Economics Letters 262 (2026) 112869

Contents lists available at ScienceDirect

Economics Letters

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

Determining the number of breaks in high-dimensional factor models with interval-valued data

Yan Guo [a,b], Jing Chen [c], Jianhong Wu [b,d,*]

[a] School of Mathematics and Statistics, Suzhou University of Technology, Suzhou 215500, China
[b] College of Mathematics and Science, Shanghai Normal University, Shanghai 200234, China
[c] College of Mathematics and Statistics, Qingdao University, Qingdao 266071, China
[d] Lab for Educational Big Data and Policymaking, Ministry of Education, China

ARTICLE INFO

JEL classification:
C1
C13

Keywords:
Approximate factor models
Interval-valued data
Structural changes
The number of breaks

ABSTRACT

The paper considers a high-dimensional interval-valued data factor model with potential structural changes. We verify that the number of factors is usually overestimated in the case with structural changes and then propose an estimator for the number of breaks by leveraging the finding. We establish the consistency of this estimator under certain conditions. Monte Carlo simulation results show that the proposed estimation procedure exhibits desired finite sample performance. In addition, an empirical application to S&P 100 stock return data further demonstrates the practical usefulness of the proposed method.

1. Introduction

Interval-valued data is currently a major type of symbolic data (Billard and Diday, 2003). Unlike traditional point-valued data, interval-valued data not only contains information about the level of the variable but also provides information about the variability of the variable. Therefore, interval-valued data generally provides richer information. For example, Fischer et al. (2016) formally define return intervals, a type of interval-valued financial data that captures the range of asset returns within a day. Building on this, Sun et al. (2020) propose a conditional heteroskedasticity model for such return intervals and show that interval-valued volatility models can outperform traditional GARCH approaches, confirming that interval-valued data indeed provides additional useful information. Over the last two decades, extensive research has been conducted in the field of interval-valued data. Classic methods for interval-valued data include the center method (Billard and Diday, 2000), the center and range method (De Carvalho et al., 2004; Lima Neto and De Carvalho, 2008), and the minimum–maximum method (Billard and Diday, 2002). Researchers have also developed many new methods based on these, as detailed in Lima Neto and De Carvalho (2010) and González-Rivera and Lin (2013), among others. Besides these point-process-based methods for interval-valued data, the last decade has also seen research treating intervals as sets and analyzing them using random set theory, such as Han et al. (2012, 2016), Sun (2017) and Sun et al. (2018, 2020).

Most of the methods mentioned above are based on regression models. In contrast, to better handle high-dimensional data, Guo et al. (2025a,b) analyze high-dimensional interval-valued data from the perspective of factor analysis. They propose an approximate factor model for high-dimensional interval-valued data and develop estimation procedures for the number of factors, interval-valued factors, and their loadings. However, in their models the loadings are set to be time-invariant. As the time dimension increases, the loadings are likely to undergo significant structural changes. Ignoring these structural changes would inevitably lead to biased estimates, thus affecting subsequent statistical inference (Hansen, 2001). To the best of our knowledge, although factor models with structural changes have been extensively studied in the context of point-valued data, there is still no related work on models with interval-valued observations.

This paper introduces a high-dimensional factor model with potential structural changes for interval-valued data and proposes the estimation procedure for determining the number of breaks. Under the interval analysis framework adopted in this paper, we verify that when breaks occur, the model can be equivalently represented as an interval-valued pseudo-factor model with stable factor loadings and then the number of factors can be usually overestimated. This finding forms the basis for the proposed method to estimate the number of breaks. Following the sample-splitting strategy of Ma and Su (2018), we divide the sample along the time dimension into several segments,

* Corresponding author at: College of Mathematics and Science, Shanghai Normal University, Shanghai 200234, China.
E-mail address: wujianhong@shnu.edu.cn (J. Wu).

https://doi.org/10.1016/j.econlet.2026.112869
Received 9 July 2025; Received in revised form 26 October 2025; Accepted 12 February 2026
Available online 14 February 2026

resulting in multiple subintervals. Inspired by Wang and Wu (2022) and leveraging the theoretical result that the number of factors tends to be overestimated in the presence of structural changes, we compare the estimated number of interval-valued factors (using the eigenvalue ratio-based estimation method proposed by Guo et al., 2025b) across adjacent subintervals to identify those containing breaks, thereby determining the number of breaks. Under certain conditions, we establish the consistency of the estimator of the number of breaks. All estimation procedures are evaluated through Monte Carlo simulations, and the results demonstrate that the proposed estimator performs well in finite samples. Furthermore, we provide a real data analysis to demonstrate the practical usefulness of the proposed method.

