J Syst Sci Syst Eng (Dec 2012) 21(4): 422-442 ISSN: 1004-3756 (Paper) 1861-9576 (Online)
DOI: 10.1007/s11518-012-5203-4 CN11-2983/N

LINEAR REGRESSION OF INTERVAL-VALUED DATA BASED ON COMPLETE INFORMATION IN HYPERCUBES*

Huiwen WANG[1] Rong GUAN[2] Junjie WU[3]

Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations,
School of Economics and Management, Beihang University, Beijing 100191, China
[1]wanghw@vip.sina.com, [2]rongguan77@gmail.com, [3]wujj@buaa.edu.cn (✉)

Abstract

Recent years have witnessed an increasing interest in interval-valued data analysis. As one of the core topics, linear regression attracts particular attention. It attempts to model the relationship between one or more explanatory variables and a response variable by fitting a linear equation to the interval-valued observations. Despite of the well-known methods such as CM, CRM and CCRM proposed in the literature, further study is still needed to build a regression model that can capture the complete information in interval-valued observations. To this end, in this paper, we propose the novel Complete Information Method (CIM) for linear regression modeling. By dividing hypercubes into informative grid data, CIM defines the inner product of interval-valued variables, and transforms the regression modeling into the computation of some inner products. Experiments on both the synthetic and real-world data sets demonstrate the merits of CIM in modeling interval-valued data, and avoiding the mathematical incoherence introduced by CM and CRM.

Keywords: Interval-valued data, linear regression, complete information method (CIM), hypercubes

1. Introduction

The explosion of databases in real-life applications is now calling for more effective statistical tools. Symbolic Data Analysis (Diday 1987, 1989, Bock & Diday 2000, Billard & Diday 2006, Diday & Noirhomme-Fraiture 2008) indicates a promising direction for solving this problem. It firstly either applies a clustering process to a huge database or directly focuses on native symbolic concepts (e.g. biological species), and then adopts symbolic data, such as multi-valued data, interval-valued data, histogram data and the like, to summarize each obtained cluster to a high-level statistical unit. As one category of the most widely used symbolic data, interval-valued data generalizes large-scale numeric data by extracting lower and upper bound values of each feature for each

* This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 71031001, 70771004, 70901002 and 71171007, the Foundation for the Author of National Excellent Doctoral Dissertation of PR China under Grant 201189, and the Program for New Century Excellent Talents in University under Grant NCET-11-0778.

cluster. Consequently, it helps reorganize a large-scale data set into a simple one with fewer statistical units, and thus makes quantitative analysis perform more effectively.

In recent years, multivariate statistics for interval-valued data has become a focus in research (de Souza & de Carvalho 2004, de Carvalho, de Souza et al. 2006, de Carvalho, Brito et al. 2006, Gioia & Lauro 2006, Lauro & Gioia 2006, Silva & Brito 2006). As one of the most popular statistical tools, linear regression aims at explaining one response variable by a linear function of p $(p \geq 1)$ explanatory variables. It has been widely studied in classical regression analysis when observations are described by numeric data (Scheffé 1959, Draper & Smith 1981, Montgomery 1982). In the case of interval-valued data, however, observations are usually represented by solid hypercubes in the space of $(p+1)$ dimensions (Bock & Diday 2000), wherein the original large-scale numeric data reside. The regression line should therefore best fit all observed hypercubes. In what follows, some existing well-known methods are briefly introduced.

Center Method (CM) may be the first attempt (Billard & Diday 2000). The method represents each hypercube by its center point and then computes a regression line that best fits all center points based on Ordinary Least Squares (OLS). The established model actually interprets the positions of center points only, but neglects other information of the observed hypercubes, such as size and shape. Besides, CM predicts values of lower/upper bounds of the response variable by applying the fitted model to values of lower/upper bounds of explanatory variables, which may lead to a mathematical incoherence that the lower bound of the predicted value is greater than the upper bound.

Lima & de Carvalho (2008) argue that the prediction of the interval-valued response variable should be accomplished from not only its center but also its range, and therefore propose Center and Range Method (CRM). Apart from the center equation, a range equation is added to CRM, aiming to interpret the size and shape information of interval-valued observations. CRM shows a significant improvement in prediction when there is a changing rule in ranges between the response variable and explanatory variables. However, since it predicts centers and ranges separately, CRM yet fails to handle the mathematical incoherence introduced by CM.

Constrained Center and Range Method (CCRM) is proposed recently, which adds a nonnegative constraint to coefficients in the range equation (Lima & de Carvalho 2010). Although the mathematical coherence is guaranteed, CCRM may result in biased estimators and therefore performs not as well as CRM in both interpretation and prediction. Consequently, a guideline is suggested that CCRM should not be adopted directly unless CRM "fails to predict the values of the lower and upper boundaries in such a way that $\hat{\underline{y}}_i \leq \hat{\overline{y}}_i$ ", where $\hat{\underline{y}}_i$ and $\hat{\overline{y}}_i$ represent the prediction values of lower and upper boundaries, respectively (Lima & de Carvalho 2010).

There are some other linear regression methods for interval-valued data in the literatures (Billard & Diday 2002, Marino & Palumo 2003, Gioia & Lauro 2005), which we will not cover in details. In general, CM, CRM

and CCRM work well in explaining the variation behavior of center points of hypercubes, and CRM and CCRM can further illustrate the changing rules of ranges. These methods, however, share a common defect that only parts of the hypercube information are employed for the regression modeling. For instance, CM only makes use of the center information, and CRM and CCRM only add the range information based on CM. As a result, further study is still needed to build a regression model that can capture the complete information in interval-valued observations, i.e., hypercubes. This is particularly interesting when the interval-valued observations are the approximation of the large-scale or missing numeric observations. In such a case, we may expect that the resulted linear function from interval-valued observations can best fit the original numeric observations.

Let us illustrate this by the following small example. Figure 1 shows a case of simple linear regression modeling for 8 interval-valued observations described by X and Y. Suppose these interval-valued observations are the approximation of the large-scale numeric observations residing inside the hypercubes, as shown by the uniformly distributed bold dots. Now, if we only use the center points of hypercubes, we will obtain the dashed regression line with circles representing the centers. However, for the bold dots, using the classic linear regression method for numeric data, we will instead obtain the solid regression line, which is obviously different from the dashed one. In other words, the regression line resulted from the center points cannot well fit the original numeric observations. Why does it happen? The reason lies in the fact that the line of center points only describes the changing rule of rectangle centers, while the line of bold dots expresses how Y changes with variation of X by considering the complete information within the 8 hypercubes.

To meet this critical challenge, in this paper, we aim to explore a new linear regression modeling method for interval-valued variables, which can make use of complete information in hypercubes, and can avoid the mathematical incoherence when predicting the response variable using the regression function. Specifically, we propose the novel Complete Information Method (CIM) for linear regression modeling. By dividing hypercubes into informative grid data, CIM defines the inner product of interval-valued variables, and transforms the regression modeling into the computation of some inner products. Experiments on both the synthetic and real-world data sets demonstrate the merits of CIM in modeling interval-valued data, and avoiding the mathematical incoherence introduced by CM and CRM.

The remainder of this paper is organized as

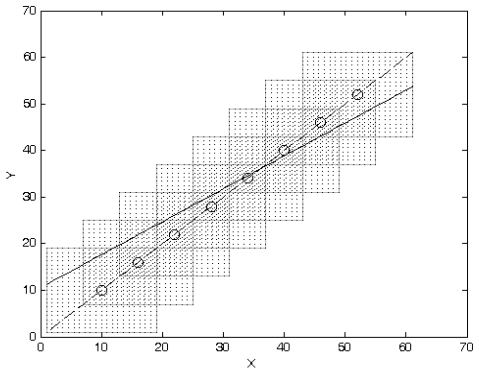

Figure 1 Comparison of regression lines for center points and numeric points

follows. Section 2 introduces CIM after giving several basic operators for interval-valued data. Experimental results on synthetic and real-world data sets are given in Sections 3 and 4, respectively. We finally conclude our work in Section 5.

2. Methodology

We begin by giving some notations of interval-valued data. Let $\mathbf{X}$ be an $n\times p$ interval-valued matrix as follows:

$$\mathbf{X}=\begin{pmatrix} [\underline{x}_{11},\overline{x}_{11}] & [\underline{x}_{12},\overline{x}_{12}] & \cdots & [\underline{x}_{1p},\overline{x}_{1p}] \\ [\underline{x}_{21},\overline{x}_{21}] & [\underline{x}_{22},\overline{x}_{22}] & \cdots & [\underline{x}_{2p},\overline{x}_{2p}] \\ \vdots & \vdots & \ddots & \vdots \\ [\underline{x}_{n1},\overline{x}_{n1}] & [\underline{x}_{n2},\overline{x}_{n2}] & \cdots & [\underline{x}_{np},\overline{x}_{np}] \end{pmatrix} \quad (1)$$

where each column is an interval-valued variable $\mathbf{X}_j\,(j=1,2,\cdots,p)$ with n observations, each row represents an interval-valued observation $\mathbf{O}_i\,(i=1,2,\cdots,n)$ described by p variables, and each data unit x_{ij} is an interval-valued data.

