performance of the proposed methods. An empirical data is analyzed, and the results are reported in Sec. 3.2. Section 4 concludes the article.

2. Methodology

For $i = 1, ..., n$ and $j = 1, ..., p$, consider the following linear regression model with p predictor variables $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_p)^\top$; $\boldsymbol{X}_{ij} = (1, X_{i1}, ..., X_{ip})^\top$ and a response variable $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top$, which are the realizations of intervals so that $X_{ij} = [X_{ij}^l, X_{ij}^u]$; $X_{ij}^l \leq X_{ij}^u$ and $Y_i = [Y_i^l, Y_i^u]$; $Y_i^l \leq Y_i^u$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\top$ is a $(p+1) \times 1$ vector of unknown parameters and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top$ denotes the error vector of dimension $n \times 1$. The first extension of the regression model given in (2.1) to the interval-valued data was suggested by Billard and Diday (2000). Their extension, called center method (CM), uses the mid-points of the intervals to fit a regression model as follows:

$$\boldsymbol{Y}^c = \boldsymbol{X}^c \beta^c + \epsilon^c, \tag{2.2}$$

where $Y_i^c = (Y_i^l + Y_i^u)/2$ and $X_{ij}^c = (X_{ij}^l + X_{ij}^u)/2$, for $i = 1, ..., n$ and $j = 1, ..., p$. The model parameters $\boldsymbol{\beta}^c$ are estimated using least squares (LS) method. Let $\widehat{\boldsymbol{\beta}}^c = (\widehat{\boldsymbol{\beta}}_0^c, \widehat{\boldsymbol{\beta}}_1^c, ..., \widehat{\boldsymbol{\beta}}_p^c)^\top$ denote the estimate of $\boldsymbol{\beta}^c$, then the lower and upper limits of the interval $\boldsymbol{Y} = [\boldsymbol{Y}^l, \boldsymbol{Y}^u]$ are predicted as follows:

$$\widehat{\boldsymbol{Y}}^l = \boldsymbol{X}^l \widehat{\boldsymbol{\beta}}^c, \quad \widehat{\boldsymbol{Y}}^u = \boldsymbol{X}^u \widehat{\boldsymbol{\beta}}^c.$$

However, the CM does not take into account the internal variations of the intervals when estimating the model parameters, see Neto and de Carvalho (2008) and Ahn et al. (2012).

To overcome this problem, Neto, de Carvalho, and Tenorio (2004) suggested the center-range method (CRM) by fitting two distinct regression models using the mid-points (as in (2.2)) and half-ranges of the intervals. Let $Y_i^r = (Y_i^u - Y_i^l)/2$ and $X_{ij}^r = (X_{ij}^u - X_{ij}^l)/2$, for $i = 1, ..., n$ and $j = 1, ..., p$, denote the half-ranges of the intervals of $(X_1, ..., X_p)$ and Y, respectively. Then, the regression equation of the half-ranges is given as follows:

$$\boldsymbol{Y}^r = \boldsymbol{X}^r \boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r, \tag{2.3}$$

where $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, ..., \beta_p^r)^\top$ and $\boldsymbol{\epsilon}^r = (\epsilon_1^r, ..., \epsilon_n^r)^\top$. The model parameters are estimated by the LS method. Let $\widehat{\boldsymbol{Y}}^r = \boldsymbol{X}^r \widehat{\boldsymbol{\beta}}^r$ denote the predicted half-ranges of the intervals. Then, the predictions of the lower and upper limits of the response variable are obtained as follows:

$$\widehat{\boldsymbol{Y}}^l = \widehat{\boldsymbol{Y}}^c - \widehat{\boldsymbol{Y}}^r, \quad \widehat{\boldsymbol{Y}}^u = \widehat{\boldsymbol{Y}}^c + \widehat{\boldsymbol{Y}}^r.$$

As pointed out by Ahn et al. (2012), the CRM assumes that mid-points and half-ranges are independent, which may not hold in general.

Billard and Diday (2007) proposed bivariate center and range method (BRCM) to take into account the effects of intervals widths. It constructs two distinct regression models using mid-points and half-ranges of the intervals. However, unlike the CRM, both the mid-points and half-ranges are used as predictors in the regression models. Let $X_i^{cr} = (1, X_{i1}^c, ..., X_{ip}^c, X_{i1}^r, ..., X_{ip}^r)^\top$. Then, the regression equations for the BCRM are given as follows:

$$\boldsymbol{Y}^c = \boldsymbol{X}^{cr} \boldsymbol{\beta}^c + \boldsymbol{\epsilon}^c, \quad \boldsymbol{Y}^r = \boldsymbol{X}^{cr} \boldsymbol{\beta}^r + \boldsymbol{\epsilon}^r, \tag{2.4}$$

where $\boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, ..., \beta_{2p}^c)^\top$ and $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, ..., \beta_{2p}^r)^\top$ denote the parameter vectors. As in CRM, the regression parameters are estimated using the LS method, accordingly the lower and upper limits of the response variable are predicted as follows:

$$\widehat{\boldsymbol{Y}}^l = \widehat{\boldsymbol{Y}}^c - \widehat{\boldsymbol{Y}}^r, \quad \widehat{\boldsymbol{Y}}^u = \widehat{\boldsymbol{Y}}^c + \widehat{\boldsymbol{Y}}^r.$$

All three regression models (CM, CRM, and BRCM), are not appropriate for statistical inference, such as coefficient and model significance tests. Ahn et al. (2012) proposed a resampling based interval-valued data regression model (MCM), which enables making

inferences on the model. Let B denote the number of Monte Carlo simulations. Then, the MCM works as follows:

Step 1. Generate single-valued predictor/s X_{ij}^* and response Y_i^* variables uniformly from the intervals $X_{ij} = [X_{ij}^l, X_{ij}^u]$ and $Y_i = [Y_i^l, Y_i^u]$, respectively, to get a random vector $(Y_i^*, X_{i1}^*, ..., X_{ip}^*)^\top$, for $i = 1, ..., n$ and $j = 1, ..., p$.

Step 2. Construct a linear regression using the generated single-valued observations $\boldsymbol{Y}^* = (Y_1^*, ..., Y_n^*)$ and $\boldsymbol{X}_j^* = (X_{1j}^*, ..., X_{nj}^*)$ in Step 1 as follows:

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^*.$$

Step 3. Estimate the regression coefficients $\widehat{\beta}^* = (\widehat{\beta}_0^*, \widehat{\beta}_1^*, ..., \widehat{\beta}_p^*)$ using the LS method:

$$\widehat{\boldsymbol{\beta}}^* = \left((\boldsymbol{X}^*)^\top \boldsymbol{X}^*\right)^{-1} (\boldsymbol{X}^*)^\top \boldsymbol{Y}^*.$$

Step 4. Repeat Steps 1–3 B times to obtain B sets of estimates, $\widehat{\boldsymbol{\beta}}^* = (\widehat{\boldsymbol{\beta}}^{*1}, ..., \widehat{\boldsymbol{\beta}}^{*B})$.

Let $\overline{\widehat{\beta}}^* = B^{-1} \sum_{b=1}^{B} \widehat{\beta}^{*b}$ be the final estimate. Then, the lower and upper values of the response variable are predicted as follows:

$$\widehat{\boldsymbol{Y}}^l = \overline{\widehat{\boldsymbol{\beta}}}^* \boldsymbol{X}^l, \quad \widehat{\boldsymbol{Y}}^u = \overline{\widehat{\boldsymbol{\beta}}}^* \boldsymbol{X}^u.$$

Note that, for CM and MCM, the lower bound $\widehat{\boldsymbol{Y}}^l$ may produce higher values than the upper bound $\widehat{\boldsymbol{Y}}^u$. Thus, the authors suggested to use $\widehat{\boldsymbol{Y}}^l = \min(\widehat{\boldsymbol{Y}}^l, \widehat{\boldsymbol{Y}}^u)$ and $\widehat{\boldsymbol{Y}}^u = \max(\widehat{\boldsymbol{Y}}^l, \widehat{\boldsymbol{Y}}^u)$ to obtain logical predictions.

2.1. Functional response model

A functional data $\{\mathcal{X}_i(t) : i = 1, ..., N, \ t \in \mathcal{T}\}$ comprises random functions which are sample elements recorded at discrete times $t = \{t_1, ..., t_J\} \subset \mathcal{T}$. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ denote a separable Hilbert space, where $\langle \cdot, \cdot \rangle$ represents the inner product which generates the norm $\| \cdot \|$. Then, the functional random variable $\mathcal{X}$ is defined as $\mathcal{X} : (\Omega, \Sigma, P) \to \mathcal{H}$ where Ω, Σ, and P represent the sample space, σ-algebra and the probability measure, respectively. Throughout this research, we consider $L^2(\mathcal{T})$-Hilbert space of square integrable functions $f : \mathcal{T} \to \mathbb{R}$ satisfying $\int_{\mathcal{T}} f^2(t)dt < \infty$ with an inner product $\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)dt, \forall f, g \in L^2$. In addition, we assume that the random variable $\mathcal{X}(t) \in L^2(\Omega)$ with finite second order moment is a second order stochastic process so that $E[|\mathcal{X}|^2] = \int_\Omega |\mathcal{X}|^2 dP < \infty$.