The remainder of this paper is organized as follows. Section 2 introduces the high-dimensional interval-valued factor model with potential structural changes. Section 3 presents the estimation procedure for the number of breaks, along with the required assumptions and theoretical result. Section 4 evaluates the finite sample performance of the estimation procedures through simulation experiments. Section 5 provides an empirical application to illustrate the practical usefulness of the proposed method. Section 6 concludes with a summary of the main contributions and a discussion of possible extensions. Detailed proofs are provided in Appendix. For the sake of statements, we introduce the following notations. For a column/row vector a, $diag(a)$ represents a diagonal matrix, with a being the vector of the diagonal elements. For an interval X, we use X^L and X^R to denote the left and right bounds of X, respectively. The integer part of real number z is denoted as $\lfloor z \rfloor$. The transpose of matrix D is denoted as D'. For any set A, $\#A$ denotes the count function of A. The notations ι_m and I_m denote the m-dimensional vector with element one and $m \times m$ identity matrix, respectively.

2. Models

Consider the high-dimensional interval-valued data factor model with $\mathcal{L}$ ($\mathcal{L} \geq 0$, unknown number) structural changes,

$$y_{it} = \lambda'_{l,i} f_t + e_{it}, \quad i = 1, 2, \ldots, N, \quad h_l < t \leq h_{l+1}, \quad l = 0, 1, \ldots, \mathcal{L}. \tag{2.1}$$

Some notations are as follows, $h_0 \equiv 0$, and $h_{\mathcal{L}+1} \equiv T$. The interval $y_{it} = [y_{it}^L, y_{it}^R]$ represents the observation for individual i at time t. The vector $f_t = (f_{t1}, f_{t2}, \ldots, f_{tr})'$ is an $r \times 1$ unobservable interval-valued factor vector, where each element is an interval-valued factor $f_{tj} = [f_{tj}^L, f_{tj}^R], j = 1, 2, \ldots, r$. The $r \times 1$ point-valued vector $\lambda_{l,i} = (\lambda_{l,i1}, \lambda_{l,i2}, \ldots, \lambda_{l,ir})'$ is the factor loading for individual i during the time period $(h_l, h_{l+1}]$, where $\lambda_{l,ij}$ denotes the point-valued loading corresponding to the jth interval-valued factor for individual i in the time period $(h_l, h_{l+1}]$. The interval-valued error term is denoted as $e_{it} = [e_{it}^L, e_{it}^R]$. In this paper, we treat an interval as a set of ordered numbers (Han et al., 2012). Therefore, the interval allows the left bound to be larger than the right bound, which is the so called extended interval (see Kaucher, 1980; Han et al., 2012, for more details). The operational rules involved in model (2.1) are adopted from Han et al. (2012) and Sun et al. (2018). For intervals X_1 and X_2, (i) Addition: $X_1 + X_2 = [X_1^L + X_2^L, X_1^R + X_2^R]$; (ii) Difference (Hukuhara, 1967): $X_1 - X_2 = [X_1^L - X_2^L, X_1^R - X_2^R]$; (iii) Scalar multiplication: $\beta X_1 = [\beta X_1^L, \beta X_1^R], \beta \in \mathbb{R}$. The rules defined for extended intervals ensure that the resulting space is linear, maintaining closure under operations and ensuring mathematical consistency.

The above model can also be expressed in the following interval-valued vector form,

$$y_t = \Lambda_l f_t + e_t, \quad h_l < t \leq h_{l+1}, \quad l = 0, 1, \ldots, \mathcal{L}, \tag{2.2}$$

where y_t is an $N \times 1$ interval vector composed of y_{it}, i.e., $y_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})'$. Similarly, $e_t = (e_{1t}, e_{2t}, \ldots, e_{Nt})'$ is an $N \times 1$ interval error vector, and $\Lambda_l = (\lambda_{l,1}, \lambda_{l,2}, \ldots, \lambda_{l,N})'$ is an $N \times r$ point-valued loading matrix for the interval-valued factor f_t.