2.1 Basic Operators of Interval-Valued Data

It is widely accepted that each interval-valued observation $\mathbf{O}_i'=([\underline{x}_{i1},\overline{x}_{i1}],\cdots,[\underline{x}_{ip},\overline{x}_{ip}])$ can be viewed as a hypercube in $\mathbb{R}^p$ (Bock & Diday 2000), with infinitely dense points uniformly distributing within it. Figure 2 shows an interval-valued observation in $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively.

If we divide each side of a p-dimensional hypercube into m equal parts, we will obtain m^p small p-dimensional hypercubes, with $(m+1)^p$ non-overlapping vertices in total. These vertices, hereinafter referred to as grid data, are single-valued elements distributing uniformly in the hypercube (see Figures 3 and 4). Clearly, grid data is a discrete form of interval-valued data. As m approaches infinity, grid data conveys the complete information of the hypercube.

Based on the concept of grid data, in what follows, we propose several basic operators for interval-valued data. First, according to Figures 3(a) and 4(a), the mean of an interval-valued data unit $x=[\underline{x},\overline{x}]$ can be defined as

$$E(x)=\frac{1}{m+1}\sum_{k=1}^{m+1}x_k\ . \quad (2)$$

When m tends to infinity, Equation (2) becomes

$$E(x)=\int_{\underline{x}}^{\overline{x}}x\cdot\frac{1}{\overline{x}-\underline{x}}dx=\frac{1}{2}(\underline{x}+\overline{x})\ . \quad (3)$$

Analogously, as shown in Figures 3(a) and 4(a), the squared norm of interval-valued data $x=[\underline{x},\overline{x}]$ can be defined as

$$\|x\|^2=\frac{1}{m+1}\sum_{k=1}^{m+1}x_k^2\ . \quad (4)$$

When m goes to infinity, Equation (4) becomes

$$\|x\|^2=\int_{\underline{x}}^{\overline{x}}x^2\cdot\frac{1}{\overline{x}-\underline{x}}dx=\frac{1}{3}(\underline{x}^2+\underline{x}\overline{x}+\overline{x}^2)\ . \quad (5)$$

Then according to Figures 3(b) and 4(b), the inner product of $x=[\underline{x},\overline{x}]$ and $y=[\underline{y},\overline{y}]$ can be defined as

$$\langle x,y\rangle=\frac{1}{(m+1)^2}\sum_{k=1}^{m+1}\sum_{l=1}^{m+1}x_k y_l\ . \quad (6)$$

When m approaches infinity, Equation (6) becomes

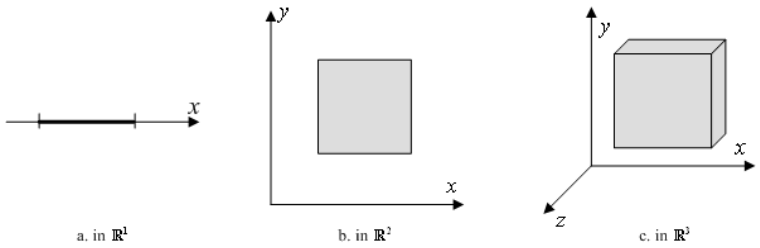

Figure 2 An interval-valued observation in $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$

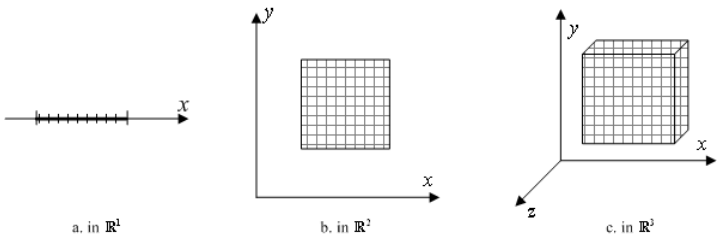

Figure 3 Grid data by $m = 10$

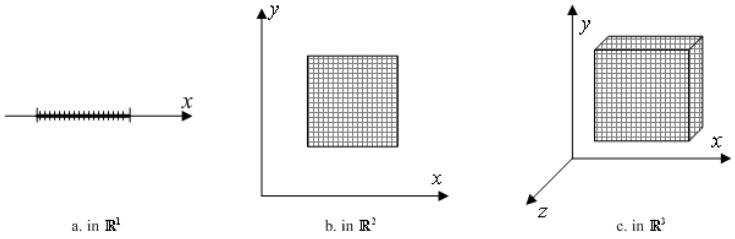

Figure 4 Grid data by $m = 20$

$$\begin{aligned}\langle x, y\rangle &= \int_{\underline{y}}^{\overline{y}}\int_{\underline{x}}^{\overline{x}} xy\cdot\frac{1}{(\overline{x}-\underline{x})(\overline{y}-\underline{y})}dxdy \\ &= \frac{1}{4}(\underline{x}+\overline{x})(\underline{y}+\overline{y}).\end{aligned} \tag{7}$$

Similar to the linear algebra for numeric data, here we define that

$$\langle x, x\rangle = \|x\|^2. \tag{8}$$

Next, we have a proposition regarding the inner product operator as follows:

Proposition 1 *For any interval-valued data x, y, z, the inner product defined by Equation (7) and (8) satisfies*

(i) Positive definiteness, i.e., $\langle x, x\rangle \geq 0$, and the equality holds iff $x = 0$;

(ii) Symmetry, i.e., $\langle x, y\rangle = \langle y, x\rangle$;

(iii) Linearity, i.e.,

$\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$,

$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$, $\forall \alpha \in \mathbb{R}$.

We leave the proof of Proposition 1 to the Appendix for clarity.

2.2 Mean, Inner Product and Norm of Interval-Valued Variables

Definition 1 *For any interval-valued variable*

$\mathbf{X}_j' = \left(\left[\underline{x}_{1j}, \overline{x}_{1j}\right], \left[\underline{x}_{2j}, \overline{x}_{2j}\right], \cdots, \left[\underline{x}_{nj}, \overline{x}_{nj}\right]\right)$, *the mean is given by*

$$E\left(\mathbf{X}_j\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(x_{ij}\right), \tag{9}$$

where $E(x_{ij}) = \frac{1}{2}\left(\underline{x}_{ij} + \overline{x}_{ij}\right)$, *as defined in Equation (3).*

Accordingly, the centralization of $x_{ij} = \left[\underline{x}_{ij}, \overline{x}_{ij}\right]$ is given by

$$x_{ij} - E\left(\mathbf{X}_j\right) = \left[\underline{x}_{ij} - E\left(\mathbf{X}_j\right), \overline{x}_{ij} - E\left(\mathbf{X}_j\right)\right]. \tag{10}$$

Definition 2 *Given any interval-valued variables* $\mathbf{X}_j' = \left([\underline{x}_{1j}, \overline{x}_{1j}], \cdots, [\underline{x}_{nj}, \overline{x}_{nj}]\right)$ *and* $\mathbf{X}_k' = \left([\underline{x}_{1k}, \overline{x}_{1k}], \cdots, [\underline{x}_{nk}, \overline{x}_{nk}]\right), j \neq k$, *the inner product is defined as*

$$\begin{aligned}\left\langle \mathbf{X}_j, \mathbf{X}_k \right\rangle &= \sum_{i=1}^{n} \left\langle x_{ij}, x_{ik} \right\rangle \\ &= \frac{1}{4}\sum_{i=1}^{n} \left(\underline{x}_{ij} + \overline{x}_{ij}\right)\left(\underline{x}_{ik} + \overline{x}_{ik}\right).\end{aligned} \tag{11}$$

Definition 3 *For an interval-valued variable* $\mathbf{X}_j$, *the squared norm is given by*

$$\begin{aligned}\left\langle \mathbf{X}_j, \mathbf{X}_j \right\rangle &= \left\|\mathbf{X}_j\right\|^2 = \sum_{i=1}^{n} \left\|x_{ij}\right\|^2 \\ &= \frac{1}{3}\sum_{i=1}^{n} \left(\underline{x}_{ij}^2 + \underline{x}_{ij}\overline{x}_{ij} + \overline{x}_{ij}^2\right).\end{aligned} \tag{12}$$

When $\mathbf{X}_j$ and $\mathbf{X}_k$ reduce to numeric vectors, Equations (11) and (12) reduce to the well-known standard form accordingly as follows:

$$\left\langle \mathbf{X}_j, \mathbf{X}_k \right\rangle = \sum_{i=1}^{n} x_{ij} x_{ik} \ . \tag{13}$$

It is interesting to note that Equations (11) and (12) correspond to the proposed covariance (Billard & Diday 2003) and variance (Bertrand & Goupil 2000) of interval-valued variables, respectively, given that the interval-valued data have been centralized initially. Moreover, we have the proposition as follows:

Proposition 2 *For any interval-valued variables* $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$, *the inner product defined by Equations (11) and (12) satisfies*

(i) Positive definiteness, i.e., $\left\langle \mathbf{X}, \mathbf{X} \right\rangle \geq 0$, *and the equality holds iff* $\mathbf{X} = \mathbf{0}$, *where* $\mathbf{0} = (0, 0, \cdots, 0)'$ *represents the zero vector;*

(ii) Symmetry, i.e., $\left\langle \mathbf{X}, \mathbf{Y} \right\rangle = \left\langle \mathbf{Y}, \mathbf{X} \right\rangle$;

(iii) Linearity, i.e.,

$\left\langle \mathbf{X} + \mathbf{Y}, \mathbf{Z} \right\rangle = \left\langle \mathbf{X}, \mathbf{Z} \right\rangle + \left\langle \mathbf{Y}, \mathbf{Z} \right\rangle$,

$\left\langle \alpha\mathbf{X}, \mathbf{Y} \right\rangle = \alpha\left\langle \mathbf{X}, \mathbf{Y} \right\rangle, \forall \alpha \in \mathbb{R}$.