The functional relationship between a functional response $\mathcal{Y}_i(t)$ and M functional predictors $\mathcal{X}_{im}(s)$; $\{(\mathcal{Y}_i(t), \mathcal{X}_{im}(s)); i = 1, ..., N, m = 1, ..., M, s \in \mathcal{T}_m, t \in \mathcal{T}\}$, where $\mathcal{T} \subset \mathbb{R}$ and $\mathcal{T}_m \subset \mathbb{R}$ are the ranges of response and predictors, respectively, can be formulated by the following multiple functional linear regression model (Ramsay and Silverman 2005):

$$\mathcal{Y}_i(t) = \beta_0(t) + \sum_{m=1}^{M} \int_{\mathcal{T}_m} \mathcal{X}_{im}(s)\beta_m(s, t)ds + \epsilon_i(t), \tag{2.5}$$

where $\beta_0(t)$ is the mean function, $\beta_m(s, t)$ is the bivariate coefficient function of the mth predictor, and $\epsilon_i(t)$ is the random error function. Without loss of generality, the

mean function can be eliminated from the model (2.5) by centering the functional response and predictors as follows:

$$\mathcal{Y}_i^*(t) = \mathcal{Y}_i(t) - \bar{\mathcal{Y}}(t),$$
$$\mathcal{X}_{im}^*(s) = \mathcal{X}_{im}(s) - \bar{\mathcal{X}}_m(s),$$

where $\bar{\mathcal{Y}}(t) = N^{-1}\sum_{i=1}^{N}\mathcal{Y}_i(t)$ and $\bar{\mathcal{X}}_m(s) = N^{-1}\sum_{i=1}^{N}\mathcal{X}_{im}(s)$ denote the mean of the functional response and mth functional predictor, respectively. Consequently, the functional regression model (2.5) can be re-written as follows:

$$\mathcal{Y}_i^*(t) = \sum_{m=1}^{M}\int_{\mathcal{T}_m}\mathcal{X}_{im}^*(s)\beta_m(s,t)ds + \epsilon_i^*(t), \tag{2.6}$$

where $\epsilon_i^*(t) = \epsilon_i(t) - \bar{\epsilon}(t)$ is the centered error function. The estimation process of the regression coefficient function $\beta_m(s,t)$ consists of three steps: (i) First, discretize the estimation problem, (ii) solve for a matrix, say $\boldsymbol{B}$, and (iii) apply smoothing step to obtain coefficient function.

A function $x(t)$ can be approximated by a linear combination of basis functions and associated coefficients for a sufficiently large number of basis functions K:

$$x(t) = \sum_{k=1}^{K} c_k\phi_k(t),$$

where, for $k = 1, ..., K$, $\phi_k(t)$ and c_k denote the basis functions and corresponding coefficients, respectively. The popular basis functions include Fourier, B-spline, and Gaussian basis functions (Ramsay and Silverman 2005; Matsui, Kawano, and Konishi 2009). In this study, we consider the B-spline basis to approximate functions.

The centered functional response and functional predictors, as well as the bivariate coefficient function, can be written as basis function expansions as follows:

$$\mathcal{Y}_i^*(t) = \sum_{k=1}^{K_y} c_{ik}\phi_k(t) = \mathrm{c}_i^\top\boldsymbol{\Phi}(t), \quad \forall t \in \mathcal{T},$$
$$\mathcal{X}_{im}^*(s) = \sum_{j=1}^{K_{m,\mathcal{X}}} d_{imj}\psi_{mj}(s) = \boldsymbol{d}_{im}^\top\boldsymbol{\Psi}_m(s), \quad \forall s \in \mathcal{T}_m,$$
$$\beta_m(s,t) = \sum_{j,k}\psi_{mj}(s)b_{mjk}\phi_k(t) = \boldsymbol{\Psi}_m^\top(s)\mathrm{B}_m\boldsymbol{\Phi}(t),$$

where $\boldsymbol{\Phi}(t) = (\phi_1(t), ..., \phi_{K_y}(t))^\top$ and $\boldsymbol{\Psi}_m(s) = (\psi_{m1}(s), ..., \psi_{mK_{m,\mathcal{X}}}(s))^\top$ represent the vector of basis functions, $\boldsymbol{c}_i = (c_{i1}, ..., c_{iK_y})^\top$ and $\boldsymbol{d}_{im} = (d_{im1}, ..., d_{imK_{m,\mathcal{X}}})^\top$ are the corresponding coefficient vectors, and $\boldsymbol{B}_m = (b_{mjk})_{j,k}$ denotes the coefficient matrix of dimension $K_{m,\mathcal{X}} \times K_{\mathcal{Y}}$. Accordingly, the regression model (2.6) can be re-expressed as follows:

$$\begin{aligned}\boldsymbol{c}_i^\top\boldsymbol{\Phi}(t) &= \sum_{m=1}^{M}\boldsymbol{d}_{im}^\top\boldsymbol{\zeta}_{\psi_m}\boldsymbol{B}_m\boldsymbol{\Phi}(t) + \epsilon_i^*(t),\\ &= \boldsymbol{z}_i^\top\boldsymbol{B}\boldsymbol{\Phi}(t) + \epsilon_i^*(t),\end{aligned} \tag{2.7}$$

where $\boldsymbol{\zeta}_{\psi_m} = \int_{\mathcal{T}_m}\boldsymbol{\Psi}_m(s)\boldsymbol{\Psi}_m^\top(s)ds$ is a $K_{m,\mathcal{X}} \times K_{m,\mathcal{X}}$ matrix, $\boldsymbol{z}_i = (\boldsymbol{d}_{i1}^\top\boldsymbol{\zeta}_{\psi_1}, ..., \boldsymbol{d}_{iM}^\top\boldsymbol{\zeta}_{\psi_M})^\top$ is a

vector of length $\sum_{m=1}^{M} K_{m,\mathcal{X}}$ and $\boldsymbol{B} = (\boldsymbol{B}_1, ..., \boldsymbol{B}_M)^\top$ is the coefficient matrix with dimension $\sum_{m=1}^{M} K_{m,\mathcal{X}} \times K_{\mathcal{Y}}$.

Suppose now that the error function $\epsilon_i^*(t)$ in (2.7) can be approximated as basis function expansion, $\epsilon_i^*(t) = \mathrm{e}_i^\top \boldsymbol{\Phi}(t)$, where $\mathrm{e}_i = (e_{i1}, ..., e_{iK})^\top$ is a vector of independently and identically distributed Gaussian random variables with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Then, the regression model in (2.7) has the following form:

$$\mathrm{c}_i^\top \boldsymbol{\Phi}(t) = \mathrm{z}_i^\top \mathrm{B}\boldsymbol{\Phi}(t) + \mathrm{e}_i^\top \boldsymbol{\Phi}(t). \tag{2.8}$$

Multiplying both sides of Eq. (2.8) from the right by $\boldsymbol{\Phi}^\top(t)$ and integrating with respect to $\mathcal{T}$ yields:

$$\begin{aligned}
\boldsymbol{c}_i^\top \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^\top(t) &= \boldsymbol{z}_i^\top \boldsymbol{B}\boldsymbol{\Phi}(t)\boldsymbol{\Phi}^\top(t) + \boldsymbol{e}_i^\top \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^\top(t),\\
\int_{\mathcal{T}} \boldsymbol{c}_i^\top \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^\top(t)dt &= \int_{\mathcal{T}} \boldsymbol{z}_i^\top \boldsymbol{B}\boldsymbol{\Phi}(t)\boldsymbol{\Phi}^\top(t)dt + \int_{\mathcal{T}} \boldsymbol{e}_i^\top \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^\top(t)dt,\\
\boldsymbol{c}_i^\top \zeta_\Phi &= \boldsymbol{z}_i^\top \boldsymbol{B}\zeta_\Phi + \boldsymbol{e}_i^\top \zeta_\Phi,\\
\boldsymbol{c}_i &= \boldsymbol{B}^\top \boldsymbol{z}_i + \boldsymbol{e}_i,
\end{aligned}$$

since ζ_Φ is nonsingular. The probability density function of a functional response $\mathcal{Y}_i$ given a functional predictor $\mathcal{X}_i$ and parameter vector $\theta = (\mathrm{B}, \boldsymbol{\Sigma})$ can be written as follow:

$$f(\mathcal{Y}_i|\mathcal{X}_i;\theta) = \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{c}_i - \boldsymbol{B}^\top \boldsymbol{z}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{c}_i - \boldsymbol{B}^\top \boldsymbol{z}_i)\},$$

since $\boldsymbol{e}_i \sim \boldsymbol{N}(0, \boldsymbol{\Sigma})$. Then, the log-likelihood function $\ell(\mathcal{Y}_i|\mathcal{X}_i;\theta) = \sum_{i=1}^{N} \log f(\mathcal{Y}_i|\mathcal{X}_i;\theta)$ is obtained as:

$$\ell(\mathcal{Y}_i|\mathcal{X}_i;\theta) = -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}tr\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{C} - \mathrm{ZB})^\top(\boldsymbol{C} - \mathrm{ZB})\},$$

where $\boldsymbol{C} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_N)^\top$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_N)^\top$. By equating the derivatives of the log-likelihood function with respect to $\theta = (\boldsymbol{B}, \boldsymbol{\Sigma})$ to $\mathbf{0}$, the ML estimators $\widehat{\theta} = (\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{\Sigma}})$ are obtained as follows:

$$\begin{aligned}
\widehat{\boldsymbol{B}} &= (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\mathrm{ZC},\\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{N}(\boldsymbol{C} - \boldsymbol{Z}\widehat{\boldsymbol{B}})^\top(\boldsymbol{C} - \boldsymbol{Z}\widehat{\boldsymbol{B}}).
\end{aligned}$$

