Knowledge-Based Systems 300 (2024) 112226

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Locally linear method for fixed effects panel interval-valued data model

Jinjin Zhang, Aibing Ji *

College of Mathematics and Information Science, Hebei University, Baoding 071002, China

ARTICLE INFO

Keywords:
Fixed effects
Panel interval-valued data
Nonparametric model
Profile least squares framework
Locally linear method

ABSTRACT

The literature on fixed effects panel interval-valued data models has been established. However, less attention has been given to models that simultaneously, consider the interval information of panel data and explore the nonlinear relationship between interval variables. To deal with this issue, this paper formulates a nonparametric fixed effects panel interval-valued data model. To estimate the fixed effects and nonparametric component, we propose a locally linear method based on the profile least squares framework. Later, experiment results on synthetic and real data sets illustrate the advantages of our proposed model.

1. Introduction

Panel data, also known as longitudinal data or cross-sectional time series data, refers to a type of data that is collected over a period of time on the same observational units. An excellent overview of panel data analysis can be found in [1–3]. One of the most important tools for analyzing this kind of data is the fixed effects panel data model. This model is designed to control for individual or unit-specific effects (fixed effects) and to analyze the impact of time-related factors. It is particularly useful when there is a concern that the observation might be influenced by factors that do not change over the period of study. For instance, in the field of economics, researchers might employ this model to investigate the economic policies' impact on economic growth, while controlling for fixed country-specific factors like geography or culture [4]. Similarly, in medicine, it could be used to study the effect of a new treatment on patient health outcomes, while accounting for individual patient characteristics that do not change over time [5–7].

While fixed effects panel data models have been extensively explored, few address panel data with measurement error that could bias predictive responses [8–10]. To solve this problem, Ji et al. [11] introduced a class of fixed effects panel interval-valued data models: the center model (P-CM), the Minmax model (P-Minmax), and the center and range model (P-CRM). In these models, the measurement error is converted to the radius of the interval, and the center of the interval represents the observation. However, the limitation of the P-CM, P-Minmax, and P-CRM models is reached when the interval variables have a complex nonlinear relationship.

In this paper, we establish a fixed effects panel interval-valued data model based on nonparametric specifications and interval-valued data analysis. The objectives of this paper are (1) to represent the uncertainty and volatility information of panel data in the form of panel interval-valued data, (2) to take into account the nonlinear relationship of interval variables. To get a consistent estimator of the nonparametric component, this paper proposes a locally linear method (P-LM) based on the profile least squares framework. Later, the experimental results on synthetic and real datasets demonstrate that our proposed model performs well compared with existing models.

The contributions of our proposed model and estimation method are listed as follows.

- Our proposed model is suitable for nonlinear panel interval-valued data, as it combines the strengths from both the interval-valued data analysis and nonparametric panel data models.
- We extend a locally linear method to our proposed model based on the profile least squares framework. The unknown nonparametric component can be estimated based on the kernel-based weighted residual sum of squares (as shown in Eq. (23)).
- The proposed method has no restriction on the form of the regression function. The proposed estimator is consistent and has a limiting normal distribution.
- Our proposed method is also suitable if the fixed effects panel interval-valued data model is linear.

The remainder of this paper is organized as follows. Section 2 provides a review of relevant literature. The current fixed effects panel interval-valued data models are introduced in Section 3. In Section 4, we first formulate a fixed effects panel interval-valued data model and then propose a locally linear method for the model. Later, we show the asymptotic properties of the proposed estimator. Sections 5 and 6 present the experimental results of synthetic and real data sets, respectively. Later, concluding remarks are provided in Section 7.

* Corresponding author.
E-mail addresses: zhangjinjin@stumail.hbu.edu.cn (J. Zhang), jab@hbu.edu.cn (A. Ji).

https://doi.org/10.1016/j.knosys.2024.112226
Received 22 April 2024; Received in revised form 10 June 2024; Accepted 7 July 2024
Available online 15 July 2024
0950-7051/

2. Literature review

In this section, we provide a review of relevant literature: fixed effects panel data models; interval-valued data analysis. We identify the strengths of existing works, and present the motivation and the details of our proposed model and estimation method.

2.1. Fixed effects panel data models

There has been much research on the estimation method of fixed effects panel data models in econometrics [1–3]. According to the treatment form of fixed effects, these methods can be classified into two main categories. One is firstly to remove the fixed effects by data transformation and then to estimate the nonparametric/semiparametric component through kernel-based methods (see, for example, [12,13]). It is worth noting that employing data transformation may alter the structure of the nonparametric/semiparametric component. Moreover, another way based on the profile least squares framework does not alter the structure of the nonparametric/semiparametric component. For instance, under the profile least squares framework, Gao and Li [14] proposed an intuitive Nadaraya–Watson kernel method of fixed effects panel data model (NWM). Lee et al. [15] offered the local-within-transformation to replace the fixed effects and presented a locally linear method for fixed effects panel data model (LLM). Many variations to the kernel-based estimation method have also been proposed over time to estimate fixed effects panel data models; see, for example [16–20]. Although these works are not specialized in the fixed effects panel interval-valued data models, they provide a theoretical basis for exploring these models.

2.2. Interval-valued data analysis

Interval-valued data, as a kind of symbolic data, was proposed by Moore [21]. The observations are not represented as single points but rather as intervals, ranges, or sets of possible values. This type of data is particularly useful when there is uncertainty or imprecision in the measurements, and it allows for a more flexible representation of the uncertain information [22]. Interval-valued data has garnered significant attention in different statistical problems, such as time series analysis [23–25], regression analysis [26–28], clustering [29–32], optimization [33–35], hypothesis testing [36], feature selection [37,38].

In the interval-valued regression analysis, the response variable (dependent variable) and the explanatory variables (independent variables) are represented as intervals [39]. The objective is to model the relationship between interval explanatory variables and interval response variable, allowing for a more realistic representation of the variability in the data. The first discussion on the regression model for interval-valued data was introduced by Billard and Diday [39], who proposed the center model (CM). This approach utilizes the midpoints of intervals to fit the regression model. Following this, Billard and Diday [40] introduced the Minmax model (Minmax), which relies on the lower and upper bounds of intervals. Later, based on the center and range of intervals, Lima Neto and de Carvalho [41] proposed the center and range model (CRM). The CRM model has received considerable attention for its effective representation of full interval information [42–44]. Recently, researchers have developed the kernel-based methods for interval-valued regression models [45,46]. To capture the nonlinear relationship between interval variables, Fagundes et al. [47] proposed a kernel method based on kernel smoothing approaches, while Sun et al. [4] and Kong et al. [48] presented a locally linear model for cross-sectional interval-valued data regression models. In conclusion, there exists much literature on the cross-sectional interval-valued regression analysis, but few develop more complex panel interval-valued data models.

In this paper, we specifically focus our attention on a locally linear method presented by Su and Ullah [49], who investigated the partially linear fixed effects panel data model (panel point-valued data). Under the profile least squares framework, the locally linear method is extended to estimate the nonparametric fixed effects panel interval-valued data model. Specifically, fixed effects are firstly treated as unknown parameters, and the nonparametric component is estimated by minimizing the kernel-based weighted residual sum of squares, as shown in Eq. (23). In this way, the estimated nonparametric component is only related to fixed effects. Then, plugging the estimated nonparametric component into the residual sum of squares, fixed effects can be estimated based on the least squares method. Later, the nonparametric component is estimated based on the estimated fixed effects.

3. Current fixed effects panel interval-valued data model

Consider the panel interval-valued data set

$$\{(x_{it}, y_{it}) : i = 1, 2, \dots, n; t = 1, 2, \dots, T\} \tag{1}$$

where $x_{it} = (x_{it1}, x_{it2}, \dots, x_{itp})'$ are a p-dimensional vector of interval explanatory variables and y_{it} is an interval response variable. A' denotes the transpose of matrix or vector A. The subscript i typically denotes individual units, such as people, households, firms, countries, or any other entities that are repeatedly observed over time; the subscript t represents time periods or the chronological sequence of observations. It could be days, months, years, etc., depending on the frequency of data collection. The observations of x_{itj} and y_{it} are two interval-valued data for $i = 1, 2, \dots, n; t = 1, 2, \dots, T; j = 1, 2, \dots, p$, $x_{itj} = [x^l_{itj}, x^u_{itj}]$ with $x^l_{itj} \le x^u_{itj}$, and $y_{it} = [y^l_{it}, y^u_{it}]$ with $y^l_{it} \le y^u_{it}$, where superscripts l and u denote the lower and upper bounds of the interval. Alternatively, each interval can be represented equivalently by its midpoint and range [41], that is, $x_{itj} = [x^c_{itj} - x^r_{itj}, x^c_{itj} + x^r_{itj}]$ and $y_{it} = [y^c_{it} - y^r_{it}, y^c_{it} + y^r_{it}]$, where

$$x^c_{itj} = \frac{1}{2}(x^l_{itj} + x^u_{itj}), \quad x^r_{itj} = \frac{1}{2}(x^u_{itj} - x^l_{itj})$$
$$y^c_{it} = \frac{1}{2}(y^l_{it} + y^u_{it}), \quad y^r_{itj} = \frac{1}{2}(y^u_{it} - y^l_{it})$$

Remark 1. If x^l_{it} and x^u_{it} are independent, identically distributed random variables across ith index, then x^c_{it} and x^r_{it} are independent, identically distributed random variables across ith index.

In the following section, we introduce three fixed effects panel interval-valued models (P-CM, P-Minmax, P-CRM) proposed by [11].

3.1. Center model for panel interval-valued data (P-CM)

In the P-CM model, suppose that y^l_{it} and y^u_{it} can be independently explained by x^l_{it} and x^u_{it}, and they follow the same regression model

$$y^l_{it} = \mu^c_i + \sum_{j=1}^{p} x^l_{itj}\beta^c_j + \varepsilon^l_{it} \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{2}$$

$$y^u_{it} = \mu^c_i + \sum_{j=1}^{p} x^u_{itj}\beta^c_j + \varepsilon^u_{it} \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{3}$$

where $\beta^c_1, \beta^c_2, \dots, \beta^c_p$ are the unknown parameters, μ^c_i is the fixed effect related to x^c_{itj} for all $t = 1, 2, \dots, T; j = 1, 2, \dots, p$. Eqs. (2) and (3) are equivalent to the following equation:

$$y^c_{it} = \mu^c_i + \sum_{j=1}^{p} x^c_{itj}\beta^c_j + \varepsilon^c_{it} \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{4}$$

where $\varepsilon^c_{it} = \frac{1}{2}(\varepsilon^l_{it} + \varepsilon^u_{it})$. Define $\beta^c = (\beta^c_1, \beta^c_2, \dots, \beta^c_p)'$. The estimator of β^c can be obtained based on least squares dummy variable (LSDV) approach [1].

$$\hat{\beta}^c = ((X^c)'\Omega^c X^c)^{-1}(X^c)'\Omega^c Y^c \tag{5}$$

where $X^c = (x^c_1, x^c_2, \dots, x^c_n)'$ is an $nT \times p$ matrix with $x^c_i = (x^c_{i1}, x^c_{i2}, \dots, x^c_{iT})$ for $i = 1, 2, \dots, n$. Each element of x^c_{it} is an $1 \times p$ vector and denoted as $x^c_{it} = (x^c_{it1}, x^c_{it2}, \dots, x^c_{itp})$ for $t = 1, 2, \dots, T$. $Y^c = (y^c_1, y^c_2, \dots, y^c_n)'$ with

$y_i^c = (y_{i1}^c, y_{i2}^c, \ldots, y_{iT}^c)$. $\Omega^c = I_{nT} - D_0(D_0'D_0)^{-1}D_0'$, $D_0 = I_n' \otimes l_T$, the operator $\otimes$ denotes the Kronecker product, I_n denotes an $n \times n$ identity matrix, l_T is an T dimensional vector with all elements equal to one. Because the intercept term μ_i^c is a constant independent of time, the corresponding estimator of μ^c can be obtained by

$$\hat{\mu}_i^c = y_{i\cdot}^c - x_{i\cdot}^c \hat{\beta}^c \tag{6}$$

where $y_{i\cdot}^c = T^{-1}\sum_{i=1}^n y_{it}^c$ and $x_{i\cdot}^c = T^{-1}\sum_{i=1}^n x_{it}^c$. Thus, the lower and upper bounds of predictive response by P-CM are given by

$$\hat{y}_{it}^l = \hat{\mu}_i^c + \sum_{j=1}^p x_{itj}^l \hat{\beta}_j^c \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{7}$$

$$\hat{y}_{it}^u = \hat{\mu}_i^c + \sum_{j=1}^p x_{itj}^u \hat{\beta}_j^c \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{8}$$

One limitation of P-CM model is that it only considers the center information of intervals in the estimation procedure, and thereby the predictive response may be biased.