We leave the proof of Proposition 2 to the Appendix for clarity.

2.3 Linear Combination Algorithm for Interval-Valued Variables

In order to avoid the problem that the predicted lower bound values of response variable are greater than the upper bound values, we adopt Moore's linear combination algorithm (Moore 1966) as follows:

Definition 4 *Given any interval-valued variables* $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$ *of n observations and* $\forall a_j \in \mathbb{R}$, $j = 1, 2, \cdots, p$, *define an interval-valued variable* $\mathbf{Y}$ *as the linear function of* $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p$, *i.e.,*

$$\begin{aligned}\mathbf{Y} &= \sum_{j=1}^{p} a_j \mathbf{X}_j \\ &= \left(\left[\underline{y}_1, \overline{y}_1\right], \left[\underline{y}_2, \overline{y}_2\right], \cdots, \left[\underline{y}_n, \overline{y}_n\right]\right)',\end{aligned} \tag{14}$$

where

$$\underline{y}_i = \sum_{j=1}^{p} a_j \left[\tau\underline{x}_{ij} + (1-\tau)\overline{x}_{ij}\right], \tag{15}$$

$$\overline{y}_i = \sum_{j=1}^{p} a_j \left[(1-\tau)\underline{x}_{ij} + \tau\overline{x}_{ij}\right], \tag{16}$$

with

$$\tau = \begin{cases} 0, & \text{if } a_j \le 0 \\ 1, & \text{otherwise.} \end{cases} \tag{17}$$

Clearly, Moore's algorithm guarantees the mathematical coherence of $\underline{y}_i \le \overline{y}_i (i = 1, 2, \cdots, n)$ by using the indicative function τ. Two methods of PCA on interval-valued data, i.e., Vertices Principal Component Analysis and Center Principal Component Analysis (Cazes et al. 1997), have also adopted this algorithm. The well-known Center Method (CM), however, simply puts the lower/upper bound values of explanatory variables into the center equation to obtain the lower/upper bound values of the response variable, and therefore cannot guarantee the mathematical coherence.

2.4 Linear Regression Model based on Complete Information in Hypercubes

Suppose there exists a linear regression relationship between interval-valued variables $\mathbf{Y}$ and $\mathbf{X}_j \left(j = 1, 2, \cdots, p \right)$, i.e.,

$$\mathbf{Y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \cdots + \beta_p \mathbf{X}_p + \boldsymbol{\varepsilon}, \tag{18}$$

where $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \right)'$ is the numeric residual error, and $\mathbf{1} = \left(1, 1, \cdots, 1 \right)'$ is a constant vector. The equation holds according to Definition 4, since both $\boldsymbol{\varepsilon}$ and $\mathbf{1}$ could be treated as interval-valued vectors.

Geometrically, the model established in Equation (18) views each interval-valued observation as a whole, i.e., a hypercube, rather than separates it in parts (center and range) as CRM/CCRM does. This is considered valuable, since in many real-world cases there is no significant linear relationship between the range variables, in which CIM still works yet CRM/CCRM will fail. The first case in Section 4 just gives such an example, where the linear relationship between the ranges of the concerning interval-valued variables, i.e., *Age* $(\mathbf{X})$ and *Cholesterol* $(\mathbf{Y})$, does not exist since the range of *Age* is fixed while the range of *Cholesterol* varies among observations.

In what follows, we estimate $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \cdots, \beta_p \right)'$ given the operators defined above. Let $\mathbf{B} = \left(b_0, b_1, \cdots, b_p \right)'$ be the estimator of $\boldsymbol{\beta}$, we have

$$\hat{\mathbf{Y}} = b_0 \mathbf{1} + b_1 \mathbf{X}_1 + b_2 \mathbf{X}_2 + \cdots + b_p \mathbf{X}_p . \tag{19}$$

As a result, the sum of squared error is

$$\begin{aligned} SSE &= \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|^2 \\ &= \left\langle \mathbf{Y} - b_0 \mathbf{1} - \sum_{l=1}^{p} b_l \mathbf{X}_l, \mathbf{Y} - b_0 \mathbf{1} - \sum_{l=1}^{p} b_l \mathbf{X}_l \right\rangle \\ &\overset{(*)}{=} \left\langle \mathbf{Y}, \mathbf{Y} \right\rangle + b_0^2 \left\langle \mathbf{1}, \mathbf{1} \right\rangle - 2 b_0 \left\langle \mathbf{1}, \mathbf{Y} \right\rangle \\ &\quad + 2 b_0 \sum_{l=1}^{p} b_l \left\langle \mathbf{1}, \mathbf{X}_l \right\rangle - 2 \sum_{l=1}^{p} b_l \left\langle \mathbf{Y}, \mathbf{X}_l \right\rangle \\ &\quad + \sum_{l=1}^{p} \sum_{k=1}^{p} b_l b_k \left\langle \mathbf{X}_l, \mathbf{X}_k \right\rangle, \end{aligned} \tag{20}$$

where $(*)$ holds due to the linearity of the inner product in Proposition 2. Furthermore, according to Ordinary Least Squares (OLS), we take partial derivatives of *SSE* w.r.t. $b_0, b_1, \cdots, b_p$, respectively, and let them be zero as follows:

$$\begin{aligned} \frac{\partial SSE}{\partial b_0} &= 2 b_0 \left\langle \mathbf{1}, \mathbf{1} \right\rangle - 2 \left\langle \mathbf{1}, \mathbf{Y} \right\rangle \\ &\quad + 2 \sum_{l=1}^{p} b_l \left\langle \mathbf{1}, \mathbf{X}_l \right\rangle = 0, \end{aligned} \tag{21}$$

$$\frac{\partial SSE}{\partial b_j} = 2b_0\left\langle \mathbf{1}, \mathbf{X}_j \right\rangle - 2\left\langle \mathbf{Y}, \mathbf{X}_j \right\rangle + 2\sum_{l=1}^{p} b_l \left\langle \mathbf{X}_j, \mathbf{X}_l \right\rangle = 0, \tag{22}$$

for $j = 1, 2, \cdots, p$.

Equations (21) and (22) can be rewritten in a matrix form as follows:

$$\begin{pmatrix} \langle \mathbf{1},\mathbf{1}\rangle & \langle \mathbf{1},\mathbf{X}_1\rangle & \cdots & \langle \mathbf{1},\mathbf{X}_p\rangle \\ \langle \mathbf{X}_1,\mathbf{1}\rangle & \langle \mathbf{X}_1,\mathbf{X}_1\rangle & \cdots & \langle \mathbf{X}_1,\mathbf{X}_p\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{X}_p,\mathbf{1}\rangle & \langle \mathbf{X}_p,\mathbf{X}_1\rangle & \cdots & \langle \mathbf{X}_p,\mathbf{X}_p\rangle \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} \langle \mathbf{1},\mathbf{Y}\rangle \\ \langle \mathbf{X}_1,\mathbf{Y}\rangle \\ \vdots \\ \langle \mathbf{X}_p,\mathbf{Y}\rangle \end{pmatrix}. \tag{23}$$

Obviously, elements in the above matrix are all numeric values. Therefore, we are easily led to the solution $\mathbf{B} = (b_0, b_1, \cdots, b_p)'$ and the corresponding regression function $\hat{\mathbf{Y}} = b_0\mathbf{1} + b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \cdots + b_p\mathbf{X}_p$. We call this method the Complete Information Method (CIM), since it employs all the information in hypercubes to establish the regression model. Given any interval-valued vector, i.e., $([\underline{x}_{i1}, \overline{x}_{i1}], \cdots, [\underline{x}_{ip}, \overline{x}_{ip}])'$, CIM achieves the predicted value $[\underline{\hat{y}}_i, \overline{\hat{y}}_i]$ according to the obtained regression function and Definition 4.