Finally, the fitted values of the response function $\mathcal{Y}(t) = (\mathcal{Y}_1(t), ..., \mathcal{Y}_N(t))^\top$ are obtained as follows:

$$\widehat{\mathcal{Y}}(t) = (\boldsymbol{Z}\widehat{\boldsymbol{B}})\boldsymbol{\Phi}(t) + \bar{\mathcal{Y}}(t).$$

2.2. Functional response model for interval-valued functional data

The interval-valued functional data consist of functions that define the interval-valued data as functions. In other words, an interval-valued functional data consist of two functions; a lower limit function $\mathcal{X}^l(t)$ and an upper limit function $\mathcal{X}^u(t)$ such that

$\mathcal{X}(t) = (\mathcal{X}^l(t), \mathcal{X}^u(t)), \mathcal{X}^l(t) \leq \mathcal{X}^u(t)$. Denote by $\mathcal{X}^c(t) = (\mathcal{X}^u(t) - X^l(t))/2$ and $\mathcal{X}^r(t) = (\mathcal{X}^u(t) - X^l(t))/2$ the center and half-range functions, respectively. Let $\mathbf{\Phi}^l(t) = (\phi_1^l(t), ..., \phi_{K^l}^l(t)), \mathbf{\Phi}^u(t) = (\phi_1^u(t), ..., \phi_{K^u}^u(t)), \mathbf{\Phi}^c(t) = (\phi_1^c(t), ..., \phi_{K^c}^c(t))$, and $\mathbf{\Phi}^r(t) = (\phi_1^r(t), ..., \phi_{K^r}^r(t))$ denote the vectors of basis functions of the lower limit, upper limit, center, and half-range functions, respectively. Then, they can be expressed as the linear combinations of basis functions as follows:

$$\mathcal{X}^l(t) = \sum_{k=1}^{K^l} c_k^l \phi_k^l(t), \quad X^u(t) = \sum_{k=1}^{K^u} c_k^u \phi_k^u(t), \quad X^r(t) = \sum_{k=1}^{K^r} c_k^r \phi_k^r(t), \quad X^c(t) = \sum_{k=1}^{K^c} c_k^c \phi_k^c(t),$$

where c_k^l, c_k^u, c_k^c, and c_k^r denote the coefficient vectors of the lower limit, upper limit, center, and half-range functions, respectively.

Before introducing the functional forms of the interval-valued data regression models, we note that for an interval-valued functional data $\mathcal{X}(t) = (\mathcal{X}^l(t), \mathcal{X}^u(t))$, one needs to find a common vector of basis functions $\mathbf{\Phi}^{cm}(t)$ that works for all the components in the interval; lower limit, upper limit, center, and half-range functions. Otherwise, the regression model may not be constructed since different components may have a different number of basis functions.

Suppose now that the (centered) interval-valued random functional response $\mathcal{Y}_i^*(t)$ and functional predictors $\mathcal{X}_{im}^*(s)$, for $i = 1, ..., N$ and $m = 1, ..., M$, be the realizations of the intervals $\mathcal{Y}_i^*(t) = (\mathcal{Y}_i^{*l}(t), \mathcal{Y}_i^{*u}(t))$ and $\mathcal{X}_{im}^*(s) = (\mathcal{X}_{im}^{*l}(s), \mathcal{X}_{im}^{*u}(s))$. Denote by $\mathcal{Y}_i^{*c}(t)$ and $\mathcal{X}_{im}^{*c}(s)$ the center functions of the response and mth predictor variables, respectively. Then, the CM defined in (2.2) can be extended to the functional data as follows:

$$\mathcal{Y}_i^{*c}(t) = \sum_{m=1}^{M} \int_{\mathcal{T}_m} \mathcal{X}_{im}^{*c}(s) \beta_m^c(s,t) ds + \epsilon_i^{*c}(t). \tag{2.9}$$

Let $\boldsymbol{C}^c = (\boldsymbol{B}^c)^\top \boldsymbol{Z}^c + \boldsymbol{e}^c$ denote the multivariate regression model constructed using the basis function expansions of the functional objects given in (2.9) and following Section (2.1). Let also $\widehat{\boldsymbol{B}}^c$ denote the ML estimate of $\boldsymbol{B}^c$. Then, the predictions of the lower and upper limit functions of the interval-valued functional response variable using the functional CM are obtained as follows:

$$\widehat{\mathcal{Y}}^l(t) = (\boldsymbol{Z}^l \widehat{\boldsymbol{B}}^c) \mathbf{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^l(t), \quad \widehat{\mathcal{Y}}^u(t) = (\boldsymbol{Z}^u \widehat{\boldsymbol{B}}^c) \mathbf{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^u(t),$$

where $\bar{\mathcal{Y}}^l(t) = N^{-1} \sum_{i=1}^N \mathcal{Y}_i^l(t), \bar{\mathcal{Y}}^u(t) = N^{-1} \sum_{i=1}^N \mathcal{Y}_i^u(t), \boldsymbol{Z}^l$ and $\boldsymbol{Z}^u$ denote the matrices (see (2.7)) obtained using the basis function expansions of the lower and upper limit functions of the predictors, respectively, and $\mathbf{\Phi}^{cm}(t)$ is the vector of basis functions used to approximate the components of the response variable.

Similarly to the traditional case, the functional CRM is based on two distinct function-on-function regression models: the regression models of the center and half-range functions of the variables. Let $\mathcal{Y}_i^{*r}(t) = (\mathcal{Y}_i^{*u}(t) - \mathcal{Y}_i^{*l}(t))/2$ and $\mathcal{X}_{im}^{*r}(s) = (\mathcal{X}_{im}^{*u}(s) - \mathcal{X}_{im}^{*l}(s))/2$ denote the half-range functions of the centered functional response and predictor variables, respectively. Then, the functional regression equation of the functional half-range variables is given by:

$$\mathcal{Y}_i^{*r}(t) = \sum_{m=1}^{M} \int_{\mathcal{T}_m} \mathcal{X}_{im}^{*r}(s)\beta_m^r(s,t)ds + \epsilon_i^{*r}(t).$$

Denote by $\boldsymbol{C}^r = (\boldsymbol{B}^r)^\top \boldsymbol{Z}^r + \boldsymbol{e}^r$ the multivariate regression model obtained using the basis function expansions of the half-range functions, and let $\widehat{\boldsymbol{B}}^r$ be the ML estimate of $\boldsymbol{B}^r$. Then, the predicted half-range of the functional response variable is obtained as follows:

$$\widehat{\mathcal{Y}}^r(t) = (\boldsymbol{Z}^r\widehat{\boldsymbol{B}}^r)\boldsymbol{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^r(t),$$

where $\boldsymbol{Z}^r$ is a matrix obtained by the basis function expansions of the half-range functions of the predictor variables and $\bar{\mathcal{Y}}^r(t) = N^{-1}\sum_{i=1}^{N}\mathcal{Y}_i^r(t)$. Subsequently, the predictions of the lower and upper limit functions of the response variable using the functional CRM are obtained as follows:

$$\widehat{\mathcal{Y}}^l(t) = \widehat{\mathcal{Y}}^c(t) - \widehat{\mathcal{Y}}^r(t), \quad \widehat{\mathcal{Y}}^u(t) = \widehat{\mathcal{Y}}^c(t) + \widehat{\mathcal{Y}}^r(t),$$

where $\widehat{\mathcal{Y}}^c(t) = (\boldsymbol{Z}^c\widehat{\boldsymbol{B}}^c)\boldsymbol{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^c(t)$ with $\bar{\mathcal{Y}}^c(t) = N^{-1}\sum_{i=1}^{N}\mathcal{Y}_i^c(t)$.