3.2. Minmax model for panel interval-valued data (P-Minmax)

P-Minmax model supposes that y_{it}^l and y_{it}^u can be independently explained by x_{it}^l and x_{it}^u, that is,

$$y_{it}^l = \mu_i^l + \sum_{j=1}^p x_{itj}^l \beta_j^l + \varepsilon_{it}^l \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{9}$$

$$y_{it}^u = \mu_i^u + \sum_{j=1}^p x_{itj}^u \beta_j^u + \varepsilon_{it}^u \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{10}$$

Based on LSDV approach, we can obtain the estimators of β^l, β^u, μ_i^l, and μ_i^u. Thus, the predictive response $\hat{y}_{it} = [\hat{y}_{it}^l, \hat{y}_{it}^u]$ is given by

$$\hat{y}_{it}^l = \hat{\mu}_i^l + \sum_{j=1}^p x_{itj}^l \hat{\beta}_j^l \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{11}$$

$$\hat{y}_{it}^u = \hat{\mu}_i^u + \sum_{j=1}^p x_{itj}^u \hat{\beta}_j^u \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{12}$$

P-Minmax model uses the interval bound information to obtain the predictive response $\hat{y}_{it}$. However, it also may lead to a biased predictive response since it omits the interval center and range information.

3.3. Center and range model for panel interval-valued data (P-CRM)

P-CRM model supposes that y_{it}^c and y_{it}^r can be independently explained by x_{it}^c and x_{it}^r, that is,

$$y_{it}^c = \mu_i^c + \sum_{j=1}^p x_{itj}^c \beta_j^c + \varepsilon_{it}^c \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{13}$$

$$y_{it}^r = \mu_i^r + \sum_{j=1}^p x_{itj}^r \beta_j^r + \varepsilon_{it}^r \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{14}$$

Using the same procedure, we can obtain the estimators of β^c, β^r, μ_i^c, and μ_i^r. Further, according to (13) and (14), we can obtain the center and range of predictive response $\hat{y}_{it}^c$ and $\hat{y}_{it}^r$. Then, the predictive response $\hat{y}_{it} = [\hat{y}_{it}^l, \hat{y}_{it}^u]$ is given by

$$\hat{y}_{it}^l = \hat{y}_{it}^c - \hat{y}_{it}^r \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{15}$$

$$\hat{y}_{it}^l = \hat{y}_{it}^c + \hat{y}_{it}^r \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{16}$$

The P-CRM model performs better than the P-CM model since it uses more interval information. However, this improvement only occurs when there is a linear dependency between the response and explanatory variables.

4. Fixed effects panel interval-valued data model

In this section, we establish a nonlinear fixed effects panel interval-valued data model, and propose the P-LM estimation method for this model. Section 4.1 provides the model specification, and the P-LM estimation method is proposed in detail in Section 4.2. Later, Section 4.3 discusses the asymptotic results of the proposed estimators.

4.1. Model specification

In this section, we assume that x_{it}^c and x_{it}^r are independently related to y_{it}^c and y_{it}^r, and we relax the assumptions on the form of the regression function, that is, there exist two sufficient smooth (i.e., twice-differentiable) multivariate functions $m_1(\cdot)$ and $m_2(\cdot)$ satisfying the following relationships

$$y_{it}^c = m_1(x_{it}^c) + \mu_i^c + \varepsilon_{it}^c \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{17}$$

$$y_{it}^r = m_2(x_{it}^r) + \mu_i^r + \varepsilon_{it}^r \quad i = 1, 2, \ldots, n; t = 1, 2, \ldots, T \tag{18}$$

where $x_{it}^c = (x_{it1}^c, x_{it2}^c, \ldots, x_{itp}^c)'$ and $x_{it}^r = (x_{it1}^r, x_{it2}^r, \ldots, x_{itp}^r)'$ are two $p \times 1$ dimensional vectors. μ_i^c and μ_i^r are the unobserved fixed effects. ε_{it}^c and ε_{it}^r are the random disturbances. For identification purpose, we assume that $\sum_{i=1}^n \mu_i^c = 0$ and $\sum_{i=1}^n \mu_i^r = 0$ [14,49,50]. Rewriting models (17) and (18) in a matrix form yields

$$Y^c = m_1(X^c) + D\mu^c + \varepsilon^c \tag{19}$$

$$Y^r = m_2(X^r) + D\mu^r + \varepsilon^r \tag{20}$$

where $m_1(X^c) = (m_1(x_1^c), m_1(x_2^c), \ldots, m_1(x_n^c))'$ is an $nT \times 1$ vector with $m_1(x_i^c) = (m_1(x_{i1}^c), m_1(x_{i2}^c), \ldots, m_1(x_{iT}^c))'$ for $i = 1, 2, \ldots, n$. $Y^c, \varepsilon^c, Y^r, m_2(X^c)$, and ε^r are similarly defined. $\mu^c = (\mu_2^c, \mu_3^c, \ldots, \mu_n^c)'$ and $\mu^r = (\mu_2^r, \mu_3^r, \ldots, \mu_n^r)'$ are two $n-1$ dimensional vectors. $D = [-l_{n-1}, I_{n-1}]' \otimes l_T$, the operator $\otimes$ denotes the Kronecker product, I_n denotes an $n \times n$ identity matrix, l_{n-1} is an $n-1$ dimensional vector with all elements equal to one, and l_T is an T dimensional vector with all elements equal to one.

4.2. Estimation procedure

In this section, we divide the problem that estimates the response $y_{it} = [y_{it}^l, y_{it}^u]$ in three parts: (1) the estimators of $m_1(\cdot)$ and $m_2(\cdot)$ with unknown parameters μ^c and μ^r, as shown in Eqs. (24) and (25), respectively; (2) the parameter estimators $\hat{\mu}^c$ and $\hat{\mu}^r$, as shown in Eqs. (27) and (28), respectively; (3) the estimators of $m_1(\cdot)$ and $m_2(\cdot)$ with known parameters μ^c and μ^r, as shown in Eqs. (29) and (30), respectively; (4) the predictive response $\hat{y}_{it} = [\hat{y}_{it}^l, \hat{y}_{it}^u]$, as shown in Eqs. (31) and (32).

Let x be a given panel interval-valued data with center x^c and range x^r. We assume that the function $m_1(\cdot)$ has continuous derivative in the neighborhood of x^c and function $m_2(\cdot)$ has continuous derivative in the neighborhood of x^r. Now by Taylor's theorem [47], for x_{it}^c in the neighborhood of x^c,

$$m_1(x_{it}^c) \approx m_1(x^c) + m_{1,x}'(x_{it}^c - x^c) \equiv \beta_0^c + (\beta_1^c)'(x_{it}^c - x^c) \tag{21}$$

where $m_{1,x} = \dfrac{\partial m_1(x^c)}{\partial x^c}$, $\beta_0^c = m_1(x^c)$, $\beta_1^c = m_{1,x}$. Similarly, for x_{it}^r in the neighborhood of x^r,

$$m_2(x_{it}^r) \approx m_2(x^r) + m_{2,x}'(x_{it}^r - x^r) \equiv \beta_0^r + (\beta_1^r)'(x_{it}^r - x^r) \tag{22}$$

where $m_{2,x} = \dfrac{\partial m_1(x^r)}{\partial x^r}$, $\beta_0^r = m_1(x^r)$, $\beta_1^r = m_{2,x}$. According to Taylor's theorem, much of our attention will be devoted to the estimators of $m_1(\cdot)$ and $m_2(\cdot)$, i.e., β_0^c and β_0^r.

Define $\beta^c = (\beta_0^c, (\beta_1^c)')'$ and $\beta^r = (\beta_0^r, (\beta_1^r)')'$. Based on the weighted least squares theory [1,2], the estimators of β^c and β^r can be obtained by solving the following optimization problem

$$\min_{\beta^c, \beta^r, \mu^c, \mu^r} \begin{array}{l} \{[Y^c - D\mu^c - \tilde{X}^c\beta^c]' W_1(x^c)[Y^c - D\mu^c - \tilde{X}^c\beta^c] \\ +[Y^r - D\mu^r - \tilde{X}^r\beta^r]' W_1(x^r)[Y^r - D\mu^r - \tilde{X}^r\beta^r]\} \end{array} \tag{23}$$

where $\tilde{X}^c = (l_{nT}, X^c - l_{nT}x^c)$, $\tilde{X}^r = (l_{nT}, X^r - l_{nT}x^r)$, l_{nT} is an $nT\times 1$ dimensional vector with all elements equal to one. $W_1(x^c)$ and $W_1(x^r)$ are two $nT\times nT$ diagonal matrixes, $W_1(x^c) = \text{diag}\{K_H(x_1^c, x^c), K_H(x_2^c, x^c), \dots, K_H(x_n^c, x^c)\}$. The diagonal element $K_H(x_i^c, x^c) = \text{diag}\left\{K_H(x_{i1}^c, x^c), K_H(x_{i2}^c, x^c), \dots, K_H(x_{iT}^c, x^c)\right\}$ of which $K_H(x_{it}^c, x^c) = |H|^{-1} K(H^{-1}(x_{it}^c, x^c))$ for $i = 1, 2, \dots, n$; $t = 1, 2, \dots, T$, where K denotes the multivariate kernel function, H is the bandwidth matrix, and in general $H = \text{diag}(h_1, h_2, \dots, h_p)$. The definition of $W_1(x^r)$ is the same definition as $W_1(x^c)$

Taking the first-order partial derivative of the objective function of the optimization problem (23) with respect to β^c and β^r, respectively, and equaling they to zero, we have

$$\tilde{\beta}^c = [(\tilde{X}^c)' W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)' W_1(x^c)(Y^c - D\mu^c)$$
$$\tilde{\beta}^r = [(\tilde{X}^r)' W_1(x^r)\tilde{X}^r]^{-1}(\tilde{X}^r)' W_1(x^r)(Y^r - D\mu^r)$$