3. Experimental Results of Synthetic Data Sets

In this section, we conduct extensive experiments on synthetic data sets to demonstrate the unique properties of CIM. Specifically, we show that CIM well interprets the relationship between the response variable and explanatory variables. Moreover, if we use interval-valued data to approximate mass numeric data, CIM will be an excellent choice for the data prediction.

3.1 Data

We firstly describe how to generate interval-valued observations in Monte Carlo experiments. Some key points are as follows:

(i) The observations of the single explanatory variable $\mathbf{X}$ and the response variable $\mathbf{Y}$ are generated by setting the center variables $\mathbf{X}^C = (x_1^C, \cdots, x_n^C)'$ and $\mathbf{Y}^C = (y_1^C, \cdots, y_n^C)'$, and the range variables $\mathbf{X}^R = (x_1^R, \cdots, x_n^R)'$ and $\mathbf{Y}^R = (y_1^R, \cdots, y_n^R)'$ of the observations, respectively. Initially, we randomly generate $\mathbf{X}^C$ according to a uniform distribution of $U[-5, 5]$. Next, in two different ways will other three variables be constructed.

• *Type I*: We have $\mathbf{Y}^C = \beta_0 + \beta_1\mathbf{X}^C + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'$ represents the residual error vector, and β_0 and β_1 are equation coefficients, both following $U[-1, 1]$. Range variables $\mathbf{X}^R$ and $\mathbf{Y}^R$ will then be randomly generated independently of the corresponding center variables $\mathbf{X}^C$ and $\mathbf{Y}^C$. The upper part of Table 1 has listed the distributions of $\mathbf{X}^R$, $\mathbf{Y}^R$ and $\boldsymbol{\varepsilon}$, respectively.

• *Type II*: Similarly, $\mathbf{Y}^C$ is computed by $\mathbf{Y}^C = \beta_0 + \beta_1\mathbf{X}^C + \boldsymbol{\varepsilon}$, where β_0 and β_1 still follow $U[-1, 1]$, while the residual error vector $\boldsymbol{\varepsilon} \sim N(0, 0.5^2)$. Then, data generation will be based on a linear relationship between center

Table 1 Twelve configurations for Monte Carlo experiments

Type	*Configs.*	$\mathbf{X}^R$	$\mathbf{Y}^R$	$\boldsymbol{\varepsilon}$
I	C_1	U[0.5,0.8]	U[0.5,0.8]	$N(0,0.5^2)$
	C_2	U[0.5,0.8]	U[0.5,0.8]	$N(0,1^2)$
	C_3	U[1,3]	U[1,3]	$N(0,0.5^2)$
	C_4	U[1,3]	U[1,3]	$N(0,1^2)$
	C_5	U[5,8]	U[5,8]	$N(0,0.5^2)$
	C_6	U[5,8]	U[5,8]	$N(0,1^2)$
Type	*Configs.*	β_0^*	β_1^*	$\boldsymbol{\varepsilon}^*$
II	C_7	U[0.1,0.16]	U[0.1,0.16]	$N(0,1^2)$
	C_8	U[0.1,0.16]	U[0.1,0.16]	$N(0,2^2)$
	C_9	U[0.2,0.6]	U[0.2,0.6]	$N(0,1^2)$
	C_{10}	U[0.2,0.6]	U[0.2,0.6]	$N(0,2^2)$
	C_{11}	U[1,1.6]	U[1,1.6]	$N(0,1^2)$
	C_{12}	U[1,1.6]	U[1,1.6]	$N(0,2^2)$

variables and range variables. More specifically, we compute $\mathbf{X}^R = \beta_0^\star + \beta_1^\star \mid \mathbf{X}^C \mid + \boldsymbol{\varepsilon}^\star$ and $\mathbf{Y}^R = \beta_0^\star + \beta_1^\star \mid \mathbf{Y}^C \mid + \boldsymbol{\varepsilon}^\star$, where $\beta_0^\star$ and $\beta_1^\star$ are the regression parameters and $\boldsymbol{\varepsilon}^\star = \left(\varepsilon_1^\star, \varepsilon_2^\star, \cdots, \varepsilon_n^\star\right)'$ represents the residual error vector. Distributions of $\beta_0^\star$, $\beta_1^\star$ and $\boldsymbol{\varepsilon}^\star$ have been shown in the lower part of Table 1.

(ii) The interval-valued data $\mathbf{X}' = \left([\underline{x}_1, \overline{x}_1], \cdots, [\underline{x}_n, \overline{x}_n]\right)$ and $\mathbf{Y}' = \left([\underline{y}_1, \overline{y}_1], \cdots, [\underline{y}_n, \overline{y}_n]\right)$ will then be constructed by having $\underline{x}_i = x_i^C - x_i^R$, $\overline{x}_i = x_i^C + x_i^R$, $\underline{y}_i = y_i^C - y_i^R$ and $\overline{y}_i = y_i^C + y_i^R$ $(i = 1, 2, \cdots, n)$, respectively.

(iii) To study the impact of parameter settings, we use 12 configurations for Monte Carlo experiments, as shown in Table 1. For instance, observations under C_1 will be those in small size and with small residual errors, while observations under C_6 correspond to hypercubes in much larger size and with higher residual errors. In fact, an association exists between configurations in the two types. The parameters $\beta_0^\star$ and $\beta_1^\star$ in *Type II* helps determine observation size, which depends on values of $\mathbf{X}^R$ and $\mathbf{Y}^R$ in *Type I*. As consequence, observations in each configuration pair (eg. C_1 and C_7, C_6 and C_{12}) are likely to appear in similar size.

(iv) For each configuration, N interval-valued observations will be generated. The first N_1 observations will work as a training set for building regression models, while the rest N_2 ones will be used as a test set. For our experiments, we have $N_1 = 250$, $N_2 = 125$, and $N = N_1 + N_2$.

(v) To simulate real-life numeric data of large scale, we further generate numeric data uniformly distributing within the interval-valued observations. That is, for the i-th observation $\left(i = 1, 2, \cdots, N\right)$, M numeric data points $\left(x_{i_l}^{NUM}, y_{i_l}^{NUM}\right)$ will be generated, with $x_{i_l}^{NUM} \sim U\left[\underline{x}_i, \overline{x}_i\right]$ and $y_{i_l}^{NUM} \sim U\left[\underline{y}_i, \overline{y}_i\right]$, $l = 1, 2, \cdots, M$. By this way, we will generate $N \times M$ numeric data points, which can be

viewed as the original data we aim to approximate using the interval-valued data. Note that to avoid biased comparison results, for each configuration, we repeat the experiments K times. For our experiments, we have $M = 100$ and $K = 100$.

3.2 Indicators

The indicator we use for interpretation assessment is Coefficient of Dissimilarity (CD), defined as

$$CD = \left| \frac{b_0^{NUM} - b_0}{b_0^{NUM}} \right| + \left| \frac{b_1^{NUM} - b_1}{b_1^{NUM}} \right|, \quad (24)$$

where $(b_0^{NUM}, b_1^{NUM})'$ is the regression coefficients for the numeric data, and $(b_0, b_1)'$ is the regression coefficients for the corresponding interval-valued observations. Apparently, CD takes $(b_0^{NUM}, b_1^{NUM})'$ as the benchmark for $(b_0, b_1)'$. A smaller CD indicates a better interpretability of the regression model.

For prediction assessment, we use two groups of indicators. The first group only contains one indicator: Absolute-Mean Error (AME), which is defined as

$$AME = \frac{1}{N_2 \times M} \sum_{i=N_1+1}^{N} \sum_{l=1}^{M} \left| y_{i_l}^{NUM} - \hat{y}_{i_l} \right|, \quad (25)$$

where $\hat{y}_{i_l}$ can be replaced by $\hat{y}_{i_l}^{CM} = b_0^{CM} + b_1^{CM} x_{i_l}^{NUM}$ or $\hat{y}_{i_l}^{CIM} = b_0^{CIM} + b_1^{CIM} x_{i_l}^{NUM}$. Apparently, the regression model with a lower AME value will have a better prediction performance. The numeric data $(x_{i_l}^{NUM},\ y_{i_l}^{NUM})\quad (1 \le l \le M,\quad N_1 + 1 \le i \le N)$ again works as a benchmark. Note that CRM and CCRM will not be involved in the comparison when using indicators CD and AME, since range equations will not appear in regression models for numeric data, and the center equations are the same for CM, CRM and CCRM.

The second group of indicators consists of the lower boundary of Root-Mean-Square Error ($RMSE_L$) and the upper boundary of Root-Mean-Square Error ($RMSE_U$) as follows (Lima & de Carvalho 2008):

$$RMSE_L = \sqrt{\frac{1}{N_2} \sum_{i=N_1+1}^{N} \left(\underline{y}_i - \underline{\hat{y}}_i \right)^2}, \quad (26)$$

$$RMSE_U = \sqrt{\frac{1}{N_2} \sum_{i=N_1+1}^{N} \left(\overline{y}_i - \overline{\hat{y}}_i \right)^2}, \quad (27)$$

where $\underline{\hat{y}}_i$ and $\overline{\hat{y}}_i$ represent the lower and upper bounds of the predicted values of the response variable, respectively. The lower the two indicators, the better the regression model. Apparently, $RMSE_L$ and $RMSE_U$ pay more attention to the boundaries of intervals, whereas AME takes the complete information in hypercubes into consideration.