The matrix form of model (2.2) can be written as follows,

$$Y_l = F_l \Lambda'_l + E_l, \quad l = 0, 1, \ldots, \mathcal{L}, \tag{2.3}$$

where $Y_l = (y_{h_l+1}, y_{h_l+2}, \ldots, y_{h_{l+1}})'$ is an $(h_{l+1} - h_l) \times N$ interval matrix whose elements are interval-valued variables. Similarly, $E_l = (e_{h_l+1}, e_{h_l+2}, \ldots, e_{h_{l+1}})'$ is an $(h_{l+1} - h_l) \times N$ interval matrix of error term, and $F_l = (f_{h_l+1}, f_{h_l+2}, \ldots, f_{h_{l+1}})'$ is an $(h_{l+1} - h_l) \times r$ interval matrix for factors. The number $\mathcal{L}$ is unknown and needs to be estimated in this paper.

3. Estimation and theoretical results

3.1. Estimation

Before introducing the estimation procedure, we first present the definition of the D_K-distance for interval, as it serves as the foundational tool for estimating the number of breaks. For intervals X_1 and X_2,

$$D_K(X_1, X_2) = \sqrt{\int_{u,v \in S^0} [s_{X_1}(u) - s_{X_2}(u)][s_{X_1}(v) - s_{X_2}(v)] dK(u, v)},$$

where the unit sphere $S^0 = \{u \in R, |u| = 1\} = \{1, -1\}$, $K(u, v)$(hereafter denoted as K) is a symmetric and positive definite kernel function, and the support function of X_1 becomes $s_{X_1}(u) = X_1^R$ if $u = 1$, $s_{X_1}(u) = -X_1^L$ if $u = -1$. Moreover, $\|X_1 - X_2\|_K^2 = \langle s_{X_1 - X_2}, s_{X_1 - X_2} \rangle_K = D_K^2(X_1, X_2)$, where $\|\cdot\|_K$ and $\langle \cdot, \cdot \rangle_K$ denote the norm and inner product with respect to the kernel K, respectively. Moreover, the D_K-distance can effectively capture the rich information within intervals through appropriate kernel selection. Indeed, the kernel K acts as a weight function. Specifically, by choosing a suitable kernel, the distance between any point pairs across two intervals can be derived as a special case, illustrating the flexibility of this framework. For more information on the D_K-distance, please refer to Han et al. (2012).

Inspired by Wang and Wu (2022), we determine the number of breaks by identifying the number of sample subintervals that contain a single structural change. We divide the entire sample along the time dimension J times, resulting in $J+1$ subintervals. The value J is a user-specified positive integer that satisfies condition $T \gg J \gg \mathcal{L}, J > 3/c_0$, where c_0 comes from Assumption 4 in the next subsection. This ensures that at most one structural change exists within any three consecutive subintervals, thereby guaranteeing the effectiveness of our estimation procedure (see e.g., Ma and Su, 2018; Wang and Wu, 2022). We denote the divided subintervals as $S_j = (\nu_j, \nu_{j+1}]$, where $\nu_j = j\lfloor \frac{T}{J+1} \rfloor, j = 0, 1, \ldots, J$ and $\nu_0 = 0, \nu_{J+1} = T$. At the same time, let the union of two adjacent subintervals be a new subinterval, denoted as $S_j^* = S_j \cup S_{j+1} = (\nu_j, \nu_{j+2}], j = 0, 1, \ldots, J-1$. Next, we will determine the number of breaks by comparing the estimated number of factors in the subintervals. Here, we adopt the eigenvalue ratio estimation method based on D_K-distance for interval-valued factors proposed in Guo et al. (2025b) to obtain the estimator of the number of interval-valued factors in each subinterval, e.g.,

$$\hat{r}_{K,j} = \underset{1 \leq i \leq r_{max}}{\operatorname{argmax}} \frac{\tilde{\mu}_{NT_j,i}}{\tilde{\mu}_{NT_j,i+1}}, \quad j = 0, 1, \ldots, J,$$

where the predetermined number r_{max} represents the maximum possible number of factors, $\tilde{\mu}_{NT_j,i}$ is the ith largest eigenvalue of the matrix $\frac{\langle s'_{Y_j}, s_{Y_j} \rangle_K}{NT_j}$, $Y_j = (y_{\nu_j+1}, y_{\nu_j+2}, \ldots, y_{\nu_{j+1}})'$ and $T_j = \nu_{j+1} - \nu_j$. Similarly, let $\hat{r}_{K,j}^*$ be the estimator of the number of factors in the subinterval S_j^*. Next, we proceed with the estimation procedure for the number of breaks.