Let e_i be a $p+1$ dimensional vector of which the first element equals to 1 and all other equals to 0, we have

$$\tilde{\beta}_0^c = e_i'[(\tilde{X}^c)' W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)' W_1(x^c)(Y^c - D\mu^c) \tag{24}$$
$$\tilde{\beta}_0^r = e_i'[(\tilde{X}^r)' W_1(x^r)\tilde{X}^r]^{-1}(\tilde{X}^r)' W_1(x^r)(Y^r - D\mu^r) \tag{25}$$

Note that $\tilde{\beta}_0^c$ and $\tilde{\beta}_0^r$ are the function of the unknown parameters μ^c and μ^r, respectively. Replacing $m_1(x_{it}^c)$ and $m_2(x_{it}^c)$ in models (17) and (18) by $\tilde{\beta}_0^c$ and $\tilde{\beta}_0^r$, respectively, and then on the basis of least squares method, we have

$$\min_{\mu^c,\mu^r} \left\{[Y^c - D\mu^c]' G_H^c [Y^c - D\mu^c] + [Y^r - D\mu^r]' G_H^r [Y^r - D\mu^r]\right\} \tag{26}$$

where $G_H^c = [I_{nT} - S_H^c]'[I_{nT} - S_H^c]$ with $S_H^c = (s_h^c(x_{11}^c), s_h^c(x_{12}^c), \dots, s_h^c(x_{nT}^c))'$ being an $nT \times nT$ matrix. The element $s_h^c(x^c)' = e_1'[(\tilde{X}^c)' W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)' W_1(x^c)$. G_H^r is similarly defined as G_H^c. Taking the first-order partial derivative of the objective function of the optimization problem (26) with respect to μ^c and μ^r, respectively, and equaling they to zero, we have

$$\hat{\mu}^c = (D' G^c D) D' G^c Y^c \tag{27}$$
$$\hat{\mu}^r = (D' G^r D) D' G^r Y^r \tag{28}$$

Now, replacing μ^c and μ^r in Eqs. (24) and (25) with $\hat{\mu}^c$ and $\hat{\mu}^r$, we can directly obtain the estimators of $m_1(x^c)$ and $m_2(x^r)$.

$$\hat{m}_1(x^c) = \hat{\beta}_0^c = s_h^c(x^c)' M^c Y^c \tag{29}$$
$$\hat{m}_2(x^r) = \hat{\beta}_0^r = s_h^r(x^r)' M^r Y^r \tag{30}$$

where $M^c = I - D[D' G^c D]^{-1} D' G^c$ and $M^r = I - D[D' G^r D]^{-1} D' G^r$. Then, we have

$$\hat{y}_{it}^c = \hat{m}_1(x^c) + \hat{\mu}_i^c = \left\{s_h^c(x^c)' M^c + e_i'[D' G^c D]^{-1} D' G^c\right\} Y^c$$
$$\hat{y}_{it}^r = \hat{m}_2(x^r) + \hat{\mu}_i^r = \left\{s_h^r(x^r)' M^r + e_i'[D' G^r D]^{-1} D' G^r\right\} Y^r$$

Further, the predictive response $\hat{y}_{it} = [\hat{y}_{it}^l, \hat{y}_{it}^u]$ can be written as

$$\hat{y}_{it}^l = \hat{y}_{it}^c - \hat{y}_{it}^r \tag{31}$$
$$\hat{y}_{it}^u = \hat{y}_{it}^c + \hat{y}_{it}^r \tag{32}$$

When the predicted response interval violates the definition of the standard interval, that is, $\hat{y}_{it}^l \geq \hat{y}_{it}^u$, we employ a transformation form below which is presented in [11]

$$\hat{y}_{it} = \begin{cases} [\hat{y}_{it}^l, \hat{y}_{it}^u] & \text{for } \hat{y}_{it}^l < \hat{y}_{it}^u \\ [\frac{1}{2}(\hat{y}_{it}^l + \hat{y}_{it}^u), \frac{1}{2}(\hat{y}_{it}^l + \hat{y}_{it}^u)] & \text{for } \hat{y}_{it}^l \geq \hat{y}_{it}^u \end{cases} \tag{33}$$

Remark 2. The proposed model (see, Eqs. (17) and (18)) is an extension of the P-CRM model (see, Eqs. (13) and (14)). We can also establish the models based on the P-CM and P-Minmax models. Below, we provide a brief description of the extended models.

(i) When x_{it}^c is related to y_{it}^c, we can establish the nonparametric fixed effects panel interval-valued data model as follows:

$$y_{it}^c = m(x_{it}^c) + \mu_i^c + \varepsilon_{it}^c \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{34}$$

where μ_i^c, $m(\cdot)$, and ε_{it}^c are similarly defined as Section 4.1. The estimation procedure of μ_i^c, $m(\cdot)$, and ε_{it}^c are similar to Section 4.2. The predictive responses $\hat{y}^l$ and $\hat{y}^u$ can be derived from

$$\hat{y}_{it}^l = \hat{m}(x_{it}^l) + \hat{\mu}_i^c \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{35}$$
$$\hat{y}_{it}^u = \hat{m}(x_{it}^u) + \hat{\mu}_i^c \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{36}$$

(ii) When x_{it}^l and x_{it}^u are independently related to y_{it}^l and y_{it}^u, we can establish the nonparametric fixed effects panel interval-valued data model as follows:

$$y_{it}^l = m_1(x_{it}^l) + \mu_i^l + \varepsilon_{it}^l \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{37}$$
$$y_{it}^u = m_2(x_{it}^u) + \mu_i^u + \varepsilon_{it}^u \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{38}$$

where μ_i^l, μ_i^u, $m_1(\cdot)$, $m_2(\cdot)$, ε_{it}^l, and ε_{it}^u are similarly defined as Section 4.1. The estimation procedure of μ_i^l, μ_i^u, $m_1(\cdot)$, $m_2(\cdot)$, ε_{it}^l, and ε_{it}^u are similar to Section 4.2. The predictive responses $\hat{y}^l$ and $\hat{y}^u$ can be derived from

$$\hat{y}_{it}^l = \hat{m}_1(x_{it}^l) + \hat{\mu}_i^l \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{39}$$
$$\hat{y}_{it}^u = \hat{m}_2(x_{it}^u) + \hat{\mu}_i^u \quad i = 1, 2, \dots, n; t = 1, 2, \dots, T \tag{40}$$

It is worth noting that, before applying the model, we should capture the location of the fixed regression reference points within the interval. Otherwise, the prediction performance by our model and estimation method be decreased. For example, the proposed model (see, Eqs. (17) and (18)) can be selected if and only if x_{it}^c and x_{it}^r are independently correlated with y_{it}^c and y_{it}^r. When the interval variables satisfy the conditions of Remark 1, the extension of P-CRM (see, Eqs. (17) and (18)) is equivalent to the extension of P-Minmax (see, Eqs. (37) and (38)). A similar discussion is represented in Section 5, when the P-CM, P-CRM, and P-Minmax models fit the linear panel interval-valued dataset (x_{it}^c and x_{it}^r are independently correlated with y_{it}^c and y_{it}^r), the P-CRM and P-Minmax models perform best.

4.3. Asymptotic results

In this section, the asymptotic distributions of $m_1(x^c)$ and $m_2(x^r)$ are derived under the following assumptions:

(A1) The continuous random variables (y_{it}^c, x_{it}^c) and (y_{it}^r, x_{it}^r) are independently and identically distributed (i.i.d) across the i individual, respectively. x_{it}^c is a strictly stationary α-mixing process with mixing coefficients $\alpha = O(k^{-(\delta+2)/\delta})$, $\mathbb{E}(\left\|x_{it}^c\right\|^{2+\delta'} \leq \infty)$, for $\delta' \geq \delta \geq 0$. x_{it}^c has common continuous density function $f_1(x^c)$ with compact support $S \subseteq \mathbb{R}^q$. x^c is the interior point of S. Also, $f_1(x^c) > 0$ holds, $f_1(x^c)$ is continuously twice differentiable, and the second-order derivatives of $m_1(x^c)$ are continuous. The assumption of x_{it}^r is similarly defined.

(A2) The unobserved fixed effects $\left\{\mu_i^c\right\}$ and $\left\{\mu_i^r\right\}$ are i.i.d. for $i = 1, 2, \dots, n; t = 1, 2, \dots, T$.

$$\mathbb{E}(\mu_i^c) = 0, \quad \mathbb{E}[(\mu_i^c)^2] = \sigma_{\mu^c}^2, \text{ and } \mathbb{E}[(\mu_i^c|x_{it}^c)] \neq 0$$
$$\mathbb{E}(\mu_i^r) = 0, \quad \mathbb{E}[(\mu_i^r)^2] = \sigma_{\mu^r}^2, \text{ and } \mathbb{E}[(\mu_i^r|x_{it}^r)] \neq 0$$

The idiosyncratic errors $\left\{\varepsilon_{it}^c\right\}$ and $\left\{\varepsilon_{it}^r\right\}$ are i.i.d. for $i = 1, 2, \dots, n; t = 1, 2, \dots, T$.

$$\mathbb{E}(\varepsilon_{it}^c|\mu_i^c, x_{it}^c) = 0 \text{ and } \mathbb{E}[(\varepsilon_{it}^c)^2|\mu_i^c, x_{it}^c] = \sigma_{\varepsilon^c}^2$$
$$\mathbb{E}(\varepsilon_{it}^r|\mu_i^r, x_{it}^r) = 0 \text{ and } \mathbb{E}[(\varepsilon_{it}^r)^2|\mu_i^r, x_{it}^r] = \sigma_{\varepsilon^r}^2$$

(A3) $K(u) = \prod_{j=0}^q k(u_s)$ is a product kernel of which $k(\cdot)$ is a bounded, symmetric univariate kernel function with compact support on $\mathbb{R}$, and such that $\int K(u)du = 0$, $\int K(u)^2 du = \varsigma_0$, and $\int u'uK(u)du = \kappa_2 I_q$, where I_q is a $q \times q$ identity matrix.

(A4) Let $H = \text{diag}\{h_1, h_2, \ldots, h_q\}$, $|H| = h_1, h_2, \ldots, h_q$ and $\|H\| = \sqrt{\sum_{j=1}^{q} h_j^2}$. As $n \to \infty$ and $T \to \infty$, we assume $h_j \to 0$ for $j = 1, 2, \ldots, p$; $nT|H| \to \infty$; $nT|H|\|H\|^4 \to O(1)$.

Assumptions (A1), (A3), and (A4) are common in the literature of kernel-type nonparametric regression [14,49,50]. In Assumption (A2), $\mathbb{E}[(\mu_i^c|x_{it}^c)] \neq 0$ and $\mathbb{E}[(\mu_i^r|x_{it}^r)] \neq 0$ imply that the models (17) and (18) are the fixed effects panel data models, respectively. Also, $\mathbb{E}(\epsilon_{it}^c|\mu_i^c, x_{it}^c) = 0$ and $\mathbb{E}(\epsilon_{it}^r|\mu_i^r, x_{it}^r) = 0$ mean that the strict exogeneity assumptions for models with fixed effects $\mathbb{E}(y_{it}^c|\mu_i^c, x_{it}^c) = m_1(x_{it}^c) + \mu_i^c$ and $\mathbb{E}(y_{it}^r|\mu_i^r, x_{it}^r) = m_2(x_{it}^r) + \mu_i^r$ hold, respectively.