3.3 Procedure

The experiments are conducted in the framework of Monte Carlo simulations as follows:

(i) For each configuration in Table 1, the experiment will be repeated K times. In the k-th replication $(1 \le k \le K)$, adopt CM, CRM, CCRM and CIM respectively to build linear regression models with the training set. Build a linear regression model for the corresponding numeric data by the classic method.

(ii) For each configuration, calculate the CD

values of the results by CM and CIM, the averaged *CD* value, i.e., $\overline{CD}$, and the frequency ratio of $CD(CIM) \le CD(CM)$ in K replications.

(iii) For each configuration, using the test set to calculate the *AME* values of the results by CM and CIM, the averaged *AME* value, i.e., $\overline{AME}$, and the frequency ratio of $AME(CIM) \le AME(CM)$ in K replications.

(iv) For each configuration, using the test set to calculate both the $RMSE_L$ and $RMSE_U$ values of the results by CM, CIM, CRM and CCRM, and the averaged values of $RMSE_L$ and $RMSE_U$ in K replications, i.e., $\overline{RMSE_L}$ and $\overline{RMSE_U}$.

3.4 Results

3.4.1 Comparison of Interpreting Performance

Here we use *CD* to measure the interpreting performance of CM and CIM. The experimental results are shown in Table 2.

As indicated by Table 2 for most configurations, $\overline{CD}(CIM)$ are much smaller than $\overline{CD}(CM)$, with the only exception: C_2. Also, the case of $CD(CIM) \le CD(CM)$ prevails in most of the replications, as indicated by the high frequency ratio (over 85%) for all configurations. These results strongly indicate that the regression coefficients obtained by CIM are much closer to the coefficients for the numeric data. In other words, CIM has a higher interpreting power than CM when interval-valued data is used to approximate original massive numeric data.

If we take a closer look at Tables 1 and 2, three findings are also notable as follows. First, it helps to improve the superiority of CIM over CM to assume a linear relationship between center and range variables in data generation process, especially when the advantage of CIM over CM is not very significant. This indicates that CIM does better in unveiling the association between centers and ranges in the regression coefficients, yet CM fails in recognizing such information. Besides, observation size has a strong impact to the comparison result. That is, with the increase of the observation size from configuration pair C_1 / C_7 to C_5 / C_{11}, the edge of CIM over CM tends to be more significant, as indicated by the increase of the frequency ratio values. This is due to the reason that CIM captures the complete information in hypercubes while CM does not. As a result, as the observation size increases, more information will be omitted by CM, which eventually leads to the downgrade of interpreting performance. The third observation is that, although the residual error $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^{\star}$ have some impact to the regression accuracy, it does not change the superiority of CIM to CM in terms of the interpreting power.

3.4.2 Comparison of Predicting Performance

Here we compare the predicting power of the regression methods using the test sets. The evaluation results of *AME*, $RMSE_L$ and $RMSE_U$ are shown in Tables 3 and 4, respectively.

As indicated by Table 3, $\overline{AME}(CIM)$ is generally smaller than $\overline{AME}(CM)$, and the gap tends to be wider with the increase of the observation size. This is, CIM has a higher predicting power than CM, especially when the hypercubes of observations appear in relatively

Table 2 Comparison between CM and CIM: $\overline{CD}$ and the frequency ratio of $CD(CIM) \le CD(CM)$ in 100 replications in the framework of a Monte Carlo experiment

Type	*Configs.*	$\overline{CD}(CM)$	$\overline{CD}(CIM)$	*Ratio (%)*
I	C_1	0.034	0.028	86
	C_2	0.048	0.068	86
	C_3	0.217	0.058	97
	C_4	0.242	0.070	99
	C_5	1.923	0.156	99
	C_6	1.988	0.296	99
Type	*Configs.*	$\overline{CD}(CM)$	$\overline{CD}(CIM)$	*Ratio (%)*
II	C_7	0.066	0.028	92
	C_8	0.144	0.033	98
	C_9	0.220	0.058	97
	C_{10}	0.346	0.062	97
	C_{11}	1.392	0.086	100
	C_{12}	1.621	0.126	100

large size. The frequency ratio of $AME(CIM) \le AME(CM)$ provides further evidences along this line. As can be seen in Table 3, although the ratio is relatively low for the first configuration pair, it approaches nearly 100% when the observation size increases for the last configuration pair.

Comparing two configurations of each pair, we find out that the assumed linear relationship between center and range variables helps CIM outperform CM in prediction, with the only exceptional pair of C_3 and C_9. This observation is more significant when the edge of CIM over CM is relatively low, which is similar to results shown in Table 2. Indeed, the improvement benefits from the fact that CIM well captures the simulated relationship between the response variable and the explanatory variable, which is originally contributed by employing complete information within hypercubes.

For indicators $RMSE_L$ and $RMSE_U$, results are more complicated. As can be seen in Table 4, although CIM still shows better performance than CM, this performance is clearly inferior to the performance of CRM and CCRM, regardless of configurations. This situation is even worse when the observation size become larger. To understand this, recall that CIM uses the complete information in hypercubes. In other words, CIM does not cast higher weights to the boundaries of intervals, which are right the focus of $RMSE_L$ and $RMSE_U$. In contrast, CRM and CCRM pay special attention to the prediction of the ranges, or equivalently, the boundaries. As a result, CRM and CCRM are more easily to get better evaluation scores from $RMSE_L$ and $RMSE_U$.

Overall, we could conclude that there is no perfect method in terms of prediction. The indicator of *AME* takes the complete information in hypercubes as a benchmark, whereas $RMSE_L$ and $RMSE_U$ pay more attention to the boundaries of intervals. Though performing slightly worse than CRM/CCRM regarding *RMSE*, CIM shows its superiority in terms of *AME*.

Table 3 Comparison between CM and CIM: $\overline{AME}$ and the frequency ratio of $AME(CIM) \le AME(CM)$ in 100 replications in the framework of a Monte Carlo experiment

Type	*Configs.*	$\overline{AME}(CM)$	$\overline{AME}(CIM)$	*Ratio (%)*
I	C_1	0.538	0.538	57
	C_2	0.881	0.881	57
	C_3	1.202	1.184	92
	C_4	1.394	1.378	84
	C_5	4.266	4.107	96
	C_6	4.395	4.186	96
Type	*Configs.*	$\overline{AME}(CM)$	$\overline{AME}(CIM)$	*Ratio (%)*
II	C_7	0.636	0.633	77
	C_8	0.865	0.857	84
	C_9	0.930	0.930	90
	C_{10}	1.195	1.164	95
	C_{11}	2.163	1.968	99
	C_{12}	2.487	2.258	98

Table 4 Comparison between CM, CRM, CCRM and CIM: $\overline{RMSE_L}$ and $\overline{RMSE_U}$ in 100 replications in the framework of a Monte Carlo experiment

Type	*Configs.*	$\overline{RMSE_L}$				$\overline{RMSE_U}$			
		CM	*CRM*	*CCRM*	*CIM*	*CM*	*CRM*	*CCRM*	*CIM*
I	C_1	0.820	0.509	0.509	0.642	0.828	0.510	0.510	0.651
	C_2	1.284	1.007	1.007	1.064	1.286	1.007	1.007	1.071
	C_3	2.217	0.770	0.776	1.498	2.211	0.766	0.771	1.487
	C_4	2.343	1.156	1.161	1.675	2.355	1.167	1.169	1.700
	C_5	6.773	1.000	1.028	5.460	6.770	0.998	1.029	5.474
	C_6	6.272	1.324	1.355	5.497	6.278	1.321	1.345	5.484
II	C_7	1.026	0.510	0.510	0.631	1.021	0.512	0.512	0.630
	C_8	1.493	0.511	0.512	0.896	1.497	0.522	0.523	0.903
	C_9	1.920	0.580	0.580	0.908	1.925	0.588	0.588	0.917
	C_{10}	2.190	0.621	0.621	1.301	2.201	0.628	0.628	1.317
	C_{11}	4.350	0.938	0.938	2.871	4.326	0.933	0.933	2.835
	C_{12}	4.946	1.025	1.025	3.376	4.949	1.030	1.030	3.400

3.4.3 Discussions

The above results indicate that, under the hypercube premise of interval-valued observations, CIM well grasps the linear relationship between the response variable and explanatory variables, but works no better than CRM or CCRM in predicting the boundaries of intervals. Consequently, a suitable method should be selected according to actual situations when establishing linear regression models for interval-valued data. We hereby propose some strategies in different scenarios as follows:

(a) If analysts attempt to use interval-valued data to approximate massive numeric data in the

regression, CIM should be adopted, as it makes use of the complete information in hypercubes.