The functional extension of the BCRM can be obtained similarly to (2.4) but using the functional variables as follows:

$$\mathcal{Y}_i^{*c}(t) = \sum_{m=1}^{M} \int_{\mathcal{T}_m} \mathcal{X}_{im}^{*cr}(s)\beta_m^{ccr}(s,t)ds + \epsilon_i^{*ccr}(t), \tag{2.10}$$

$$\mathcal{Y}_i^{*r}(t) = \sum_{m=1}^{M} \int_{\mathcal{T}_m} \mathcal{X}_{im}^{*cr}(s)\beta_m^{rcr}(s,t)ds + \epsilon_i^{*rcr}(t), \tag{2.11}$$

where $\mathcal{X}_{im}^{*cr}(s) = (\mathcal{X}_{im}^{*c}(s), \mathcal{X}_{im}^{*r}(s))$, for $i = 1, ..., N$ and $m = 1, ..., M$. Let us denote by $\boldsymbol{C}^{ccr} = (\boldsymbol{B}^{ccr})^\top \boldsymbol{Z}^{ccr} + \boldsymbol{e}^{ccr}$ and $\boldsymbol{C}^{rcr} = (\boldsymbol{B}^{rcr})^\top \boldsymbol{Z}^{rcr} + \boldsymbol{e}^{rcr}$ the regression models constructed by the basis function expansions of the functional objects given in (2.10) and (2.11), respectively. Let $\widehat{\boldsymbol{B}}^{ccr}$ and $\widehat{\boldsymbol{B}}^{rcr}$ denote the ML estimates of $\boldsymbol{B}^{ccr}$ and $\boldsymbol{B}^{rcr}$, respectively. Then, the predictions of the lower and upper limit functions of the response variable using the functional BCRM are obtained as follows:

$$\widehat{\mathcal{Y}}^l(t) = \widehat{\mathcal{Y}}^c(t) - \widehat{\mathcal{Y}}^r(t), \quad \widehat{\mathcal{Y}}^u(t) = \widehat{\mathcal{Y}}^c(t) + \widehat{\mathcal{Y}}^r(t),$$

where $\widehat{\mathcal{Y}}^c(t) = (\boldsymbol{Z}^{crc}\widehat{\boldsymbol{B}}^{crc})\boldsymbol{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^c(t)$ and $\widehat{\mathcal{Y}}^r(t) = (\boldsymbol{Z}^{rrc}\widehat{\boldsymbol{B}}^{rrc})\boldsymbol{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^c(t)$.

To extend the traditional MCM to the functional case, we provide the following algorithm.

Step 1. For $i = 1, ..., N$ and $m = 1, ..., M$, generate a functional response $\mathcal{Y}_i(t)$ and N sets of M functional predictors $\mathcal{X}_{im}(s)$ uniformly from their intervals, $(\mathcal{Y}_i^l(t), \mathcal{Y}_i^u(t))$ and $(\mathcal{X}_{im}^l(s), \mathcal{X}_{im}^u(s))$.

Step 2. Center the generated functional response and functional predictors and construct a functional regression model as follows:

$$\mathcal{Y}_i^*(t) = \sum_{m=1}^{M} \int_{\mathcal{T}_m} \mathcal{X}_{im}^*(s)\beta_m(s,t)ds + \epsilon_i^*(t).$$

Step 3. Obtain the multivariate regression model $\boldsymbol{C} = (\boldsymbol{B})^{\top}\boldsymbol{Z} + \boldsymbol{e}$ using the basis function expansions of the centered functional objects. Then, apply the ML method to estimate the regression coefficients matrix, $\widehat{\boldsymbol{B}} = (\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\mathrm{ZC}$ as explained in Sec. 2.1.

Step 4. Repeat Steps 1–3 B times to obtain B sets of coefficient matrices $\widehat{\boldsymbol{B}}^{\mathrm{MCM}} = (\widehat{\boldsymbol{B}}^{1}, ..., \widehat{\boldsymbol{B}}^{B})$.

Let $\overline{\widehat{\boldsymbol{B}}}^{\mathrm{MCM}} = B^{-1}\sum_{b=1}^{B}\widehat{\boldsymbol{B}}^{b}$ be the final estimate of the functional MCM. Then, the lower and upper limit functions of the response variable are predicted as follows:

$$\widehat{\mathcal{Y}}^{l}(t) = (\boldsymbol{Z}^{l}\overline{\widehat{\boldsymbol{B}}}^{\mathrm{MCM}})\boldsymbol{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^{l}(t), \quad \widehat{\mathcal{Y}}^{u}(t) = (\boldsymbol{Z}^{u}\overline{\widehat{\boldsymbol{B}}}^{\mathrm{MCM}})\boldsymbol{\Phi}^{cm}(t) + \bar{\mathcal{Y}}^{u}(t).$$

Similarly to the traditional case, for the functional CM and MCM, the calculated lower limit functions of the response variable, $\widehat{\mathcal{Y}}_{i}^{l}(t)$ may be greater than the upper limit functions, $\widehat{\mathcal{Y}}_{i}^{u}$. To overcome this problem, we recommend to take the pointwise minimum and maximum values of the functions as follows:

$$\widehat{\mathcal{Y}}_{i}^{l}(t_j) = \min\left(\widehat{\mathcal{Y}}_{i}^{l}(t_j), \widehat{\mathcal{Y}}_{i}^{u}(t_j)\right), \quad \widehat{\mathcal{Y}}_{i}^{u}(t_j) = \max\left(\widehat{\mathcal{Y}}_{i}^{l}(t_j), \widehat{\mathcal{Y}}_{i}^{u}(t_j)\right),$$

for $i = 1, ..., N$ and $j = 1, ..., J$.

The functional MCM can also be used to construct prediction intervals for the lower and upper limit functions of the response variable. Let $\widehat{\epsilon}^{l}(t) = \mathcal{Y}^{l}(t) - \widehat{\mathcal{Y}}^{l}(t)$ and $\widehat{\epsilon}^{u} = \mathcal{Y}^{u}(t) - \widehat{\mathcal{Y}}^{u}(t)$, respectively, denote the estimated error functions for the lower and upper limits obtained using the functional MCM. Then, for $b = 1, ..., B$, one can calculate B sets of fitted lower $(\widehat{\mathcal{Y}}^{l,1}, ..., \widehat{\mathcal{Y}}^{l,B})$ and upper $(\widehat{\mathcal{Y}}^{u,1}, ..., \widehat{\mathcal{Y}}^{u,B})$ limit functions as follows:

$$\widehat{\mathcal{Y}}^{l,b} = (\boldsymbol{Z}^{l}\widehat{\boldsymbol{B}}^{b})\boldsymbol{\Phi}^{cm}(t) + \epsilon^{l,*} + \bar{\mathcal{Y}}^{l}(t), \quad \widehat{\mathcal{Y}}^{u,b} = (\boldsymbol{Z}^{u}\widehat{\boldsymbol{B}}^{b})\boldsymbol{\Phi}^{cm}(t) + \epsilon^{u,*} + \bar{\mathcal{Y}}^{u}(t),$$

where, $\epsilon^{l,*}$ and $\epsilon^{u,*}$ denote random samples from $\widehat{\epsilon}^{l}$ and $\widehat{\epsilon}^{u}$, respectively. Denote by $\mathbb{Q}_{\alpha}^{l}(t)$ and $\mathbb{Q}_{\alpha}^{u}(t)$ the α^{th} quantiles of the generated B sets of MCM replicates of the fitted lower and upper limit functions, respectively. Then, the $100(1-\alpha)\%$ functional MCM prediction intervals for $\mathcal{Y}^{l}(t)$ and $\mathcal{Y}^{u}(t)$ can be computed as $[\mathbb{Q}_{\alpha/2}^{l}(t), \mathbb{Q}_{1-\alpha/2}^{l}(t)]$ and $[\mathbb{Q}_{\alpha/2}^{u}(t), \mathbb{Q}_{1-\alpha/2}^{u}(t)]$, respectively.

3. Numerical results

Various Monte Carlo simulations and empirical data analysis were conducted to investigate the finite sample performance of the proposed interval-valued functional data regression models. We note that in our calculations, the generalized inverse was used to estimate the model parameter matrix $\boldsymbol{B}$ to avoid the singular-matrix problem. All the numerical analyses were performed using R 3.6.0 (an example R code can be found at https://github.com/UfukBeyaztas/FLM_interval_valued_data).