Theorem 1. *Under Assumptions* $(A1)-(A4)$,

$$\sqrt{nTh_1h_2\cdots h_q}\left(\hat{m}_1(x^c) - m_1(x^c) - \frac{1}{2}\kappa_2 tr\left\{Hm_{1,xx}H\right\}\right) \xrightarrow{d} \left(0, \frac{\varsigma_0\sigma_{\epsilon^c}^2}{f(x^c)}\right) \quad (41)$$

$$\sqrt{nTh_1h_2\cdots h_q}\left(\hat{m}_2(x^r) - m_2(x^r) - \frac{1}{2}\kappa_2 tr\left\{Hm_{2,xx}H\right\}\right) \xrightarrow{d} \left(0, \frac{\varsigma_0\sigma_{\epsilon^r}^2}{f(x^r)}\right) \quad (42)$$

Moreover, the bias is zero if function $m_1(x^c)$ ($m_1(x^r)$) is linear.

The proof is provided in Appendix.

Theorem 2. *Under Assumptions* $(A1)-(A4)$, *when $H = hI_q$, the mean integrated square error (MISE) will be*

$$MISE = h^4\frac{\kappa_2^2}{4}\left\{\int m_{1,xx}m'_{1,xx}dx^c + \int m_{2,xx}m'_{2,xx}dx^r\right\} + \frac{\varsigma_0}{nTh^q}\left\{\sigma_{\epsilon^c}^2\int (f(x^c))^{-1}dx^c + \sigma_{\epsilon^r}^2\int (f(x^r))^{-1}dx^r\right\} + O_p(h^4) + O_p((nT)^{-\frac{1}{2}}h^{1-\frac{q}{2}}) + O((nT)^{-1}h^{-q+1})$$

where

$$h^4\frac{\kappa_2^2}{4}\left\{\int m_{1,xx}m'_{1,xx}dx^c + \int m_{2,xx}m'_{2,xx}dx^r\right\} + \frac{\varsigma_0}{nTh^q}\left\{\sigma_{\epsilon^c}^2\int (f(x^c))^{-1}dx^c + \sigma_{\epsilon^r}^2\int (f(x^r))^{-1}dx^r\right\}$$

is called the asymptotical mean integrated square error (AMISE). The optimal smoothing bandwidth is chosen by minimizing AMISE. Namely,

$$h_{opt} = \left\{\frac{q}{nT}\cdot\frac{\left\{\sigma_{\epsilon^c}^2\int (f(x^c))^{-1}dx^c + \sigma_{\epsilon^r}^2\int (f(x^r))^{-1}dx^r\right\}}{\left\{\int m_{1,xx}m'_{1,xx}dx^c + \int m_{2,xx}m'_{2,xx}dx^r\right\}}\right\}$$

The proof is provided in Appendix.

5. Monte Carlo simulations

In this section, we use simulations to assess the performance of our proposed model (P-LM) and compare its prediction accuracy with other fixed effects models (P-CM [11], P-Minmax [11], P-CRM [11], NWM [14], LLM [15]).

5.1. Measurements

To evaluate the prediction accuracy of the four models P-LM, P-CM, P-Minmax, P-CRM, we introduce the following three popular evaluation measurements [11,48].

(1) The root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{2nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[(\hat{y}_{it}^l - y_{it}^l)^2 + (\hat{y}_{it}^u - y_{it}^u)^2\right]}$$

(2) The mean absolute error (MAE):

$$\text{MAE} = \frac{1}{2nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[|\hat{y}_{it}^l - y_{it}^l| + |\hat{y}_{it}^u - y_{it}^u|\right]$$

(3) The mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{1}{2nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\frac{|\hat{y}_{it}^l - y_{it}^l|}{y_{it}^l} + \frac{|\hat{y}_{it}^u - y_{it}^u|}{y_{it}^u}\right]\cdot 100\%$$

In practice, the classical fixed effects panel data models, i.e., NWM and LLM, are fitted using the interval midpoint (that is the mean value of observations), and thereby correspond to the single predictive response $\hat{y}_{it}$. To evaluate the prediction accuracy of the models NWM and LLM, we introduce corresponding evaluation measurements RMSE, MAE, and MAPE.

$$\text{RMSE} = \sqrt{\frac{1}{2nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[(\hat{y}_{it} - y_{it}^l)^2 + (\hat{y}_{it} - y_{it}^u)^2\right]}$$

$$\text{MAE} = \frac{1}{2nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[|\hat{y}_{it} - y_{it}^l| + |\hat{y}_{it} - y_{it}^u|\right]$$

$$\text{MAPE} = \frac{1}{2nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\frac{|\hat{y}_{it} - y_{it}^l|}{y_{it}^l} + \frac{|\hat{y}_{it} - y_{it}^u|}{y_{it}^u}\right]\cdot 100\%$$

5.2. Synthetic datasets

We simulate two Data Generation Processes (DGP); one for the nonlinear relationship (namely, DGPI) and another for the linear relationship (namely, DGPII). In nonlinear cases, the center of response variable is generated from the models NWM [14], and the range of response variable is based on the exponential function [48] (guarantee nonnegative range).

Configuration (Fixed effects panel data model based on center and range of intervals)

step i. Generate the center and range of explanatory variable, respectively.

$$x_{it}^c \sim U[-1, 1]$$
$$x_{it}^r \sim U[0, 1]$$

step ii. Generate the center and range of fixed effects

$$\mu_i^c = v_i + T^{-1}\sum_{i=1}^{T} x_{it}^c$$

$$\mu_i^r = v_i + T^{-1}\sum_{i=1}^{T} x_{it}^r$$

where v_i is i.i.d. uniform$[0, 1]$.

step iii. Derive the center and range of response variable as

DGPI: $$y_{it}^c = \mu_i^c + \sin(\pi x_{it}^c) + \epsilon_{it}^c$$ $$y_{it}^r = \mu_i^c + \exp(x_{it}^r) + \epsilon_{it}^r$$

DGPII: $$y_{it}^c = \mu_i^c + x_{it}^c\beta^c + \epsilon_{it}^c$$ $$y_{it}^r = \mu_i^r + x_{it}^r\beta^r + \epsilon_{it}^r$$

where the error term $\epsilon_{it}^c, \epsilon_{it}^r \sim N(0.1)$.

step iv. Derive the interval lower and upper bounds of the response and explanatory variables as

$$x_{it}^l = x_{it}^c - x_{it}^r \quad y_{it}^l = y_{it}^c - y_{it}^r$$
$$x_{it}^u = x_{it}^c + x_{it}^r \quad y_{it}^u = y_{it}^c + y_{it}^r$$

We take $n = 50, 60, 80, 110$, and 200 and $T = 3, 4, 5, 10$, and 20. We employ the Gaussian kernel to fit the nonparametric models P-LM, NWM, and LLM, and the optimal bandwidth is selected through biased cross-validation. To ensure sufficient statistical power, the number of Monte Carlo replications is set to 1000.

5.3. Comparison results

The simulation results are reported in Tables 1 and 2. Table 1 reports simulation results for DGPI (panel interval-valued data follows

Table 1
The RMSE, MAE, and MAPE for DGPI.

T	n	measures	P-LM	P-CM	P-CRM	P-Minmax	NWM	LLM
3	50	RMSE	1.4885	6.4865	4.6428	4.6583	8.5431	8.2460
		MAE	0.5094	2.5864	1.7282	1.7061	3.7529	3.7529
		MAPE	62.89%	290.72%	251.05%	242.52%	477.24%	521.72%
	60	RMSE	1.1315	6.6571	4.6532	4.6177	8.3858	8.0344
		MAE	0.3910	2.6038	1.6865	1.6552	3.6806	3.6804
		MAPE	57.75%	233.47%	177.54%	504.63%	401.86%	472.64%
	80	RMSE	1.0485	6.3658	4.4811	4.4756	8.1655	7.8309
		MAE	0.3684	2.5442	1.6146	1.6046	3.5968	3.5964
		MAPE	42.96%	244.19%	201.04%	270.33%	729.48%	1123.00%
	110	RMSE	0.6129	6.7753	4.7070	4.6536	8.5183	8.1142
		MAE	0.2187	2.6741	1.6745	1.6665	3.7312	3.7312
		MAPE	30.25%	240.70%	262.57%	394.14%	847.66%	1043.33%
	200	RMSE	0.3050	6.6565	4.6826	4.6254	8.5833	8.1799
		MAE	0.1056	2.6297	1.7794	1.7143	3.7751	3.7748
		MAPE	22.44%	281.76%	223.42%	310.56%	561.79%	7.9771
4	50	RMSE	1.3476	8.7788	6.3774	6.3631	11.0874	10.5588
		MAE	0.3549	2.5881	1.7217	1.6918	3.6354	3.6347
		MAPE	67.73%	411.74%	399.19%	340.79%	555.02%	666.35%
	60	RMSE	0.9575	8.3818	6.4136	6.3557	11.0783	10.5065
		MAE	0.2583	2.4837	1.7374	1.6900	3.6442	3.6442
		MAPE	52.85%	369.17%	385.83%	427.53%	512.52%	724.59%
	80	RMSE	0.9012	9.1751	6.6479	6.4663	11.4770	10.9055
		MAE	0.2456	2.7057	1.8493	1.7535	3.7679	3.7679
		MAPE	42.75%	493.39%	563.47%	366.48%	668.29%	779.03%
	110	RMSE	0.6786	9.0223	6.7077	6.5811	11.5489	10.9196
		MAE	0.1794	2.6686	1.8379	1.7598	3.7742	3.7728
		MAPE	34.24%	355.19%	315.79%	578.84%	502.77%	665.81%
	200	RMSE	0.2339	9.4260	6.8257	6.5914	11.5383	10.9348
		MAE	0.0622	2.8121	1.9014	1.8266	3.7849	3.7828
		MAPE	25.92%	545.75%	525.67%	3840.41%	721.89%	932.25%
5	50	RMSE	1.3976	11.1335	8.0977	7.9533	14.0285	13.2830
		MAE	0.3008	2.6140	1.8020	1.7424	3.6725	3.6724
		MAPE	49.38%	279.00%	288.97%	487.88%	340.87%	449.90%
	60	RMSE	0.7389	11.4702	8.4991	8.2285	14.1760	13.4179
		MAE	0.1583	2.7361	1.9256	1.8176	3.7208	3.7173
		MAPE	39.39%	341.03%	289.62%	282.19%	391.73%	515.92%
	80	RMSE	0.7176	12.2329	8.5463	8.3889	14.5388	13.8041
		MAE	0.1439	2.9031	1.9158	1.8250	3.8240	3.8240
		MAPE	27.43%	330.84%	278.25%	730.42%	432.22%	559.53%
	110	RMSE	0.5824	11.6102	8.5570	8.3539	14.4113	13.6339
		MAE	0.1068	2.7677	1.9434	1.8363	3.7690	3.7685
		MAPE	13.57%	264.40%	229.75%	543.45%	312.70%	398.99%
	200	RMSE	0.2669	11.0806	8.1079	7.7729	13.8221	13.0660
		MAE	0.0456	2.6438	1.8182	1.7094	3.6218	3.6209
		MAPE	13.95%	586.85%	576.79%	515.46%	496.98%	807.62%
10	50	RMSE	1.1203	20.2995	15.6622	14.5914	24.7968	22.8664
		MAE	0.1023	2.3727	1.7810	1.6260	3.1578	3.1488
		MAPE	16.13%	426.58%	1066.80%	313.60%	608.16%	1029.91%
	60	RMSE	0.7281	18.7881	15.3222	14.2872	23.5393	21.5348
		MAE	0.0634	2.1991	1.7204	1.5959	2.9665	2.9585
		MAPE	15.23%	560.06%	498.05%	412.03%	792.66%	1094.93%
	80	RMSE	0.5782	20.0287	15.5950	14.4913	24.5732	22.7300
		MAE	0.0510	2.3632	1.7720	1.6288	3.1355	3.1303
		MAPE	9.08%	493.00%	346.63%	766.65%	504.31%	656.37%
	110	RMSE	0.3750	18.9581	15.1083	13.9556	23.7643	21.8498
		MAE	0.0303	2.2188	1.7062	1.5569	3.0175	3.0116
		MAPE	6.13%	615.14%	689.25%	640.20%	601.42%	1050.15%
	200	RMSE	0.1549	17.5049	13.9856	13.0382	22.3069	20.3454
		MAE	0.0135	2.0353	1.5784	1.4516	2.8053	2.7961
		MAPE	2.80%	644.37%	528.91%	638.41%	496.67%	784.75%

Table 2
The RMSE and MAE for DGPII.