(b) If analysts want to explain only the variation behavior of the center points of hypercubes, CM will be a better choice.

(c) If the observations are measuring results with measuring errors, analyst should adopt CRM, which can help to discover changing rules of error ranges.

(d) CCRM should be employed only when CRM fails to predict the values of the lower and upper boundaries and leads to $\underline{\hat{y}}_i > \hat{\overline{y}}_i$.

4. Experimental Results of Real Data Sets

In this section, we use two real data sets to show the performance of CIM in real-world applications. Specifically, we aim to illustrate that CIM provides a linear regression modeling that best fits the real-world numeric data.

4.1 Cholesterol-Age Data Set

The first data set is the well-known Cholesterol-Age data set (Billard & Diday 2006). As shown by Table 5, this data set contains the interval-valued records of *Age* (**X**) and *Cholesterol* (**Y**) for a certain population. Each record consists of various individuals of the same age decade, and the number of individuals is listed in the right-most column of the table.

Assume there exists a linear relationship between $\mathbf{Y}$ and $\mathbf{X}$, i.e., $\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X} + \boldsymbol{\varepsilon}$. We first decide which method should be used to model this data set. It is interesting to note that, in Table 5, the *Cholesterol* range changes among different observations while the *Age* range is fixed to 5. This implies that the linear relationship between the ranges of $\mathbf{Y}$ and $\mathbf{X}$ does not exist. As a result, CRM and CCRM should not be employed for the regression modeling. Apparently, this real-life case is just the example strongly against the application of CRM/CCRM. This, indeed, has well supported the virtue of CIM, i.e., modeling each interval-valued observation as a whole instead of handling it separately in center and range. In what follows, we only present the modeling results of CM and CIM.

Table 6 shows the estimated regression coefficients by CM and CIM, respectively. Obviously, both of the two models indicate that the variation of *Age* gives positive contributions to the variation of *Cholesterol* level. This well illustrates that CIM and CM indeed have some commonplaces in general. As a result, in what follows, we focus on their differences.

As mentioned above, each interval-valued observation in Table 5 is actually the summary

Table 5 Cholesterol-Age data set

Obs. Group	*Cholesterol* **Y**	*Age* **X**	*Num. of individuals*
1	[114,192]	[20,30)	43
2	[103,189]	[30,40)	66
3	[120,191]	[40,50)	75
4	[136,223]	[50,60)	43
5	[149,234]	[60,70)	59
6	[142,229]	[70,80)	35
7	[140,254]	[80,90)	18

Table 6 Estimated regression coefficients by CM and CIM

Methods	$\hat{\beta}_0$	$\hat{\beta}_1$
CM	124.054	0.882
CIM	125.044	0.864

of numeric data of individuals in the same age decade. Therefore, one may expect that the regression model can be used to predict the *Cholesterol* level of any individual rather than the group. CIM, as discussed in Section 3.4.3, will be a better choice than CM in this case, since it makes use of the complete information in hypercubes.

To illustrate this, we divide each interval-valued observation into $(m+1)^p$ equal parts, and use the grid data to simulate the individuals' numeric data. We denote the regression coefficients for the grid data as $\hat{\boldsymbol{\beta}}^{GRID}(m)$ for different values of m, and compare $\hat{\boldsymbol{\beta}}^{GRID}(m)$ with $\hat{\boldsymbol{\beta}}^{CM}$ and $\hat{\boldsymbol{\beta}}^{CIM}$ listed in Table 6.

Figure 5 shows the comparison results. As can be seen in the figure, compared with $\hat{\beta}_0^{CM}$ ($\hat{\beta}_1^{CM}$), $\hat{\beta}_0^{CIM}$ ($\hat{\beta}_1^{CIM}$) is indeed much closer to $\hat{\beta}_0^{GRID}$ ($\hat{\beta}_1^{GRID}$) for any given m. Moreover, as the increase of m, $\hat{\beta}_0^{GRID}$ ($\hat{\beta}_1^{GRID}$) shows a trend more and more consistent with $\hat{\beta}_0^{CIM}$ ($\hat{\beta}_1^{CIM}$).

Surely, the grid data is just a simulation of the real individual data, which we do not know except for the number of individuals in each group. Nonetheless, this real example still demonstrates that if we aim to use interval-valued data to approximate numeric data, CIM is a better choice for the regression modeling.

4.2 Basketball Data Set

In this subsection, a real-world case with numeric data is used to illustrate the effectiveness of CIM even if the uniform distribution assumption does not hold. In this case, three variables, i.e., performance of basketball players on *Points* per minute $(\mathbf{Y})$, *Assists* per minute $(\mathbf{X}_1)$ and *Time Played* $(\mathbf{X}_2)$, have been concerned (Simonoff 1994). The original data set involves numeric records of 96 players.

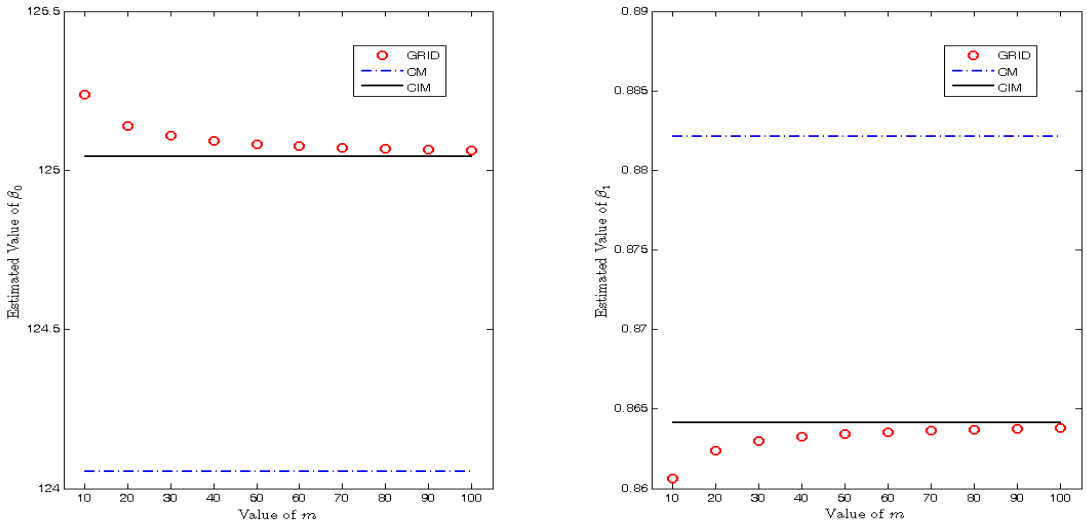

Figure 5 Comparison between CM and CIM using the grid data

To obtain the interval-valued data set, we firstly classify all players into 13 groups by age, and then summarize records of each group by interval-valued data, with the minimum or maximum value of each group respectively being the lower or upper bound of intervals (see Table 7). Note that in this case, the numeric data no longer distribute uniformly within the hypercubes of interval-valued data.

Suppose $\mathbf{Y}$ is linearly correlated with $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e., $\mathbf{Y} = \beta_0 + \beta_1\mathbf{X}_1 + \beta_2\mathbf{X}_2 + \boldsymbol{\varepsilon}$. We firstly establish a linear regression model for the original numeric data by classical method. CM, CIM, CRM and CCRM are also used for the regression modeling. Denote the estimated regression coefficients for numeric data as $\hat{\beta}_j^{NUM}$ $(j = 0,1,2)$, which is the benchmark for the modeling results of interval-valued data. In what follows, we aim to find out which method is better for the linear regression modeling of interval-valued data, under the premise that interval-valued data is an approximation of the original numeric data.