3.1. Simulation studies

Throughout the simulations, the multiple function-on-function regression model with $M = 3$ functional predictors was considered, and the following process was used to generate interval-valued functional variables.

a. For $i = 1, ..., N = 200$ and $m = 1, 2, 3$, use the following process to generate center functions of the predictors $\mathcal{X}_m^c(s)$; $\mathcal{X}_m^c(s) = 10 + V_m(s)$, where $s \in [0, 1]$ and $V_m(s)$ s are generated from the Gaussian process with zero mean and a positive variance-covariance function $\mathbf{\Sigma}_V(s, s') = \exp(-100(s - s'))^2$.
b. Generate the center functions of the error process; $\epsilon^c(t)$, where $t \in [0, 1]$, from the normal distribution with zero mean and variance four; $N(0, 4)$. Then, generate the center functions of the response variable as follows:

$$\mathcal{Y}^c(t) = \sum_{m=1}^{3} \int_0^1 \mathcal{X}_m^c(s)\beta_m^c(s, t) + \epsilon^c(t),$$

where

$$\beta_1^c(s, t) = (1 - s)^2(t - 0.5)^2$$
$$\beta_2^c(s, t) = \exp(-3(s - 1)^2)\exp(-5(t - 0.5)^2)$$
$$\beta_3^c(sit) = \exp(-5(s - 0.5)^2 - 5(t - 0.5)^2) + 8\exp(-5(s - 1.5)^2 - 5(t - 0.5)^2)$$

c. Generate the range functions of the response $\mathcal{Y}^r(t)$ and predictors $\mathcal{X}_m^r(s)$ as follows:

$$\mathcal{Y}^r(t) = \mathcal{Y}^c(t) + U(a, b), \quad \mathcal{X}_m^r(s) = \mathcal{X}_m^r(s) + U(c, d).$$

For the range functions, four different cases were considered:

$$(a, b) = [(1, 1.5), (1, 3), (3, 5), (8, 20)] \text{and} (c, d) = [(1, 1.5), (1, 3), (5, 8), (6, 15)].$$

d. Calculate the lower and upper limit functions of the response and predictors as follow:

$$\mathcal{Y}^l(t) = \mathcal{Y}^c(t) - \mathcal{Y}^r(t)/2, \quad \mathcal{Y}^u(t) = \mathcal{Y}^c(t) + \mathcal{Y}^r(t)/2,$$
$$\mathcal{X}_m^l(s) = \mathcal{X}_m^c(s) - \mathcal{X}_m^r(s)/2, \quad \mathcal{X}_m^l(s) = \mathcal{X}_m^c(s) - \mathcal{X}_m^r(s)/2.$$

All the components of the functional variables were generated at 100 equally spaced points in the interval [0,1], and the components of the generated predictor variables were distorted by the Gaussian noise $u(s) \sim N(0, 4)$ before fitting the interval-valued functional regression models. A graphical display of the generated lower and upper limit functions are presented in Figure 1.

For each simulation setting, MC = 250 Monte Carlo simulations were performed. For each simulation, the first 100 functions of the generated functional variables were used to construct the interval-valued functional regression models. Note that for each simulation setting, eight number of basis functions was used to convert all the components of the generated noisy data to their functional forms. The last 100 functions of the data were used to evaluate the prediction performances of the proposed methods. The

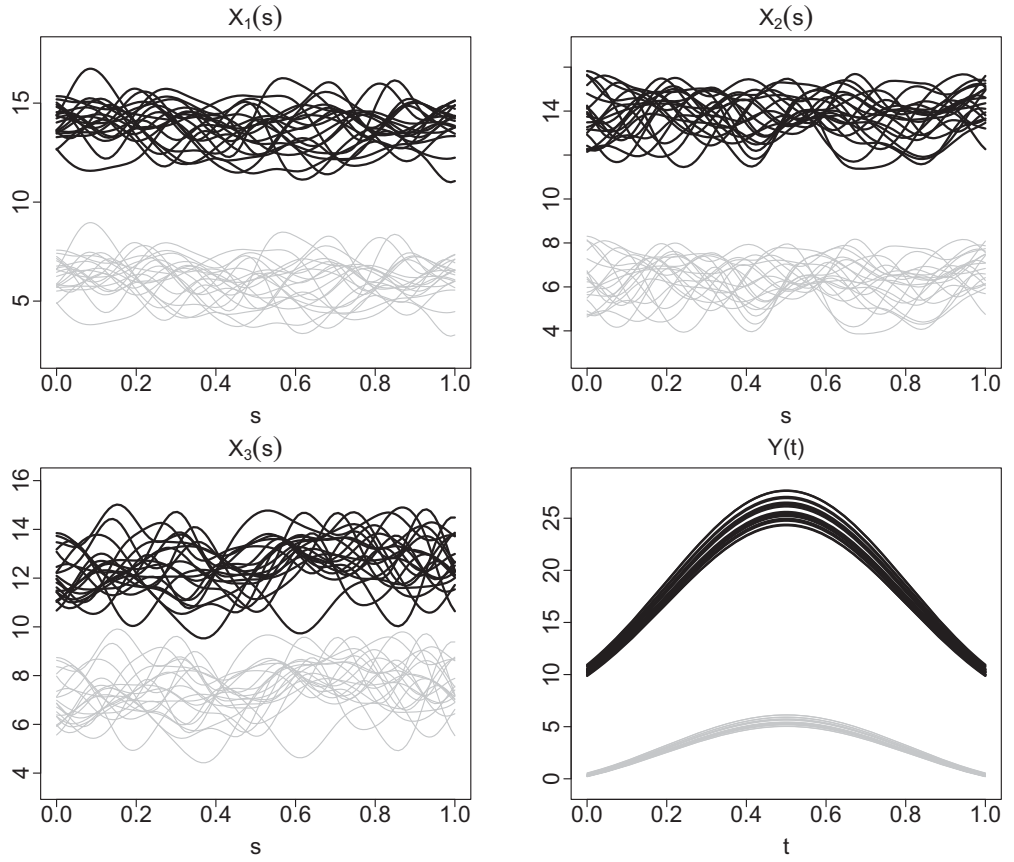


Figure 1. Plots of the generated 20 lower (gray lines) and upper (black lines) limit functions when $(a, b) = (3, 5)$ and $(c, d) = (5, 8)$.

prediction performance of the methods was evaluated using the average mean squared errors (AMSE) of the lower (AMSE^l) and upper (AMSE^u) limit functions as follows:

$$\text{AMSE}^l = \frac{1}{100}\sum_{i=1}^{100} ||\mathcal{Y}_i^l(t) - \widehat{\mathcal{Y}}_i^l(t)||_{L^2},$$

$$\text{AMSE}^u = \frac{1}{100}\sum_{i=1}^{100} ||\mathcal{Y}_i^u(t) - \widehat{\mathcal{Y}}_i^u(t)||_{L^2},$$

where $||\cdot||_{L^2}$ is the L^2 norm, which is approximated by the Riemann sum (Luo and Qi 2019). Note that, for the functional MCM, the results were obtained based on a $B = 100$ re-sampling procedure. The nominal significance level α was set to 0.05 to construct functional MCM-based prediction intervals for the lower and upper limit functions. The coverage probabilities (CP) for the prediction intervals of lower (CP^l) and upper (CP^u) limit functions were calculated to evaluate the accuracy of the constructed prediction intervals:

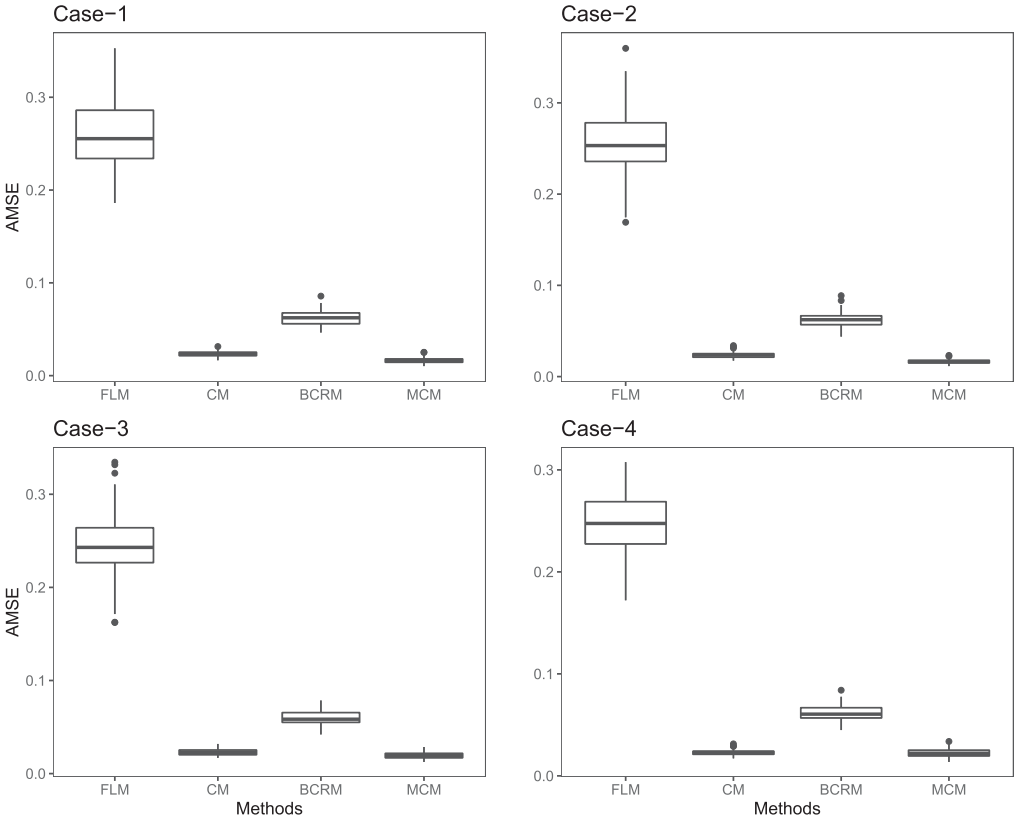


Figure 2. Calculated AMSE values of the FLM, CM, BCRM, and MCM for the lower limit function of the response variable in all cases: Case-1: $(a, b) = (1, 1.5), (c, d) = (1, 1.5)$, Case-2: $(a, b) = (1, 3), (c, d) = (1, 3)$, Case-3: $(a, b) = (3, 5), (c, d) = (5, 8)$, and Case-4: $(a, b) = (8, 20), (c, d) = (6, 15)$.