T	n	measures	P-LM	P-CM	P-CRM	P-Minmax	NWM	LLM
3	50	RMSE	1.6847	6.5357	1.7033	1.7062	11.0164	10.8856
		MAE	0.4682	2.6402	0.4624	0.4436	4.9368	4.9369
		MAPE	46.37%	217.03%	62.18%	69.89%	251.47%	263.99%
	60	RMSE	1.5579	6.6775	1.5547	1.5595	10.9793	10.8463
		MAE	0.4313	2.6911	0.4313	0.4084	4.9429	4.9429
		MAPE	41.12%	218.47%	62.95%	61.51%	271.70%	288.24%
	80	RMSE	1.4352	6.5901	1.5598	1.5520	10.7789	10.6441
		MAE	0.3936	2.6580	0.4466	0.4068	4.8512	4.8512
		MAPE	38.91%	246.56%	56.37%	54.63%	266.51%	281.29%
	110	RMSE	1.5046	6.6675	1.4906	1.4902	10.9434	10.7899
		MAE	0.4452	2.7280	0.3935	0.3794	1.8272	4.9189
		MAPE	32.55%	229.75%	55.39%	53.46%	289.99%	309.40%
	200	RMSE	1.4451	6.6765	1.4416	1.4443	10.9533	10.7829
		MAE	0.4554	2.6932	0.3976	0.3795	4.9269	4.9260
		MAPE	25.55%	228.67%	50.39%	52.43%	287.38%	365.30%
4	50	RMSE	1.5159	8.6493	2.3411	2.3577	14.4736	14.2641
		MAE	0.3864	2.6404	0.4498	0.4212	4.8779	4.8779
		MAPE	42.29%	320.94%	69.71%	71.19%	282.39%	371.18%
	60	RMSE	1.5214	9.0040	2.3151	2.3122	14.4930	14.2453
		MAE	0.4101	2.7071	0.4694	0.4467	4.9002	4.8990
		MAPE	37.95%	336.69%	63.57%	65.26%	212.65%	344.26%
	80	RMSE	1.3944	8.8546	2.3381	2.3436	14.7135	14.4496
		MAE	0.3771	2.7172	0.4621	0.4416	4.9530	4.9530
		MAPE	31.54%	382.93%	61.26%	58.72%	285.49%	449.03%
	110	RMSE	1.1812	8.8191	2.3524	2.3523	14.6146	14.3075
		MAE	0.3227	2.6974	0.4642	0.4597	4.8954	4.8951
		MAPE	24.21%	392.94%	59.76%	62.99%	306.96%	522.96%
	200	RMSE	0.8801	8.8151	2.2785	2.2791	14.6868	14.3963
		MAE	0.2283	2.6818	0.441	0.4316	4.9457	4.9443
		MAPE	18.17%	323.18%	54.76%	55.86%	330.88%	385.86%
5	50	RMSE	1.3820	10.7204	3.4341	3.4367	18.1689	17.7923
		MAE	0.2835	2.5994	2.6147	0.5398	0.5307	4.9083
		MAPE	22.54%	438.27%	62.18%	68.69%	422.70%	504.89%
	60	RMSE	1.3524	11.0094	3.3265	3.3225	18.1558	17.7991
		MAE	0.2962	2.6815	0.5337	0.5147	4.9007	4.9007
		MAPE	22.88%	275.46%	63.51%	58.12%	276.57%	331.28%
	80	RMSE	1.2942	11.1237	3.2700	3.2662	18.4882	18.1171
		MAE	0.2621	2.7341	0.5368	0.5225	4.9975	4.9959
		MAPE	17.84%	251.74%	62.27%	55.43%	262.59%	309.38%
	110	RMSE	1.0946	10.9671	3.2403	3.2299	18.1839	17.7603
		MAE	0.2210	2.6720	0.5318	0.4956	4.8887	4.8880
		MAPE	14.75%	223.46%	51.58%	54.36%	339.59%	405.88%
	200	RMSE	0.7475	11.1279	3.2320	3.2322	18.5084	18.0813
		MAE	0.1326	2.7102	0.5102	0.5046	4.9741	4.9722
		MAPE	9.53%	273.44%	49.65%	48.36%	292.46%	353.00%
10	50	RMSE	1.1828	22.9577	7.9839	7.9793	36.1096	35.3445
		MAE	0.1036	2.8142	0.6417	0.6232	4.8968	4.8968
		MAPE	6.76%	427.04%	179.89%	174.56%	271.54%	421.42%
	60	RMSE	1.1739	22.7227	7.6207	7.6329	34.9582	34.1529
		MAE	0.1124	2.7621	0.6114	0.5942	4.7460	4.7460
		MAPE	3.17%	432.68%	157.70%	141.16%	291.40%	436.63%
	80	RMSE	1.0737	21.5598	7.4378	7.4531	36.8868	36.1311
		MAE	0.0996	2.6463	0.6193	0.5906	5.0145	5.0143
		MAPE	2.63%	202.20%	95.29%	102.87%	236.03%	260.05%
	110	RMSE	0.8497	21.7974	7.2795	7.2758	36.6258	35.8193
		MAE	0.0737	2.6567	0.5817	0.5666	4.9731	4.9731
		MAPE	1.98%	224.43%	105.43%	120.32%	253.01%	281.34%
	200	RMSE	0.4425	22.2898	6.9306	6.9307	36.4988	35.6825
		MAE	0.0274	2.7261	0.5416	0.5389	4.9603	4.9603
		MAPE	1.69%	267.63%	119.96%	136.54%	281.01%	326.99%

nonlinear relationships). Table 2 report simulation results for DGPII (panel interval-valued data with linear relationships).

From Table 1, we have the following three findings. (1) The proposed estimator is consistent when dealing with the panel interval-valued data with nonlinear relationship, since the RMSE, MAE, and MAPE of the proposed model P-LM decrease as both n and T grow. For example, when fixing $T = 3$ and let $n = 50, 60, 80, 110$, and 200, the RMSE of the proposed model are 1.4885, 1.1315 (−0.357), 1.0485 (−0.083), 0.6129 (−0.4356), and 0.3050 (−0.3079), and the MAPE of the proposed model are 62.89%, 57.75% (−5.14%), 42.96% (−19.93%), 30.25%(−32.64%), 22.44% (−40.45%) (the number in parentheses represent the gap between current sample size and previous sample size). When fixing $n = 80$ and let $T = 3, 4, 5, 10$, the RMSE of the proposed model are 0.3891, 0.3154 (−0.1473), 0.7176 (−0.1784), and 0.5782 (−0.1394), and the MAPE of the proposed model are 42.96%, 42.75% (−0.21%), 27.43% (−15.53%), 9.08% (−33.88%)(the number in parentheses represent the gap between current sample size and previous sample size). (2) The estimators derived from P-CM, P-Minmax, P-CRM, NWM, and LLM may be not consistent when dealing with the nonlinear panel interval-valued data, since the RMSE, MAE, and MAPE of the models P-CM, P-Minmax, P-CRM, NWM, and LLM have no apparent trend of decreasing in most case as n or T or both n and T grow. For the models P-CM, P-Minmax and P-CRM, the reason is that they are constructed based on linear specification. For the models NWM and LLM, the reason may be that the features used by these models are insufficient to describe the true nature of the interval-valued data, that is, the interval range feature is overlooked. (3) For given n and T, our proposed model P-LM performs well compared with the models P-CM, P-Minmax, P-CRM, NWM, and LLM. For example, when $n = 110$ and $T = 10$, the MAE of the models P-LM, P-CM, P-Minmax, P-CRM, NWM, and LLM are 0.0303, 2.2188 (+2.1885), 1.7062 (+1.6759), 1.5569 (+1.5266), 3.0175 (+2.9872), and 3.0116 (+2.9812), and the MAPE of the models P-LM, P-CM, P-Minmax, P-CRM, NWM, and LLM are 6.13%, 615.14% (+609.01%), 689.25% (+683.12%), 640.20% (+634.07%), 601.42% (+595.29%), and 1050.15% (+1044.02%) (the number in parentheses represent the gap between the proposed model (P-LM) and the corresponding model). Moreover, the models P-LM, P-CM, P-Minmax, P-CRM, NWM, and LLM show an ever lower predictive validity. The range of MAPE of the models P-LM, P-CM, P-Minmax, P-CRM, NWM, and LLM are greater than 100%.

Table 2 displays the RMSE, MAE, and MAPE for DGPII. Four main findings are listed as follows. (1) The RMSE, MAE, and MAPE of the models P-LM, P-CRM, P-Minmax decrease in most case as n fix and T grow, indicating that the three corresponding estimators are consistent. For instance, when $T = 5$, $n = 50, 60, 80, 110$, and 200, the RMSE of the model P-LM are 1.3820, 1.3524 (−0.0296), 1.2942 (−0.0582), 1.0946 (−0.1996), and 0.7475 (−0.3471) (the number in parentheses represent the gap between current sample size and previous sample size), the RMSE of the model P-CRM are 3.4341, 3.3262 (−0.1079), 3.2700 (−0.0562), 3.2403 (−0.0297), and 3.2320(−0.0083), and the RMSE of the model P-Minmax are 3.4367, 3.3225 (−0.1135), 3.2662 (−0.0563), 3.2299 (−0.0363), 3.2322 (+0.0023). While, the RMSE and MAE of the models P-CM, NWM, and LLM have no apparent trend of increasing or decreasing in most case as n or T or both n and T grow, indicating that the estimators may be not consistent when dealing the panel interval-valued data with linear relationship. (2) The RMSE and MAE of the model P-LM decrease in most case as n or T or both n and T grow, indicating that the corresponding estimator is consistent. (3) For given the small sample n and T, the models P-LM, P-CRM, P-Minmax can be outperform the other two models NWM and LLM. For example, when $T = 3$, and n increases, the prediction accuracy for the models P-LM, P-CRM, P-Minmax have a small gap, and the RMSE and MAE of the models P-LM, P-CRM, P-Minmax are significantly smaller than those of other three models, P-CM, NWM and LLM. The reason for the simulation result of the models is that the interval range feature for these models is overlooked. (4) The proposed model P-LM is better than all the other models as T increases. For instance, when $T = 3$, and n increases, the MAPE range of P-LM is less than 46.37%, but the MAPE range of the models P-CM, P-CRM, P-Minmax, NWM, and LLM is greater than or equal to 50.39%, and when $T = 10$, and n increases, the MAPE range of P-LM is less than 6.76%, but the MAPE range of the models P-CM, P-CRM, P-Minmax, NWM, and LLM is greater than 100%.