We use the indicators *CD*, *AME*, $RMSE_L$ and $RMSE_U$ for the comparison of interpretation and prediction performances. Results are shown in Table 8. Note that CRM and CCRM obtain the same values as CM in *CD* and *AME*, since their center equations are identical. As can be seen, although the numeric data do not distribute uniformly within the hypercubes, CIM still performs better than CM, CRM and CCRM in terms of *CD* and *AME*. In other words, the regression model generated by CIM gets closer to the one for numeric data even if the uniform distribution assumption does not hold. If we turn to the indicators $RMSE_L$ and $RMSE_U$, we will find that both CRM and CCRM outperform CIM. This, however, is not surprising. As has been pointed out in the conclusion of Section 3.4.2, CRM or CCRM is more concerned with the prediction of boundaries, which is right the focus of $RMSE_L$ and $RMSE_U$. In contrast, CIM attempts to make the best use of the information

Table 7 Interval-valued basketball data set

Age	*Points per min.* $\mathbf{Y}$	*Assists per min.* $\mathbf{X}_1$	*Time played* $\mathbf{X}_2$	*Num. of Individuals*
under 23	[0.2683,0.5437]	[0.0528,0.2244]	[11.81,36.55]	9
24	[0.2381,0.5668]	[0.1010,0.2282]	[10.08,33.88]	10
25	[0.3004,0.5059]	[0.0805,0.2495]	[12.63,35.22]	10
26	[0.2719,0.5769]	[0.0747,0.2383]	[17.41,38.80]	8
27	[0.2578,0.5523]	[0.0728,0.2681]	[17.46,39.53]	12
28	[0.2894,0.5885]	[0.0888,0.2771]	[18.49,38.40]	12
29	[0.4007,0.6244]	[0.1227,0.2521]	[27.87,38.43]	4
30	[0.3498,0.8291]	[0.0896,0.2130]	[12.24,40.71]	10
31	[0.2185,0.5835]	[0.0550,0.3437]	[12.12,34.91]	7
32	[0.1593,0.6318]	[0.0494,0.2327]	[13.37,36.52]	6
33	[0.2406,0.4035]	[0.1317,0.1528]	[16.36,17.46]	2
34	[0.3890,0.6318]	[0.0898,0.1236]	[13.37,28.81]	4
over 35	[0.2471,0.2989]	[0.1668,0.2127]	[14.38,14.57]	2

Table 8 CD, AME, $RMSE_L$ and $RMSE_U$ of the results by CM, CRM, CCRM and CIM

Methods	*CM*	*CRM*	*CCRM*	*CIM*
CD	1.726	1.726	1.726	1.195
AME	0.079	0.079	0.079	0.070
$RMSE_L$	0.145	0.056	0.055	0.104
$RMSE_U$	0.144	0.066	0.066	0.127

inside the hypercubes, and therefore is better at approximating the regression results of numeric data, which is exactly the main theme of this case.

To further understand the differences between these methods, we shall pay attention to the regression results for numeric data and for interval-valued data, respectively, by CM, CRM, CCRM and CIM (see Table 9).

Apparently, all the equations show that *Points* receives negative contributions from *Assists* but changes in line with *Time Played*. The only exception lies in the range equation by CCRM, where the weight of *Assists* range is set to zero due to the nonnegative constraint of regression coefficients. While this result has successfully avoided the mathematical incoherence for interval-valued prediction, it is yet considered biased without taking into account the contribution of *Assists* range to *Points* range. This, again, reveals the disadvantage of modeling the range relationship separately in real-world applications.

5. Conclusions

Building regression models for interval-valued data can be regarded as looking for a line that best fits a cloud of infinitely dense points distributing within the observed hypercubes. Under such premise, a new linear regression modeling method for interval-valued data, referred to as CIM, has been proposed in this paper. CIM employs complete information in interval-valued observations to establish a linear regression model based on the novel inner product operator of interval-valued data. Due to the adoption of Moore's linear combination algorithm, CIM further avoids the mathematical incoherence of the predicted lower and upper bounds of the response variable. Extensive experiments on both the synthetic and real-life data sets have been conducted for the comparison of CM, CRM, CCRM and CIM. Results show that CIM has the best interpretation and prediction performance when we expect to use interval-valued data to approximate the original numeric data. Moreover, even if the uniform distribution assumption does not hold for the numeric data, CIM still has a better performance than CM by using the grid data as the complete information.

Table 9 Estimated regression equations for numeric and interval-valued data

Methods	*Equations*
for numeric data	$\mathbf{Y} = 0.3116 - 0.5870\mathbf{X}_1 + 0.0078\mathbf{X}_2$
CM	$\mathbf{Y}^L = 0.3384 - 1.2530\mathbf{X}_1^L + 0.0118\mathbf{X}_2^L,\ \mathbf{Y}^U = 0.3384 - 1.2530\mathbf{X}_1^U + 0.0118\mathbf{X}_2^U$
CRM	$\mathbf{Y}^C = 0.3384 - 1.2530\mathbf{X}_1^C + 0.0118\mathbf{X}_2^C,\ \mathbf{Y}^R = 0.0468 - 0.0281\mathbf{X}_1^R + 0.0108\mathbf{X}_2^R$
CCRM	$\mathbf{Y}^C = 0.3384 - 1.2530\mathbf{X}_1^C + 0.0118\mathbf{X}_2^C,\ \mathbf{Y}^R = 0.0463 + 0.0106\mathbf{X}_2^R$
CIM	$\mathbf{Y} = 0.3511 - 0.2173\mathbf{X}_1 + 0.0044\mathbf{X}_2$

Appendix 1

Proof to Proposition 1 is given as follows.

Proof.

(i) Positive definiteness:

As defined in Equation (5), we have

$$\langle x,x\rangle = \|x\|^2 = \frac{1}{3}\left(\underline{x}^2 + \underline{x}\overline{x} + \overline{x}^2\right). \tag{28}$$

Since $\underline{x}^2 + \overline{x}^2 \geq 2|\underline{x}\overline{x}| \geq |\underline{x}\overline{x}|$, we have

$$\langle x,x\rangle = \frac{1}{3}\left(\underline{x}^2 + \underline{x}\overline{x} + \overline{x}^2\right) \geq 0, \tag{29}$$

and $\langle x,x\rangle = 0$ if and only if $\underline{x} = \overline{x} = 0$.

(ii) Symmetry:

Based on Equation (7), we obtain

$$\begin{aligned}\langle x,y\rangle &= \frac{1}{4}\left(\underline{x}+\overline{x}\right)\left(\underline{y}+\overline{y}\right)\\ &= \frac{1}{4}\left(\underline{y}+\overline{y}\right)\left(\underline{x}+\overline{x}\right) = \langle y,x\rangle.\end{aligned} \tag{30}$$

(iii) Linearity:

Using the grid data of 2-dimensional hypercubes (see Figure 3(b) or 4(b)), we have

$$\begin{aligned}\langle \alpha x,y\rangle &= \frac{1}{(m+1)^2}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}\alpha x_i y_j\\ &= \frac{\alpha}{(m+1)^2}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} x_i y_j\\ &= \alpha\langle x,y\rangle.\end{aligned} \tag{31}$$

When m trends to infinity, Equation (31) becomes

$$\begin{aligned}\langle \alpha x,y\rangle &= \int_{\underline{y}}^{\overline{y}}\int_{\underline{x}}^{\overline{x}} \alpha xy \frac{1}{\left(\overline{x}-\underline{x}\right)\left(\overline{y}-\underline{y}\right)}dxdy\\ &= \frac{\alpha}{4}\left(\underline{x}+\overline{x}\right)\left(\underline{y}+\overline{y}\right) = \alpha\langle x,y\rangle.\end{aligned} \tag{32}$$

Moreover, with the grid data of 3-dimensional hypercubes (see Figures 3(c) or 4(c)), we can obtain

$$\begin{aligned}\langle x+y,z\rangle &= \frac{1}{(m+1)^3}\sum_{k=1}^{m+1}\sum_{j=1}^{m+1}\sum_{i=1}^{m+1}\left(x_i+y_j\right)z_k\\ &= \frac{1}{(m+1)^2}\sum_{k=1}^{m+1}\sum_{i=1}^{m+1} x_i z_k + \frac{1}{(m+1)^2}\sum_{k=1}^{m+1}\sum_{j=1}^{m+1} y_j z_k\\ &= \langle x,z\rangle + \langle y,z\rangle.\end{aligned} \tag{33}$$

When m trends to infinity, Equation (33) becomes

$$\begin{aligned}&\langle x+y,z\rangle\\ &= \int_{\underline{z}}^{\overline{z}}\int_{\underline{y}}^{\overline{y}}\int_{\underline{x}}^{\overline{x}}(x+y)z\frac{1}{\left(\overline{x}-\underline{x}\right)\left(\overline{y}-\underline{y}\right)\left(\overline{z}-\underline{z}\right)}dxdydz\\ &= \int_{\underline{z}}^{\overline{z}}\int_{\underline{x}}^{\overline{x}} xz\frac{1}{\left(\overline{x}-\underline{x}\right)\left(\overline{z}-\underline{z}\right)}dxdz\int_{\underline{y}}^{\overline{y}} y\frac{1}{\overline{y}-\underline{y}}dy\\ &+\int_{\underline{z}}^{\overline{z}}\int_{\underline{y}}^{\overline{y}} yz\frac{1}{\left(\overline{y}-\underline{y}\right)\left(\overline{z}-\underline{z}\right)}dxdz\int_{\underline{x}}^{\overline{x}} x\frac{1}{\overline{x}-\underline{x}}dx\\ &= \frac{1}{4}\left(\underline{x}+\overline{x}\right)\left(\underline{z}+\overline{z}\right)+\frac{1}{4}\left(\underline{y}+\overline{y}\right)\left(\underline{z}+\overline{z}\right)\\ &= \langle x,z\rangle + \langle y,z\rangle.\end{aligned} \tag{34}$$

■

Appendix 2

We prove Proposition 2 as follow.