$$\mathrm{CP}^l = \frac{1}{100 \times 100} \sum_{i=1}^{100} \sum_{j=1}^{100} 1\{\mathbb{Q}^l_{\alpha/2,i,j} \leq \mathcal{Y}^l_i(t_j) \leq \mathbb{Q}^l_{1-\alpha/2,i,j}\},$$

$$\mathrm{CP}^u = \frac{1}{100 \times 100} \sum_{i=1}^{100} \sum_{j=1}^{100} 1\{\mathbb{Q}^u_{\alpha/2,i,j} \leq \mathcal{Y}^u_i(t_j) \leq \mathbb{Q}^u_{1-\alpha/2,i,j}\},$$

where $1\{\cdot\}$ denotes the indicator function. The finite sample performance of the proposed methods was compared with the classical functional linear model (FLM). While doing so, two FLMs for the lower and upper limit functions of the response and predictor variables were constructed as follows:

$$\mathcal{Y}^l_i(t) = \int_0^1 \mathcal{X}^l_{im}(s)\beta^l_m(s,t)ds + \epsilon^l_i(t),$$

$$\mathcal{Y}^u_i(t) = \int_0^1 \mathcal{X}^u_{im}(s)\beta^u_m(s,t)ds + \epsilon^u_i(t).$$

Our findings showed that functional CRM and BCRM produce very close results for each other, and thus, we presented the results of the functional BCRM only. The results

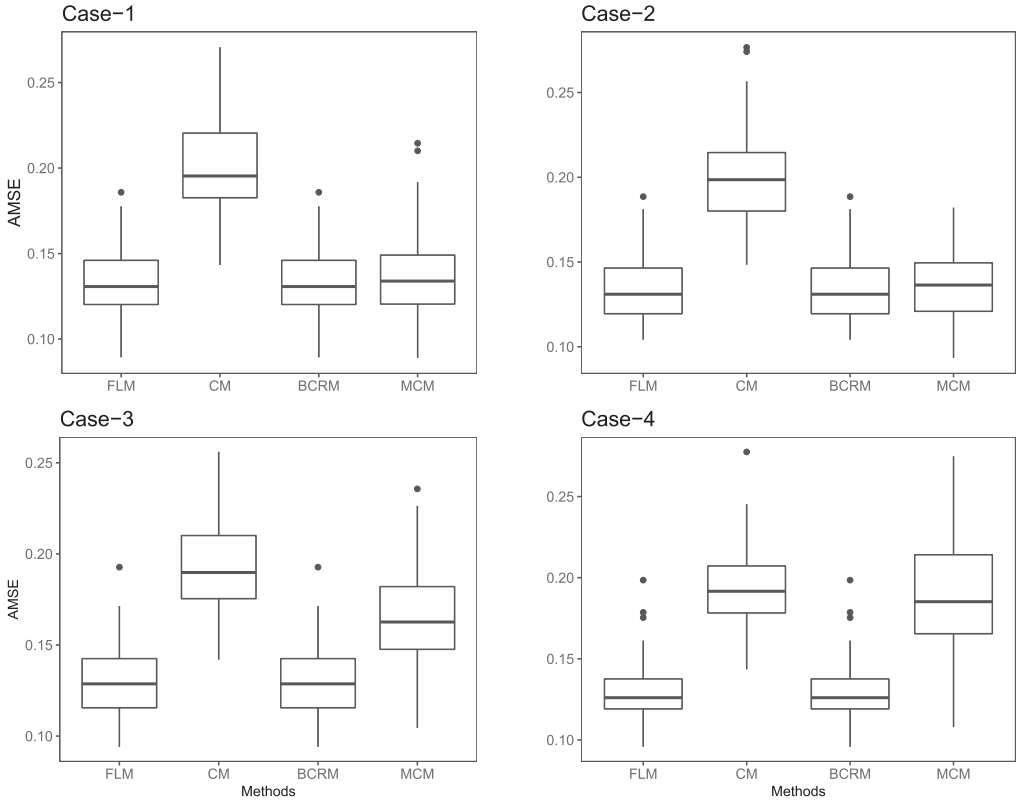


Figure 3. Calculated AMSE values of the FLM, CM, BCRM, and MCM for the upper limit function of the response variable in all cases: Case-1: $(a, b) = (1, 1.5), (c, d) = (1, 1.5)$, Case-2: $(a, b) = (1, 3), (c, d) = (1, 3)$, Case-3: $(a, b) = (3, 5), (c, d) = (5, 8)$, and Case-4: $(a, b) = (8, 20), (c, d) = (6, 15)$.

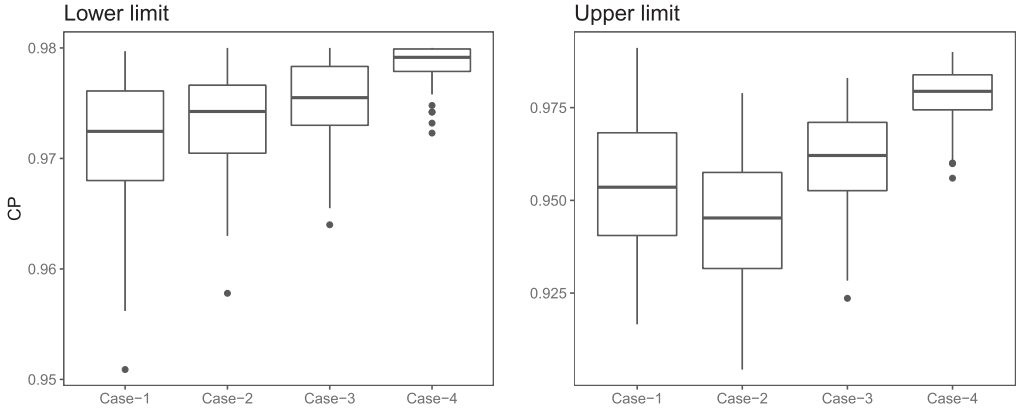


Figure 4. Calculated coverage probabilities of the MCM-based prediction intervals for the lower and upper limit functions in all cases: Case-1: $(a, b) = (1, 1.5), (c, d) = (1, 1.5)$, Case-2: $(a, b) = (1, 3), (c, d) = (1, 3)$, Case-3: $(a, b) = (3, 5), (c, d) = (5, 8)$, and Case-4: $(a, b) = (8, 20), (c, d) = (6, 15)$.

for the lower and upper limit functions are presented in Figures 2 and 3, respectively. The results demonstrate that, for the lower limit functions, the proposed interval-valued functional regression models outperform the FLM for all cases. Compared with

Table 1. Station names for the Oman monthly weather data.

Station	Station	Station	Station	Station
Adam Airport	Bidiyah	Khasab Port	Muscat City	Shalim
AlJubah	Bowsher	Liwa	Qairoon Hairiti	Sohar Airport
Al Amrat	Bukha	Madha	Qalhat	Sunaynah
Al Hamra	Dhank Qumaira	Mahdah	Qarn alam	Sur
Al Jazir	Diba	Majis	Qurayyat	Suwaiq
Al Mazyunah	Duqum Airport	Masirah	Ras AlHaad	Taqah
Al Mudhaibi	Ibra	Mina Salalah	Rustaq	Thamrayt
Al-Buraymi	Ibri	Mina Sultan Qaboos	Sadah	Yanqul
Bahla	Izki	Mirbat	Salalah Airport	
Bidbid	Khasab Airport	Muqshin	Samail	

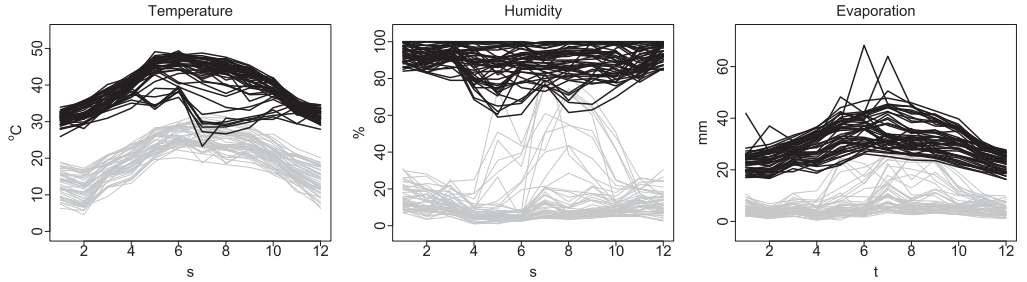


Figure 5. Time series plots of the averaged minimum (gray lines) and maximum (black lines) monthly Oman weather variables.

functional BCRM, the functional CM and MCM produce better AMSE values (see Figure 2). On the other hand, for the upper limit functions, the functional CM performs less among others for all cases. The FLM, BCRM, and MCM tend to produce similar AMSE values when the range between the lower and upper limit functions is small (Case-1 and Case-2) while the FLM and BCRM produce better results compared to the MCM as the range increases (Case-2 and Case-3). The results presented in Figures 2 and 3 also show that the FLM and proposed interval-valued functional regression models (except functional MCM) are not affected by the range when predicting lower and upper limit functions. Only the performance of the functional MCM gets worse as the range increases when predicting upper limit functions.