Table 3
The RMSE and MAE for air quality dataset.

Models	RMSE	MAE	MAPE
P-LM	0.3982	0.3761	2.37%
P-CM	0.9450	1.6544	35.41%
P-CRM	0.4227	0.4359	2.57%
P-Minmax	0.4226	0.4336	2.57%
NWM	1.4370	2.5942	55.32%
LLM	1.4267	2.5942	55.80%

6. Real data analysis

This section uses two real panel datasets (air quality dataset and stock price dataset) to compare the models P-LM, P-CM, P-Minmax, P-CRM, NWM, and LLM.

6.1. Air quality dataset

Interval-valued observations about air quality are considered in [11]. This dataset collects four cities (Hangzhou, Chongqing, Beijing, and Tianjin for the past nine years) over nine years (2014–2022) with five variables: the annual air quality index (AQI), particulate matter of less than 10 and 2.5 μm in diameter (PM 10 and PM 2.5), sulfur dioxide (SO2), nitrogen dioxide (NO2). We treat the AQI (y_{it}) as response variable, the PM 2.5 (x_{it1}), PM 10 (x_{it2}), SO2 (x_{it3}), and NO2 (x_{it4}) as explanatory variables. All data are normalized for ease of calculation.

We establish our proposed model P-LM and other models P-CM, P-CRM, P-Minmax, NWM, LLM, and we calculate the RMSR, MAE, and MAPE for these models. The experiment results are shown in Table 3. It is clear that our proposed model P-LM has the best performance. Except for model P-LM, the models P-CRM and P-Minmax show better prediction performance than the models P-CM, NWM, LLM. In fact, it proves once again that ignoring the range inherent in interval-valued data can potentially lead to biased predictive responses.

6.2. Stock price dataset

This dataset contains the daily highest and lowest stock prices of 35 transportation companies in the Chinese road transportation industry from January 1st, 2021 to January 10th, 2021 (8 trading days). The datasets can be found at https://vip.stock.finance.sina.com.cn/mkt/. The highest and lowest prices of a stock serve several purposes in the stock market, such as trend analysis, support and resistance levels, and volatility assessment. They are essential indicators in the stock market, offering insights into price movements and market activity and assisting investors in making informed investment decisions. We study the influence of the previous day's highest and lowest stock prices on the next day's highest and lowest prices. Let the previous day's highest and lowest prices as interval explanatory variables (x_{it}) and the next day's highest and lowest prices as interval response variable (y_{it}), we construct a first-order lag panel interval-valued data model. Fig. 1 illustrates the stock price dataset.

The comparison results between the models P-LM, P-CM, P-CRM, P-Minmax, NWM, and LLM for stock price dataset are shown in Table 4. It presents that our proposed model P-LM exhibits the best prediction accuracy than other models, since our proposed model P-LM has the smallest values of RMSE, MAE, and MAPE than other models P-CM, P-CRM, P-Minmax, NWM, and LLM.

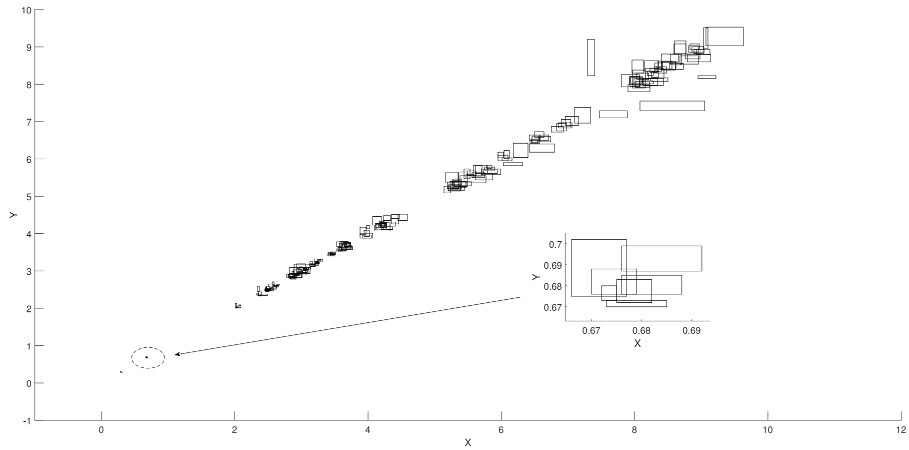

Fig. 1. Panel interval-valued data plot for stock price.

Table 4
The RMSE and MAE for stock price dataset.

Models	RMSE	MAE	MAPE
P-LM	0.1252	0.1386	2.16%
P-CM	4.8636	5.7796	104.01%
P-CRM	0.1620	0.1668	2.52%
P-Minmax	0.1725	0.1725	2.60%
NWM	0.2116	0.2245	3.52%
LLM	0.1645	0.1879	3.07%

7. Conclusions

For the observations of panel data that are characterized by intervals rather than single point values, we established a fixed effects panel interval-valued data model based on the center and range of intervals. Under the profile least squares framework, we further proposed a locally linear method for the proposed model. As far as we know, our proposed model and corresponding estimation method present the first endeavor to use the full interval information dealing with panel interval-valued data with the nonlinear relationship. Thus, compared with P-CM, P-Minmax, and P-CRM, our proposed model and corresponding estimation method allow for more complex relationships between the response and explanatory variables. In addition, our proposed model does not make any assumptions on the form of a nonlinear component. Later, experimental results demonstrated that our proposed model and corresponding estimation method achieve high prediction and outperform the other models in most case.

Although our proposed estimation method considers the nonnegativity of the range of predictive response, as shown in Eq. (33), the employing transformation in the last procedure is bunt and thereby may be leading to a biased predictive response. Following this, our future work will consider further exploring the nonnegativity of the range of predictive response, similar to the cross-sectional model for interval-valued data [51,52], imposing the restriction or adding the nonnegative constraint on the proposed model.

CRediT authorship contribution statement

Jinjin Zhang: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Aibing Ji:** Writing – review & editing, Validation, Resources, Project administration, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and company that could be construed as influencing the position presented in, or the review of, the manuscript.

Data availability

Data will be made available on request.

Acknowledgments

The authors gratefully acknowledge the editor and three anonymous reviewers for their useful feedback that improved this paper. This paper was supported by the National Statistical Science Research Project (No. 2023LY025); the Operations Research and Management Innovation Team of Hebei University (No. IT2023C02); and the Innovation Capacity Enhancement Program-Science and Technology Platform Project, Hebei Province (No. 22567623H).

Appendix

A.1. Proof of Theorem 1

Considering the estimator of $m_1(x^c)$

$$\begin{aligned}\hat{m}_1(x^c) &= s_h^c(x^c)' M^c Y^c\\ &= s_h^c(x^c)' M^c (m_1(X^c) + D\mu^c + \varepsilon)\\ &= s_h^c(x^c)' M^c m_1(X^c) + s_h^c(x^c)' M^c \varepsilon\end{aligned}$$

where $s_h^c(x^c)' = e_1'[(\tilde{X}^c)' W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)' W_1(x^c)$ and $M^c = I - D[D'G^c D]^{-1} D' G^c$. In fact, $M^c m_1(X^c) D\mu^c = 0$ is used in the third equality. Define $\hat{m}_{11}(x^c) = s_h^c(x^c)' M^c m_1(X^c)$, and $m_{12}(x^c) = s_h^c(x^c)' M^c \varepsilon$. To derive the asymptotic distribution of $m_1(x^c)$, we provide the following Lemmas 1, 2, and 3, and derive the asymptotic distribution of $m_{11}(x^c)$ and $m_{12}(x^c)$.

Lemma 1. *Under Assumptions* $(A1)-(A4)$*,*

$$(nT)^{-1}(\tilde{X}^c)' W_1(x^c)\tilde{X}^c = \begin{pmatrix} f(x) + o_p(1) & \kappa_2 m_1(x^c)' H^2 + o_p(H^2 l_p) \\ \kappa_2 H^2 m_1(x^c) + o_p(H^2 l_p) & \kappa_2 f(x^c) H^2 + o_p(H^2) \end{pmatrix} \tag{A.1}$$

$$(D'G^cD)^{-1} = (D'D)^{-1} + O_p(c_n) \tag{A.2}$$

where $c_n = \|H\|^2 + \sqrt{(nT|H|)^{-1/2}}$

Proof. Recall the kernel density estimation

$$(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} K_H(x_{it}^c, x^c) = f(x^c) + O_p(\|H\|^2) + O_p((nT|H|)^{-1/2}) \tag{A.3}$$

Under Assumption (A4), the right-hand side of equality turns into $f(x^c) + o_p(1)$. The detailed proofs of (A.1) and (A.2) can be refer to [14,15,49]. □

Lemma 2. *Under Assumptions (A1) − (A4),*

$$\hat{m}_{11}(x^c) = m_1(x^c) + B_H \tag{A.4}$$

where $B_H = \frac{1}{2}\kappa_2 tr\left\{Hm_{1,xx}H\right\} + O_p(\|H\|^4) + O_p((nT|H|)^{-1/2}\|H\|)$ *is the bias term of* $\hat{m}_1(x^c)$.

Proof. We decompose $m_{11}(x^c)$ into two parts below.

$$\begin{aligned}\hat{m}_{11}(x^c) =& s_h^c(x^c)'M^c m_1(X^c)\\ =& s_h^c(x^c)'m_1(X^c) - s_h^c(x^c)'D[D'G^cD]^{-1}D'G^c m_1(X^c)\\ \equiv& \hat{\theta}_1(x^c) - \hat{\theta}_2(x^c)\end{aligned}$$

where $\hat{\theta}_1(x^c) = s_h^c(x^c)'m_1(X^c)$ and $\hat{\theta}_2(x^c) = s_h^c(x^c)'D[D'G^cD]^{-1}D'G^c m_1(X^c)$. Further,

$$\hat{\theta}_1(x^c) = m_1(x^c) + \frac{1}{2}\kappa_2 tr\left\{Hm_{1,xx}H\right\} + O_p(\|H\|^4) + O_p((nT|H|)^{-1/2}\|H\|) \tag{A.5}$$

$$\begin{aligned}\hat{\theta}_2(x^c) =& \frac{1}{n^2T}(1 - \kappa_2 f(x^c)^{-2}m'_{1,x}H^2 m_{1,x})\sum_{i=2}^{n}\sum_{t=1}^{T} m_1(x_{it}^c)(1 + c_n)\\ &- 2\frac{n-1}{n^3T}(1 - \kappa_2 f(x^c)^{-2}m'_{1,x}H^2 m_{1,x})^2\sum_{i=2}^{n}\sum_{t=1}^{T} m_1(x_{it}^c)(1 + c_n)\\ &+ \frac{(n-1)^2}{n^4T}(1 - \kappa_2 f(x^c)^{-2}m'_{1,x}H^2 m_{1,x})^3\sum_{i=2}^{n}\sum_{t=1}^{T} m_1(x_{it}^c)(1 + c_n)\end{aligned} \tag{A.6}$$