To clearly present the method, we first consider a simple case where structural changes occur only in the factor loadings. The following four cases will occur: (i) when $\hat{r}_{K,j} > \hat{r}_{K,j+1}$, there is a break in S_j; (ii) when $\hat{r}_{K,j} < \hat{r}_{K,j+1}$, there is a break in S_{j+1}; (iii) when $\hat{r}_{K,j} = \hat{r}_{K,j+1} \neq \hat{r}_{K,j}^*$, there is a break in S_j^*, and the break occurs at the right end of S_j or the left end of S_{j+1}; (iv) when $\hat{r}_{K,j} = \hat{r}_{K,j+1} = \hat{r}_{K,j}^* = \hat{r}_{K,j-1}^*$, there is no

break in S_j. Then the estimator of the number of breaks can be denoted as

$$\hat{\mathcal{L}} = \#\{S_j : \hat{r}_{K,j} > \hat{r}_{K,j+1}, j = 0, 1, \dots, J-1\} + \#\{S_j^* : \hat{r}_{K,j} = \hat{r}_{K,j+1} \neq \hat{r}_{K,j}^*, j = 0, 1, \dots, J-1\},$$

where $\#A$ denotes the count function of set A.

Then, we can consider a more general case where structural changes may occur in the number of factors and/or in the factor loadings. Then the following three cases may occur: (i) when $\hat{r}_{K,j} \neq \hat{r}_{K,j+1}$, there is a break in S_j^*; (ii) when $\hat{r}_{K,j} = \hat{r}_{K,j+1} \neq \hat{r}_{K,j}^*$, there is a break in S_j^*; (iii) when $\hat{r}_{K,j} = \hat{r}_{K,j+1} = \hat{r}_{K,j}^* = \hat{r}_{K,j-1}^*$, there is no break in S_j. Based on cases (i) and (ii), we can identify the subintervals that contain structural changes. Suppose there are n subintervals S_j^* which can be denoted by $S_{J_1}^*, S_{J_2}^*, \dots, S_{J_n}^*$ along the time order, and define the set $\Omega = \{S_{J_1}^*, S_{J_2}^*, \dots, S_{J_n}^*\}$. However, these subintervals may overlap. To address this, we eliminate the overlapping parts: if $S_{J_j}^* \cap S_{J_{j+1}}^* \neq \emptyset$, then both $S_{J_j}^*$ and $S_{J_{j+1}}^*$ are removed from Ω and replaced by the intersection $S_{J_j}^* \cap S_{J_{j+1}}^*$. This adjustment is justified by our partitioning rule, which theoretically ensures that at most one structural change exists among any three consecutive subintervals. Consequently, the estimator of the number of breaks is given by

$$\hat{\mathcal{L}} = \#\Omega.$$

3.2. Theoretical result

For $\eta < \frac{1}{J+1}$, the time sample subinterval $[t_1, t_2]$ satisfies that $1 \leq t_1 < t_2 \leq T, t_2 - t_1 \geq \eta T$. Let r_l represent the true number of factors on the subinterval $(h_l, h_{l+1}]$, c_1, c_2 be positive constants, and $\psi_j(D)$ denote the jth largest eigenvalue of the point-valued matrix D.

Assumption 1. For $j = 1, 2, \dots, r_l$, $p\lim_{m\to\infty} \psi_{j,l}(\frac{\Gamma_l' \Gamma_l \sum_{t=t_1}^{t_2} \langle s_{f_t}, s_{f_t}' \rangle_K}{N(t_2 - t_1)}) = \mu_{j,l}$, where $0 < \mu_{j,l} < \infty$, and the number of factors $r_l, l = 0, 1, \dots, \mathcal{L}$, is finite.

Assumption 2. (i) For all $t = 1, 2, \dots, T, i = 1, 2, \dots, N, l = 0, 1, 2, \dots, \mathcal{L}$, it holds that $\mathbb{E}(\|f_t\|_K^4) \leq c_1$ and $\|\lambda_{l,i}\| \leq c_1$.
(ii) $\mathbb{E}(\frac{\|\sum_{t=t_1}^{t_2} \langle s_{e_t}, s_{f_t}' \rangle_K\|^2}{N(t_2 - t_1)}) < c_1$.
(iii) $\text{rank}(\begin{bmatrix} \Lambda_{l-1} & \Lambda_l \end{bmatrix}) > r_l, l = 1, 2, \dots, \mathcal{L}$.