The coverage performances of the functional MCM-based prediction intervals are presented in Figure 4. It is clear from this figure that the functional MCM is capable of producing valid prediction intervals for upper limit functions and generally produces coverage probabilities close to the nominal significance level. On the other hand, for the lower limit functions, over-coverage is observed for all cases.

3.2. Empirical data example: Oman weather data

The empirical data example represents the monthly Oman weather data spanning from January 2017 to December 2018. The dataset contain three interval-valued variables: monthly evaporation (mm), humidity (%), and temperature (°C), and were collected from 48 stations across Oman (dataset is available from the National Center for Statistics & Information: https://data.gov.om). The list of stations is reported in Table 1.

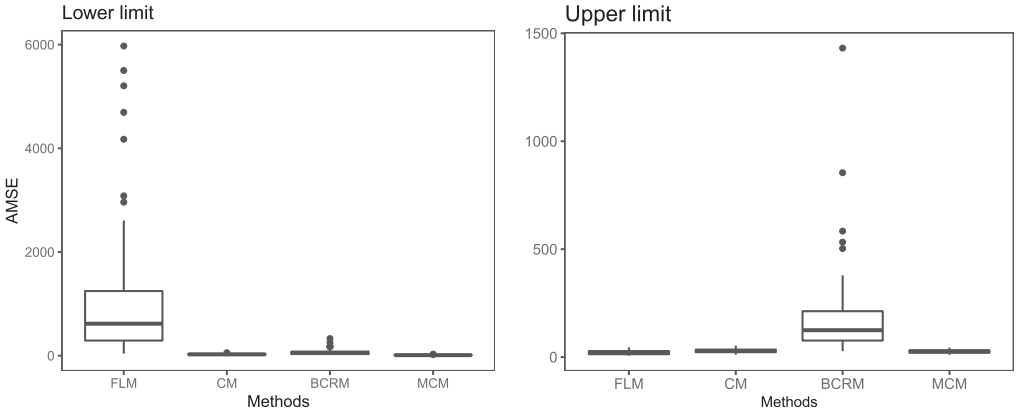


Figure 6. Estimated AMSE values for the monthly Oman weather data.

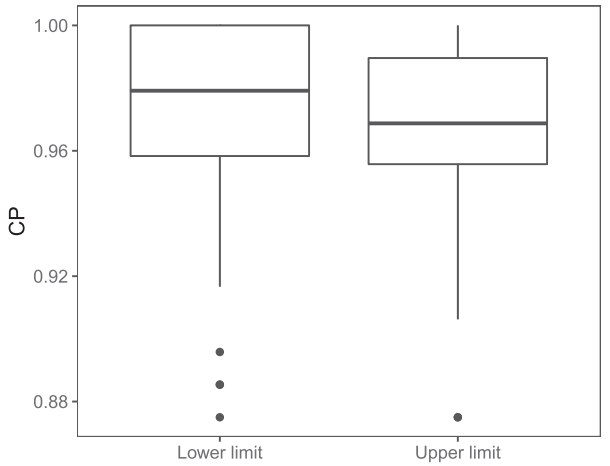


Figure 7. Estimated AMSE values for the monthly Oman weather data.

The data were averaged for each interval-valued variable over the whole time, and the observations were considered as the functions of months, $1 \leq s, t \leq 12$. The plots of the averaged interval-valued functional variables for all 48 stations are presented in Figure 5.

We considered predicting monthly minimum and maximum evaporation using the minimum and maximum temperature and humidity variables. For this purpose, the values of the discretely observed minimum and maximum variables (and their center and range points) were first converted to functional forms by B-spline basis function expansion using 10 number of basis functions. The following procedure was repeated 100 times to evaluate the predictive performance of the proposed interval-valued functional regression models as well as the traditional FLM. In each repeat, the dataset was divided into two parts; the models were constructed based on the functions of 40 randomly selected stations to predict the minimum and maximum evaporations of the remaining eight stations. For each model, the parameter matrix $\mathbf{B}$ was estimated using the ML method, as explained in Sec. 2. For each replicate, the AMSE^l and AMSE^u values were calculated. Our findings are presented in Figure 6.

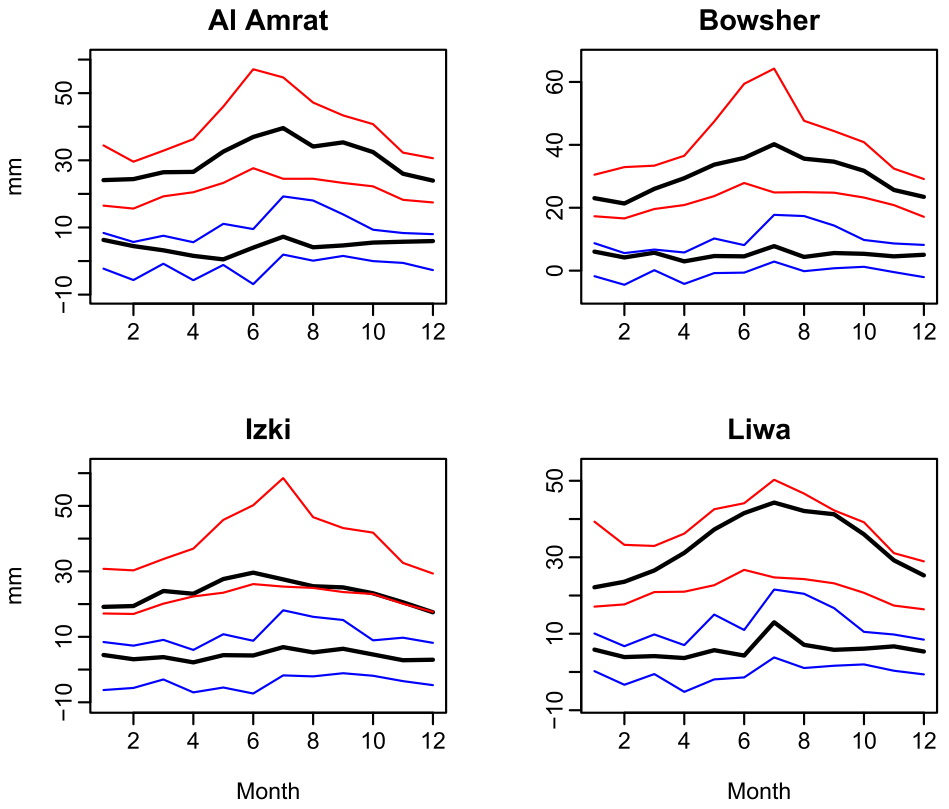


Figure 8. Plots of the observed lower and upper limit functions (black lines) and their calculated MCM-based 95% prediction intervals for four stations; Al Amrat, Bowsher, Izki, and Liwa. Prediction intervals for the lower and upper limit functions are given in blue and red colors, respectively.

The results show that all the proposed methods outperform the FLM for the lower limit functions. The functional CM and MCM produce slightly less $AMSE^{l}$ compared with BCRM. On the other hand, for the upper limit functions, the functional BCRM performs less, among others. The functional MCM and FLM produce slightly better performances compared with the functional CM. Also, the functional MCM-based prediction intervals were calculated, and the results are given in Figure 7. This figure shows that the functional MCM generally produce valid prediction intervals for both lower and upper limit functions. However, as it is seen from Figure 7, MCM produces low coverage probabilities for some cases. Because the dataset includes some outlying observations, the performance of MCM is affected by the presence of outlying observations. A robust estimation method may be used to overcome this problem.

An example of the constructed MCM-based prediction intervals is presented in Figure 8.

4. Conclusion

Interval-valued data regression models have become a natural framework to analyze datasets that are collected in an interval. On the other hand, recent technological advances in data collection tools cause complex and high dimensional datasets, which may not be analyzed by

existing traditional methods. FDA tools are one of the frequently used methods for visualizing and analyzing such data. Particularly, the FLMs have been one of the most commonly used techniques to explore the association between the functional response and predictor variables. Although the functional extensions of many traditional methods are available in the literature, the functional forms of the interval-valued data regression models have not yet been studied.

In this study, we present the functional forms of some well-known interval-valued data regression models and examine their prediction performances. The proposed methods are based on the function-on-function regression model, where both the response and predictors are functions which is common in real-life situations. The finite sample performance of the proposed interval-valued functional regression models is evaluated via Monte Carlo simulations and an empirical data analysis. Also, we compare the finite sample performance of the proposed methods with the traditional FLM. Our findings show that the proposed regression models are superior to the traditional FLM when predicting lower limit functions. Also, they produce competitive performance with FLM when predicting upper limit functions. Moreover, the proposed functional MCM is used to construct prediction intervals for both lower and upper limit functions of the response variable. All the numerical analyses considered in this study have shown that the functional MCM is generally capable of producing valid prediction intervals.