It shows that the expectation of $\hat{\theta}_2(x^c)$ converges to 0 as $n \to \infty$. We proof (A.5) and (A.6) below. We first consider $\hat{\theta}_1(x^c)$. Using a Taylor expansion for the function $m_1(x_{it}^c)$ around this point x^c and converting into a matrix form, we have

$$\begin{aligned}m_1(X^c) =& l_{nT}m_1(x^c) + (X^c - l_{nT}x^c)m_{1,x} + \frac{1}{2}Q_m + R_m\\ =& \tilde{X}^c\begin{bmatrix} m_1(x^c)\\ m_{1,x}\end{bmatrix} + \frac{1}{2}Q_m + R_m\end{aligned}$$

where $Q_m = (X^c - l_{nT}x^c)m_{1,XX}(X^c - l_{nT}x^c)'$, $m_{1,XX}$ is a $p\times p$ dimensional diagonal matrix whose diagonal elements are all Hessian matrix $m_1(x^c)$ evaluated at x^c and denoted as $m_{1,xx}$. R_m is the Taylor series remainder terms. Plugging this back to $\hat{\theta}_1(x^c)$,

$$\begin{aligned}\hat{\theta}_1(x^c) =& s_h^c(x^c)'m_1(X^c)\\ =& m_1(x^c) + \frac{1}{2}s_h^c(x^c)'Q_m + s_h^c(x^c)'R_m\end{aligned}$$

R_m left multiplied by $e_1'[(\tilde{X}^c)'W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)'W_1(x^c)$ is of negligible order compared to the term Q_m left multiplied by same, and equals to $O(\|H\|^4)$. Also,

$$\begin{aligned}s_h^c(x^c)'Q_m =& e_1'[(\tilde{X}^c)'W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)'W_1(x^c)(X^c - l_{nT}x^c)m_{1,XX}(X^c - l_{nT}x^c)'\\ =& e_1'[(nT)^{-1}(\tilde{X}^c)'W_1(x^c)\tilde{X}^c]^{-1}\\ &\begin{bmatrix}(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} K_H(x_{it}^c, x^c)(x_{it}^c - x^c)'m_{1,xx}(x_{it}^c - x^c)\\ (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\{K_H(x_{it}^c, x^c)(x_{it}^c - x^c)'m_{1,xx}(x_{it}^c - x^c)\right\}(x_{it}^c - x^c)\end{bmatrix}\end{aligned}$$

Using Lemma 1, the expectation of $s_h^c(x^c)'Q_m$ is

$$\mathbb{E}(s_h^c(x^c)'Q_m) = \kappa_2 tr\left\{Hm_{1,xx}H\right\} + O(\|H\|^4)$$

and for the variance, using the same method, it is straightforward to show that $\text{Var}(s_h^c(x^c)'Q_m) = O((nT|H|)^{-1}\|H\|^2)$. Hence,

$$\hat{\theta}_1(x^c) = m_1(x^c) + \frac{1}{2}\kappa_2 tr\left\{Hm_{1,xx}H\right\} + O_p(\|H\|^4) + O_p((nT|H|)^{-1/2}\|H\|)$$

The proof of (A.5) is complete.

Let us consider $\hat{\theta}_2(x^c)$ below,

$$\begin{aligned}\hat{\theta}_2(x^c) &= s_h^c(x^c)'D[D'G^cD]^{-1}D'G^c m_1(X^c)\\ &= s_h^c(x^c)'D[D'G^cD]^{-1}D'[I_{nT} - S_H^c]'[I_{nT} - S_H^c]m_1(X^c)\\ &\equiv \hat{\theta}_{21}(x^c) - \hat{\theta}_{22}(x^c) - \hat{\theta}_{23}(x^c) + \hat{\theta}_{24}(x^c) + o_p(1)\end{aligned}$$

where

$$\begin{aligned}\hat{\theta}_{21}(x^c) =& s_h^c(x^c)'D[D'D]^{-1}D'm_1(X^c)\\ \hat{\theta}_{22}(x^c) =& s_h^c(x^c)'D[D'D]^{-1}D'(S_H^c)'m_1(X^c)\\ \hat{\theta}_{23}(x^c) =& s_h^c(x^c)'D[D'D]^{-1}D'S_H^c m_1(X^c)\\ \hat{\theta}_{24}(x^c) =& s_h^c(x^c)'D[D'D]^{-1}D'(S_H^c)'S_H^c m_1(X^c)\end{aligned}$$

The common matrix $s_h^c(x^c)'D[D'D]^{-1}D'$ in estimator $\hat{\theta}_{21}(x^c)$, $\hat{\theta}_{22}(x^c)$, $\hat{\theta}_{23}(x^c)$, and $\hat{\theta}_{24}(x^c)$ is given by

$$s_h^c(x^c)'D[D'D]^{-1}D' = \frac{1}{n^2T}(1 - \kappa_2 f(x^c)^{-2}m'_{1,x}H^2 m_{1,x})e'(1 + c_n) \tag{A.7}$$

where e is an nT dimensional vector whose the first T elements are zeros and the rest elements are ones. In fact,

$$s_h^c(x^c)'D[D'D]^{-1}D' = e_1'[(\tilde{X}^c)'W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)'W_1(x^c)D[D'D]^{-1}D'$$

where the derivation of $(nT)^{-1}(\tilde{X}^c)'W_1(x^c)$ is given in Box I, and using Lemma 1, we can obtain (A.7). Thus,

$$\hat{\theta}_{21}(x^c) = \frac{1}{n^2T}(1 - \kappa_2 f(x^c)^{-2}m'_{1,x}H^2 m_{1,x})\sum_{i=2}^{n}\sum_{t=1}^{T} m_1(x_{it}^c)(1 + c_n)$$

$$\hat{\theta}_{22}(x^c) = \hat{\theta}_{23}(x^c) = \frac{n-1}{n^3T}(1 - \kappa_2 f(x^c)^{-2}m'_{1,x}H^2 m_{1,x})^2 \times \sum_{i=1}^{n}\sum_{t=1}^{T} m_1(x_{it}^c)(1 + c_n)$$

$$\hat{\theta}_{24}(x^c) = \frac{(n-1)^2}{n^4T}(1 - \kappa_2 f(x^c)^{-2}m'_{1,x}H^2 m_{1,x})^3\sum_{i=1}^{n}\sum_{t=1}^{T} m_1(x_{it}^c)(1 + c_n)$$

The proof of (A.6) is complete. □

Lemma 3. *Under Assumptions (A1) − (A4),*

$$\sqrt{nT|H|}\hat{m}_{12}(x^c) \xrightarrow{d} N(0, \frac{\varsigma_0\sigma_{\epsilon^c}^2}{f(x^c)}) \tag{A.8}$$

Proof. Since

$$\begin{aligned}\hat{m}_{12}(x^c) &= s_h^c(x^c)'M^c\varepsilon\\ &= e_1'[(\tilde{X}^c)'W_1(x^c)\tilde{X}^c]^{-1}(\tilde{X}^c)'W_1(x^c)M^c\varepsilon\end{aligned}$$

Using the kernel density function, we obtain

$$\mathbb{E}(\hat{m}_{12}(x^c)) = 0$$

$$\text{Var}(\hat{m}_{12}(x^c)) = \frac{\varsigma_0\sigma_{\epsilon^c}^2}{f(x^c)} + O(\|H\|)$$

Combining Lemmas 1, 2, and 3, we prove Eq. (41) in Theorem 1. Similarly, using the same procedure, we can prove Eq. (42) in Theorem 1. Note that the asymptotic distribution of $m_1(x^r)$ can be derived in the same way as the asymptotic distribution of $m_1(x^c)$. Then, the proof of in Theorem 1 is complete. □

$$(nT)^{-1}(\tilde{X}^c)'W_1(x^c) = \begin{bmatrix} (nT)^{-1}K_H(x_{11}^c, x^c) & (nT)^{-1}K_H(x_{12}^c, x^c) & \cdots & (nT)^{-1}K_H(x_{nT}^c, x^c) \\ (nT)^{-1}K_H(x_{11}^c, x^c)(x_{11}^c - x^c) & (nT)^{-1}K_H(x_{12}^c, x^c)(x_{12}^c - x^c) & \cdots & (nT)^{-1}K_H(x_{nT}^c, x^c)(x_{nT}^c - x^c) \end{bmatrix}$$

Box I.

A.2. Proof of Theorem 2

Under Assumptions $(A1)-(A4)$, when $H = hI_q$, by Eqs. (A.5) and (A.8), we have

$$\mathbb{E}(\hat{m}_1(x^c) - m_1(x^c)) = \frac{1}{2}\kappa_2 h^2 m_{1,xx} + O(h^4) + O((nT)^{-\frac{1}{2}}h^{1-\frac{q}{2}})$$

$$\mathrm{Var}(\hat{m}_1(x^c) - m_1(x^c)) = \frac{\varsigma_0 \sigma_{\varepsilon^c}^2}{nTh^q f(x^c)} + O((nT)^{-1}h^{-q+1})$$

Using the same method in deriving $\mathbb{E}(\hat{m}_1(x^c) - m_1(x^c))$ and $\mathrm{Var}(\hat{m}_1(x^c) - m_1(x^c))$, we have

$$\mathbb{E}(\hat{m}_2(x^r) - m_2(x^r)) = \frac{1}{2}\kappa_2 h^2 m_{2,xx} + O(h^4) + O((nT)^{-\frac{1}{2}}h^{1-\frac{q}{2}})$$

$$\mathrm{Var}(\hat{m}_2(x^r) - m_2(x^r)) = \frac{\varsigma_0 \sigma_{\varepsilon^c}^2}{nTh^q f(x^c)} + O((nT)^{-1}h^{-q+1})$$

The goodness-of-fit criterion between $(m_1(x^c), m_2(x^r))$ and $(\hat{m}_1(x^c), \hat{m}_2(x^r))$ is usual mean integrated square error (MISE),

$$\mathrm{MISE}(\hat{m}_1(x^c), \hat{m}_2(x^r)) = \mathbb{E}\left\{\int (m_1(x^c) - \hat{m}_1(x^c))^2 dx^c + \int (m_2(x^r) - \hat{m}_2(x^r))^2 dx^r\right\}$$

By using the Fubini's theorem, we can show that

$$\mathrm{MISE}(\hat{m}_1(x^c), \hat{m}_2(x^r)) = \mathrm{IV}(\hat{m}_1(x^c)) + \mathrm{IV}(\hat{m}_2(x^r)) + \mathrm{IBS}(\hat{m}_1(x^c)) + \mathrm{IBS}(\hat{m}_2(x^r))$$

where $\mathrm{IV}(\cdot)$ is the integrated variance and $\mathrm{IBS}(\cdot)$ is the integrated square bias.

$$\mathrm{IV}(\hat{m}_1(x^c)) = \int \mathrm{Var}(\hat{m}_1(x^c) - m_1(x^c))dx^c$$

$$\mathrm{IV}(\hat{m}_2(x^r)) = \int \mathrm{Var}(\hat{m}_2(x^r) - m_2(x^r))dx^c$$

$$\mathrm{IBS}(\hat{m}_1(x^c)) = \int (\mathbb{E}(\hat{m}_1(x^c)) - m_1(x^c))^2 dx^c$$

$$\mathrm{IBS}(\hat{m}_2(x^r)) = \int (\mathbb{E}(\hat{m}_2(x^r)) - m_2(x^r))^2 dx^r$$