Proof. Without loss of generality, assume that $\mathbf{X} = (x_1,\cdots,x_n)'$, $\mathbf{Y} = (y_1,\cdots,y_n)'$ and $\mathbf{Z} = (z_1,\cdots,z_n)'$, where $x_i = \left[\underline{x}_i,\overline{x}_i\right]$, $y_i = \left[\underline{y}_i,\overline{y}_i\right]$, $z_i = \left[\underline{z}_i,\overline{z}_i\right]$, $(i = 1,2,\cdots,n)$. Since the inner product of interval-valued variables is the sum of the inner product of data units, the proof to Proposition 2 will be based on Proposition 1 heavily.

(i) Positive definiteness:

According to positive definiteness in Proposition 1, we have $\langle x_i,x_i\rangle \geq 0$, and the equality holds if and only if $\underline{x}_i = \overline{x}_i = 0$, $i = 1,2,\cdots,n$. Then

$$\langle \mathbf{X}, \mathbf{X} \rangle = \|\mathbf{X}\|^2 = \sum_{i=1}^{n} \langle x_i, x_i \rangle \geq 0, \tag{35}$$

and $\langle \mathbf{X}, \mathbf{X} \rangle = \mathbf{0}$ if and only if $\underline{x}_i = \overline{x}_i = 0$, i.e., $\mathbf{X} = \mathbf{0}$.

(ii) Symmetry:

According to symmetry in Proposition 1, we have $\langle x_i, y_i \rangle = \langle y_i, x_i \rangle$, $i = 1, 2, \cdots, n$. Then

$$\begin{aligned} \langle \mathbf{X}, \mathbf{Y} \rangle &= \sum_{i=1}^{n} \langle x_i, y_i \rangle \\ &= \sum_{i=1}^{n} \langle y_i, x_i \rangle = \langle \mathbf{Y}, \mathbf{X} \rangle. \end{aligned} \tag{36}$$

(iii) Linearity:

According to linearity in Proposition 1, we have $\langle x_i + y_i, z_i \rangle = \langle x_i, z_i \rangle + \langle y_i, z_i \rangle$, and $\langle \alpha x_i, y_i \rangle = \alpha \langle x_i, y_i \rangle, \forall \alpha \in \mathbb{R} (i = 1, 2, \cdots, n)$.

Then

$$\begin{aligned} \langle \mathbf{X} + \mathbf{Y}, \mathbf{Z} \rangle &= \sum_{i=1}^{n} \langle x_i + y_i, z_i \rangle \\ &= \sum_{i=1}^{n} \left(\langle x_i, z_i \rangle + \langle y_i, z_i \rangle \right) \\ &= \sum_{i=1}^{n} \langle x_i, z_i \rangle + \sum_{i=1}^{n} \langle y_i, z_i \rangle \\ &= \langle \mathbf{X}, \mathbf{Z} \rangle + \langle \mathbf{Y}, \mathbf{Z} \rangle, \end{aligned} \tag{37}$$

$$\begin{aligned} \langle \alpha \mathbf{X}, \mathbf{Y} \rangle &= \sum_{i=1}^{n} \langle \alpha x_i, y_i \rangle = \sum_{i=1}^{n} \alpha \langle x_i, y_i \rangle \\ &= \alpha \sum_{i=1}^{n} \langle x_i, y_i \rangle = \alpha \langle \mathbf{X}, \mathbf{Y} \rangle. \end{aligned} \tag{38}$$

■

References

[1] Bertrand, P. & Goupil, F. (2000). Descriptive statistics for symbolic data. In: Bock, H., Diday, E. (eds.), Analysis of symbolic data: exploratory methods for extracting statistical information from complex data, pp. 106-124. Berlin: Springer-Verlag

[2] Billard, L. & Diday, E. (2000). Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., Schader, M. (eds.), Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies: 369-374, Namur, July 11-14, 2000, Springer

[3] Billard, L. & Diday, E. (2002). Symbolic regression analysis. In: Jajuga, K., Sokolowski, A., Bock, H.H. (eds.), Data Analysis, Classification and Related Methods: Proceedings of the Eighth Conference of the International Federation of Classification Societies: 281-288, Cracow, July 14-15, 2002, Springer

[4] Billard, L. & Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. Journal of the American Statistical Association, 98 (462): 470-487

[5] Billard, L. & Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley, Chichester

[6] Bock, H.H. & Diday, E. (2000). Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Berlin

[7] Cazes, P., Chouakria, A., Diday, E. & Schektman, Y. (1997). Extension de l'analyse en composantes principales à des donnés de type intervalle. Revue de Statistique Apliquée, 45 (3): 5-24

[8] de Carvalho, F.A.T., Brito, P. & Bock, H.H. (2006). Dynamic clustering for interval data based on L2 distance. Computational Statistics, 21 (2): 231-250

[9] de Carvalho, F.D.T., de Souza, R.M.C.R., Chavent, M. & Lechevallier, Y. (2006). Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27 (3): 167-179

[10] de Souza, R.M.C.R. & de Carvalho, F.D.T. (2004). Clustering of interval data based on city-block distances. Pattern Recognition Letters, 25 (3): 353-365

[11] Diday, E. (1987). The symbolic approach in clustering and related methods of data analysis. In: Bock, H.H., (ed.), Classification and Related Methods of Data Analysis. Amsterdam: North-Holland

[12] Diday, E. (1989). Introduction à l'approche symbolique en analyse des données. Revue Francaise d'automatique, d'informatique et de Recherche Opérationnelle: Recherche Opérationnelle, 23 (2): 193-236

[13] Diday, E. & Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis and the SODAS Software. Wiley-Interscience, Chichester

[14] Draper, N. & Smith, H. (1981). Applied Regression Analysis. John Wiley, New York

[15] Gioia, F. & Lauro, C.N. (2005). Basic statistical methods for interval data. Statistical Application, 17 (1): 1-29

[16] Gioia, F. & Lauro, C.N. (2006). Principal component analysis on interval data. Computational Statistics, 21 (2): 343-363

[17] Lauro, C.N. & Gioia, F. (2006). Dependence and interdependence analysis for interval-valued variables. In: Batagelj, V., Bock, H.H., Ferligoj, A., Ziberna, A. (eds.), Data Science and Classification. Berlin: Springer-Verlag

[18] Lima, E.D. & de Carvalho, F.D.T. (2008). Centre and Range method for fitting a linear regression model to symbolic interval data. Computational Statistics & Data Analysis, 52 (3): 1500-1515

[19] Lima, E.D. & de Carvalho, F.D.T. (2010). Constrained linear regression models for symbolic interval-valued variables. Computational Statistics & Data Analysis, 54 (2): 333-347

[20] Marino, M. & Palumo, F. (2003). Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression. Statistica Applicata, 14 (3): 277-291

[21] Montgomery, D. (1982). Introduction to Linear Regression Analysis. John Wiley, New York

[22] Moore, R. (1966). Interval Analysis. Prentice Hall, Englewood Cliffs, NJ

[23] Scheffé, H. (1959). The Analysis of Variance

[24] Silva, A.P.D. & Brito, P. (2006). Linear discriminant analysis for interval data. Computational Statistics, 21: 289-308

[25] Simonoff, J. (1994). Smoothing Methods in Statistics. Springer, New York

Huiwen Wang received her B.Sc. degree from Beihang University (BHU), China, in 1982, DEA MASE, from Paris XI, France, in 1989, and Ph.D. degree in engineering system from BHU in 1992. She is currently a professor in Management Science and Engineering Department, School of Economics and Management (SEM), BHU. Also, she is dean of SEM, director of SEM Academic Degrees Committee, and director of Research Center of Complex Data Analysis in BHU. Prof. Wang received National Science Fund for Distinguished Young Scholars. Her general area of research is statistics and data analysis, with a

recent focus on multivariate analysis for high-dimension complex data. She is an IASC member, a member of National Statistics Teaching Materials Review Committee, executive director of China Marketing Association, editorial member of Journal of Symbolic Data Analysis.

Rong Guan is a Ph.D. candidate from the School of Economics and Management at Beihang University, China. She received her B.S. degree in industrial engineering from the same university in 2008. Her research interests are in the area of computational statistics and data analysis, currently focus on multivariate analysis on interval-valued data.

Junjie Wu, the contact author of the paper, received his Ph.D. degree in management science and engineering from Tsinghua University, China, in 2008. He is currently an associate professor in Information Systems Department, School of Economics and Management, Beihang University, China. He is also the director of Social Computing and Sentiment Analysis Center, the vice director of Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, and the outside research fellow of Research Center for Contemporary Management, Key Research Institute of Humanities and Social Sciences at Universities, Tsinghua University. His general area of research is data mining and complex networks, with a special interest in solving the problems raised from the emerging data-intensive applications. He is the recipient of the National Excellent Doctoral Dissertation award (2010) and the New Century Excellent Talents in University award (2011), and the choices of the Microsoft Star-Track program and the Springer Thesis Prize. He is a member of ACM, IEEE, INFORMS, AIS, and CCF.