For future studies, the followings may be considered; (1) throughout this study, only the function-on-function regression is considered, but the proposed methods can also be extended to other functional regression models such as scalar-on-function and function-on-scalar, (2) we consider only the B-spline basis function to convert discretely observed data to functional form, other basis functions, such as Fourier, wavelet, and radial basis functions may also be considered, and (3) the model parameters of the proposed interval-valued functional regression models are estimated using the ML method, other estimation methods, such as partial least squares and principal component regression, may also be used.

Acknowledgments

We would like to thank an anonymous reviewer for her/his careful reading of our manuscript and valuable suggestions and comments, which have helped us produce an improved version of our manuscript. The publication of this article was funded by the Qatar National Library.

ORCID

Ufuk Beyaztas http://orcid.org/0000-0002-5208-4950
Han Lin Shang http://orcid.org/0000-0003-1769-6430
Abdel-Salam G. Abdel-Salam http://orcid.org/0000-0003-4905-6489

References

Ahn, J., M. Peng, C. Park, and Y. Jeon. 2012. A resampling approach for interval-valued data regression. *Statistical Analysis and Data Mining* 5 (4):336–48. doi:10.1002/sam.11150.

Alfonso, F., L. Billard, and E. Diday. 2004. Symbolic linear regression with taxonomies. In *Classification and clustering and data mining applications*, ed. D. Banks, L. House, F. McMorris, P. Arabie and W. Gaul, 429–37. Berlin: Springer.

Amato, U., A. Antoniadis, and I. De Feis. 2006. Dimension reduction in functional regression with applications. *Computational Statistics & Data Analysis* 50 (9):2422–46. doi:10.1016/j.csda.2004.12.007.

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

Billard, L. 2011. Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining* 4 (2):149–56. doi:10.1002/sam.10115.

Billard, L., and E. Diday. 2000. Regression analysis for interval-valued data. In *Data analysis, classification and related methods*, ed. H. A. L. Kiers, P. J. Rassoon, F. P. J. Groenen and M. Schaderl, 369–74. Berlin: Springer.

Billard, L., and E. Diday. 2002. Symbolic regression analysis. In *Proceedings of the 8th Conference of the International Federation of Classification Societies*, 281–8. Polan: Springer.

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

Cardot, H., F. Ferraty, and P. Sarda. 1999. Functional linear model. *Statistics & Probability Letters* 45 (1):11–22. doi:10.1016/S0167-7152(99)00036-X.

Cardot, H., F. Ferraty, and P. Sarda. 2003. Spline estimators for the functional linearmodel. *Statistica Sinica* 13:571–91.

Chen, D., P. Hall, and H. G. Müller. 2011. Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics* 39 (3):1720–47. doi:10.1214/11-AOS882.

Chiou, J. M., Y. F. Yang, and Y. T. Chen. 2016. Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis* 146:301–12. doi:10.1016/j.jmva.2015.10.003.

Cook, R. D., L. Forzani, and A. F. Yao. 2010. Necessary and sufficient conditions for consistency of a method for smoothed functional inverse regression. *Statistica Sinica* 20:235–8.

Cuevas, A. 2014. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference* 147:1–23.

Dou, W. W., D. Pollard, and H. H. Zhou. 2012. Estimation in functional regression for general exponential families. *The Annals of Statistics* 40 (5):2421–51. doi:10.1214/12-AOS1027.

Fan, J., and W. Zhang. 1999. Statistical estimation in varying coefficient models. *The Annals of Statistics* 27:1491–518. doi:10.1214/aos/1017939139.

Febrero-Bande, M., and W. Gonzalez-Manteiga. 2013. Generalized additive models for functional data. *Test* 22 (2):278–92. doi:10.1007/s11749-012-0308-0.

Ferraty, F., and P. Vieu. 2006. *Nonparametric functional data analysis*. New York: Springer.

Ferraty, F., and P. Vieu. 2009. Additive prediction and boosting for functional data. *Computational Statistics & Data Analysis* 53 (4):1400–13. doi:10.1016/j.csda.2008.11.023.

Goia, A., and P. Vieu. 2015. A partitioned single functional index model. *Computational Statistics* 30 (3):673–92. doi:10.1007/s00180-014-0530-1.

Hall, P., and J. L. Horowitz. 2007. Methodology and convergence rates for functional linear regression. *The Annals of Statistics* 35 (1):70–91. doi:10.1214/009053606000000957.

Harezlak, J., B. A. Coull, N. M. Laird, S. R. Magari, and D. C. Christiani. 2007. Penalized solutions to functional regression problems. *Computational Statistics & Data Analysis* 51 (10):4911–25. doi:10.1016/j.csda.2006.09.034.

Horvath, L., and P. Kokoszka. 2012. *Inference for functional data with applications*. New York: Springer.

Hu, Z., N. Wang, and R. J. Carroll. 2004. Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika* 31:251–62.

Ivanescu, A. E., A. M. Staicu, F. Scheipl, and S. Greven. 2015. Penalized function-on-function regression. *Computational Statistics* 30 (2):539–68. doi:10.1007/s00180-014-0548-4.

James, G. M. 2002. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3):411–32. doi:10.1111/1467-9868.00342.

Jiang, C., and J. L. Wang. 2011. Functional single index models for longitudinal data. *The Annals of Statistics* 39 (1):362–88. doi:10.1214/10-AOS845.

Kokoszka, P., and M. Reimherr. 2017. *Introduction to functional data analysis*. Boca Raton: CRC Press.

Lim, C. 2016. Interval-valued data regression using nonparametric additive models. *Journal of the Korean Statistical Society* 45 (3):358–70. doi:10.1016/j.jkss.2015.12.003.

Luo, R., and X. Qi. 2019. Interaction model and model selection for function-on-function regression. *Journal of Computational and Graphical Statistics* 28 (2):309–22. doi:10.1080/10618600.2018.1514310.

Maia, A., and F. D. Carvalho. 2008. Fitting a least absolute deviation regression model on symbolic interval data. In: *Lecture Notes in Artificial Intelligence: Proceedings of the Ninth Brazilian Symposium on Artificial Intelligence*, 207–16. Berlin: Springer.

Matsui, H., S. Kawano, and S. Konishi. 2009. Regularized functional regression modeling for functional response and predictors. *Journal of Math-for-Industry* 1:17–25.

Müller, H. G. 2005. Functional modelling and classification of functional and longitudinal data. *Scandinavian Journal of Statistics* 32:223–40.

Müller, H. G., and U. Stadtmuller. 2005. Generalized functional linear models. *The Annals of Statistics* 33 (2):774–805. doi:10.1214/009053604000001156.

Neto, E. A. L., G. M. Cordeiro, F. A. T. Carvalho, U. Anjos, and A. Costa. 2009. Bivariate generalized linear model for interval-valued variables. In *Proceedings IEEE International Joint Conference on Neural Networks, Atlanta, USA*, pp. 2226–9.

Neto, E. d. A. L., G. M. Cordeiro, and F. d. A.T. de Carvalho. 2011. Bivariate symbolic regression models for interval-valued variables. *Journal of Statistical Computation and Simulation* 81 (11): 1727–44. doi:10.1080/00949655.2010.500470.

Neto, E. L., and F. de Carvalho. 2008. Center and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis* 52:1500–15.

Neto, E. L., and F. A. T. de Carvalho. 2010. Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics and Data Analysis* 54:333–47.

Neto, E. L., and F. A. T. de Carvalho. 2017. Nonlinear regression applied to interval-valued data. *Pattern Analysis and Applications* 20:809–24.

Neto, E. A. L., F. A. T. de Carvalho, and C. P. Tenorio. 2004. Univariate and multivariate linear regression methods to predict interval-valued features. In: *Lecture Notes in Computer Science and Advances in Artificial Intelligence*, 526–37. Berlin: Springer.

Noirhomme-Fraiture, M., and P. Brito. 2011. Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining* 4 (2):157–70. doi:10.1002/sam.10112.

Ramsay, J. O., and C. Dalzell. 1991. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* 53:539–72. doi:10.1111/j.2517-6161.1991.tb01844.x.

Ramsay, J. O., and B. W. Silverman. 2002. *Applied functional data analysis*. New York: Springer.

Ramsay, J. O., and B. W. Silverman. 2005. *Functional data analysis*. New York: Springer.

Şentürk, D., and H. G. Müller. 2005. Covariate adjusted correlation analysis via varying coefficient models. *Scandinavian Journal of Statistic* 32:365–83.

Şentürk, D., and H. G. Müller. 2008. Generalized varying coefficient models for longitudinal data. *Biometrika* 95:653–66.

Valderrama, M. J., F. A. Ocana, A. M. Aguilera, and F. M. Ocana-Peinado. 2010. Forecasting pollen concentration by a two-step functional model. *Biometrics* 66 (2):578–85.

Yao, F., H. G. Müller, and J. L. Wang. 2005. Functional linear regression analysis for longitudinal data. *The Annals of Statistics* 33 (6):2873–903. doi:10.1214/009053605000000660.

Zhang, Y., J. Thakar, D. J. Topham, A. R. Falsey, D. Zeng, and X. Qiu. 2018. Some equivalence relationships of regularized regressions. *Mathematics for Application* 7 (1):3–10. doi:10.13164/ma.2018.01.