Now putting both bias and variance together, we obtain the MISE of $\hat{m}_1(x^c)$ and $\hat{m}_2(x^r)$:

$$\begin{aligned}\mathrm{MISE}(\hat{m}_1(x^c), \hat{m}_2(x^r)) = {} & \frac{\kappa_2^2}{4}\cdot h^4 \cdot \left\{\int m_{1,xx}m'_{1,xx}dx^c + \int m_{2,xx}m'_{2,xx}dx^r\right\} \\ & + \frac{\varsigma_0}{nTh^q}\cdot\left\{\sigma_{\varepsilon^c}^2\int (f(x^c))^{-1}dx^c + \sigma_{\varepsilon^r}^2 \int (f(x^r))^{-1}dx^r\right\} \\ & + O_p(h^4) + O_p((nT)^{-\frac{1}{2}}h^{1-\frac{q}{2}}) + O((nT)^{-1}h^{-q+1})\end{aligned}$$

The first two term,

$$\frac{\kappa_2^2}{4}\cdot h^4 \cdot \left\{\int m_{1,xx}m'_{1,xx}dx^c + \int m_{2,xx}m'_{2,xx}dx^r\right\} + \frac{\varsigma_0}{nTh^q} \cdot\left\{\sigma_{\varepsilon^c}^2\int (f(x^c))^{-1}dx^c + \sigma_{\varepsilon^r}^2 \int (f(x^r))^{-1}dx^r\right\}$$

is called the asymptotical mean integrated square error (AMISE). Thus, the smoothing bandwidth minimizing the AMISE is

$$h_{opt} = \left\{\frac{q}{nT}\cdot\frac{\left\{\sigma_{\varepsilon^c}^2\int (f(x^c))^{-1}dx^c + \sigma_{\varepsilon^r}^2\int (f(x^r))^{-1}dx^r\right\}}{\left\{\int m_{1,xx}m'_{1,xx}dx^c + \int m_{2,xx}m'_{2,xx}dx^r\right\}}\right\}$$

Thus, the proof of Theorem 2 is complete.

References

[1] J.M. Wooldridge, Econometric Analysis of Cross Section and Panel Data, The MIT Press, 2010.
[2] B.H. Baltagi, The one-way error component regression model, in: Econometric Analysis of Panel Data, Springer International Publishing, Cham, 2021, pp. 15–45.
[3] C. Hsiao, Analysis of Panel Data, fourth ed., in: Econometric Society Monographs, Cambridge University Press, 2022.
[4] Y. Sun, B. Huang, A. Ullah, S. Wang, Nonparametric estimation and forecasting of interval-valued time series regression models with constraints, Expert Syst. Appl. 249 (2024) 123385.
[5] S.E. Culver, D.H. Papell, Is there a unit root in the inflation rate? Evidence from sequential break and panel data models, J. Appl. Econometrics 12 (4) (1997) 435–444.
[6] B. Hamiye Beyaztas, S. Bandyopadhyay, Robust estimation for linear panel data models, Stat. Med. 39 (29) (2020) 4421–4438.
[7] H.H.Z. Thom, C.H. Jackson, D. Commenges, L.D. Sharples, State selection in Markov models for panel data with application to psoriatic arthritis, Stat. Med. 34 (16) (2015) 2456–2475.
[8] T.D. Hill, A.P. Davis, J.M. Roos, M.T. French, Limitations of fixed-effects models for panel data, Sociol. Perspect. 63 (3) (2020) 357–369.
[9] M. Collischon, A. Eberl, Let's talk about fixed effects: Let's talk about all the good things and the bad things, Sociol. Perspect. 72 (2020) 289–299.
[10] C.J.D.A. Hanck, J.-S. Pischke, Mostly harmless econometrics: An empiricist's companion, Statist. Papers 52 (2009) 503–504.
[11] A. Ji, J. Zhang, X. He, Y. Zhang, Fixed effects panel interval-valued data models and applications, Knowl.-Based Syst. 237 (2022) 107798.
[12] D.J. Henderson, R.J. Carroll, Q. Li, Nonparametric estimation and testing of fixed effects panel data models, J. Econometrics 144 (1) (2008) 257–275.
[13] J. Qian, L. Wang, Estimating semiparametric panel data models by marginal integration, J. Econometrics 167 (2) (2012) 483–493, Fourth Symposium on Econometric Theory and Applications (SETA).
[14] Y. Gao, K. Li, Nonparametric estimation of fixed effects panel data models, J. Nonparametr. Stat. 25 (3) (2013) 679–693.
[15] Y. Lee, D. Mukherjee, A. Ullah, Nonparametric estimation of the marginal effect in fixed-effect panel data models, J. Multivariate Anal. 171 (2019) 53–67.
[16] G. Li, H. Peng, T. Tong, Simultaneous confidence band for nonparametric fixed effects panel data models, Econom. Lett. 119 (3) (2013) 229–232.
[17] L.A. Arteaga-Molina, J.M. Rodriguez-Poo, Empirical likelihood based inference for fixed effects varying coefficient panel data models, J. Statist. Plann. Inference 196 (2018) 144–162.
[18] S.C. Halder, E. Malikov, Smoothed LSDV estimation of functional-coefficient panel data models with two-way fixed effects, Econom. Lett. 192 (2020) 109239.
[19] Q. Xu, Z. Cai, Y. Fang, Semiparametric inferences for panel data models with fixed effects via nearest neighbor difference transformation, Econometric Rev. 40 (10) (2021) 919–943.
[20] Z. Lin, Q. Li, Y. Sun, A consistent nonparametric test of parametric regression functional form in fixed effects panel data models, J. Econometrics 178 (2014) 167–179, Annals Issue: Misspecification Test Methods in Econometrics.
[21] R.E. Moore, Interval Analysis, Prentice-Hall, 1966.
[22] A. Sengupta, T.K. Pal, On comparing interval numbers, European J. Oper. Res. 127 (1) (2000) 28–43.
[23] X. Wang, Z. Zhang, S. Li, Set-valued and interval-valued stationary time series, J. Multivariate Anal. 145 (2016) 208–223.
[24] A.L.S. Maia, F. de A.T. de Carvalho, T.B. Ludermir, Forecasting models for interval-valued time series, Neurocomputing 71 (16) (2008) 3344–3352, Advances in Neural Information Processing (ICONIP 2006) / Brazilian Symposium on Neural Networks (SBRN 2006).
[25] P. Wang, S.H. Gurmani, Z. Tao, J. Liu, H. Chen, Interval time series forecasting: A systematic literature review, J. Forecast. (2023) 1–37.
[26] L.C. Souza, R.M. Souza, G.J. Amaral, T.M. Silva Filho, A parametrized approach for linear regression of interval data, Knowl.-Based Syst. 131 (2017) 149–159.
[27] F.d.A. de Carvalho, E.d.A. Lima Neto, K.C. da Silva, A clusterwise nonlinear regression algorithm for interval-valued data, Inform. Sci. 555 (2021) 357–385.
[28] W. Silva, R. Souza, F. Cysneiros, Bivariate elliptical regression for modeling interval-valued data, Comput. Statist. 37 (2022) 2003–2028.

[29] S.I. Rizo Rodríguez, F.d.A. Tenório de Carvalho, Clustering interval-valued data with adaptive Euclidean and City-Block distances, Expert Syst. Appl. 198 (2022) 116774.

[30] S.I.R. Rodríguez, F.d.T. de Carvalho, Clustering interval-valued data with automatic variables weighting, in: 2019 International Joint Conference on Neural Networks, IJCNN, 2019, pp. 1–8.

[31] M. Liu, H. Zhao, Z. Xu, et al., A novel projection-based distance measure for interval-valued intuitionistic multiplicative clustering algorithm, Soft Comput. 27 (2023) 2369–2383.

[32] S.I. Rodríguez, F. de A.T. de Carvalho, Soft subspace clustering of interval-valued data with regularizations, Knowl.-Based Syst. 227 (2021) 107191.

[33] J. Zhang, S. Liu, L.e.a. Li, The KKT optimality conditions in a class of generalized convex optimization problems with an interval-valued objective function, Optim. Lett. 8 (2014) 607–631.

[34] M.S. Rahman, A.A. Shaikh, A.K. Bhunia, Necessary and sufficient optimality conditions for non-linear unconstrained and constrained optimization problem with interval valued objective function, Comput. Ind. Eng. 147 (2020) 106634.

[35] N. Abdulaleem, Optimality conditions in a class of generalized convex optimization problems with the multiple interval-valued objective function, Syst. Soft Comput. 5 (2023) 200056.

[36] H. Choi, J. Lim, D.e.a.J. Yu, Two-sample tests for interval-valued data., J. Korean Statist. Soc. 50 (2021) 233–271.

[37] J. Dai, Y. Liu, J. Chen, X. Liu, Fast feature selection for interval-valued data through kernel density estimation entropy, Int. J. Mach. Learn. Cybern. 11 (2020) 2607–2624.

[38] B. Sang, H. Chen, L. Yang, T. Li, W. Xu, C. Luo, Feature selection for dynamic interval-valued ordered data based on fuzzy dominance neighborhood rough set, Knowl.-Based Syst. 227 (2021) 107223.

[39] L. Billard, E. Diday, Regression analysis for interval-valued data, in: H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, M. Schader (Eds.), Data Analysis, Classification, and Related Methods, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 369–374.

[40] L. Billard, E. Diday, Symbolic regression analysis, in: K. Jajuga, A. Sokołowski, H.-H. Bock (Eds.), Classification, Clustering, and Data Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 281–288.

[41] E.d.A. Lima Neto, F.d.A. de Carvalho, Centre and range method for fitting a linear regression model to symbolic interval data, Comput. Statist. Data Anal. 52 (3) (2008) 1500–1515.

[42] P. Giordani, Lasso-constrained regression analysis for interval-valued data, Adv. Data Anal. Classif. 9 (2015) 5–19.

[43] M. Xu, Z. Qin, A bivariate Bayesian method for interval-valued regression models, Knowl.-Based Syst. 235 (2022) 107396.

[44] E.A. Lima Neto, F.d.A.T. de Carvalho, Nonlinear regression applied to interval-valued data, Pattern Anal. Appl. 20 (2017) 809–824.

[45] E.d.A. Lima Neto, F.d.A. de Carvalho, An exponential-type kernel robust regression model for interval-valued variables, Inform. Sci. 454–455 (2018) 419–442.

[46] F.d.A. de Carvalho, E.d.A. Lima Neto, U.d.N. Rosendo, Interval joint robust regression method, Neurocomputing 465 (2021) 265–288.

[47] R.A. Fagundes, R.M. de Souza, F.J.A. Cysneiros, Interval kernel regression, Neurocomputing 128 (2014) 371–388.

[48] L. Kong, X. Song, X. Wang, Nonparametric regression for interval-valued data based on local linear smoothing approach, Neurocomputing 501 (2022) 834–843.

[49] L. Su, A. Ullah, Profile likelihood estimation of partially linear panel data models with fixed effects, Econom. Lett. 92 (1) (2006) 75–81.

[50] C.F. Parmeter, J.S. Racine, Chapter 4 - Nonparametric estimation and inference for panel data models, in: M. Tsionas (Ed.), Panel Data Econometrics, Academic Press, 2019, pp. 97–129.

[51] Y. Zhong, Z. Zhang, S. Li, A constrained interval-valued linear regression model: A new heteroscedasticity estimation method, J. Syst. Sci. Complex. 33 (2020) 2048–2066.

[52] E.d.A. Lima Neto, F.d.A. de Carvalho, Constrained linear regression models for symbolic interval-valued variables, Comput. Statist. Data Anal. 54 (2) (2010) 333–347.