Assumption 3. For some $d^c \in (0, 1]$, it holds that $\psi_1(\frac{\sum_{t=t_1}^{t_2} \langle s_{e_t}, s_{e_t}' \rangle_K}{M}) = O_p(1)$ and $\psi_{\lfloor d^c m \rfloor}(\frac{\sum_{t=t_1}^{t_2} \langle s_{e_t}, s_{e_t}' \rangle_K}{M}) \geq c_2 + o_p(1)$, where $m = \min\{N, T\}$, $M = \max\{N, T\}$.

Assumption 4. There exists a positive constant c_0 such that $I_{min} \geq c_0 T$, where $I_{min} = \min_{0 \leq l \leq \mathcal{L}}(h_{l+1} - h_l)$.

Most parts of Assumptions 1–3 follow Guo et al. (2025b) and represent the standard conditions typically imposed in approximate factor models. For instance, Assumption 2(ii) requires orthogonality between interval-valued factors and interval-valued errors, while Assumption 3 specifies the dependence structure of the error term by allowing for weak cross-sectional and serial correlations. These settings are consistent with Ahn and Horenstein (2013) and constitute the standard assumptions in factor models, thereby ensuring the consistency of the eigenvalue ratio-based estimators constructed under the D_K-distance framework. Assumption 4 further guarantees that each regime contains a sufficiently large number of observations, in line with the requirements commonly adopted in the structural break literature (see Bai and Perron, 1998; Wang and Wu, 2022).

Theorem 1. *Suppose that Assumptions 1–4 hold, we then have*

$$\lim_{N,T\to\infty} P(\hat{\mathcal{L}} = \mathcal{L}) = 1.$$

Remark 1. Note that we allow the number of breaks $\mathcal{L} = 0$. That is, we can conclude that there is no structural changes happening on the considered time range if $\hat{\mathcal{L}} = 0$.

4. Monte Carlo simulation study

In this section, we demonstrate the finite sample properties of the proposed estimator through some Monte Carlo simulation experiments. We consider the data generating processes (DGPs) and parameter settings similar to those in Wang and Wu (2022). Specifically, for any $s \in \{L, R\}$,

$$y_{it}^s = \lambda_{j,i}' f_t^s + e_{it}^s, \quad h_j < t \leq h_{j+1}, \quad j = 0, 1, 2, \ i = 1, 2, \dots, N,$$

where the left and right bounds of the factors are generated as follows,

$$f_t^s = diag(0.6, 0.3) f_{t-1}^s + u_t^s,$$

with $f_t^s = (f_{1t}^s, f_{2t}^s)'$, $u_t^s = (u_{t,1}^s, u_{t,2}^s)' \overset{i.i.d.}{\sim} N(0_{2\times1}, I_2)$, $f_1^s = (f_{11}^s, f_{21}^s)' \overset{i.i.d.}{\sim} N(0_{2\times1}, diag((1-0.6^2)^{-1}, (1-0.3^2)^{-1}))$. Moreover, we set $\text{Cov}(f_{jt}^L, f_{k\tau}^R) = 0.4$ when $j = k$ and $t = \tau$, and 0 otherwise. The generation process for the left and right bounds of the error terms is as follows,

$$e_t^s = \rho e_{t-1}^s + v_t^s,$$

where $e_t^s = (e_{1t}^s, e_{2t}^s, \dots, e_{Nt}^s)'$, $e_1^s \sim N(0_{N\times1}, \frac{1}{1-\rho^2}\Omega)$, and $v_t^s = (v_{t,1}^s, v_{t,2}^s, \dots, v_{t,N}^s)' \overset{i.i.d.}{\sim} N(0_{N\times1}, \Omega)$. Here, the $N \times N$ matrix Ω characterizes the cross-sectional dependence of the disturbance terms at time t, with the (i, j)th element given by $\Omega_{ij} = \alpha^{|i-j|}$. In the simulations presented in this section, we consider the following six settings for the error terms: (C_1) independent and identically distributed errors ($\rho = 0, \alpha = 0$); (C_2) cross-sectional heterogeneity errors ($\rho = 0, \Omega = diag(\omega_1, \omega_2, \dots, \omega_N)$, $\omega_i \overset{i.i.d.}{\sim} U(0.5, 1.5)$, $i = 1, 2, \dots, N$); (C_3) serially correlated errors ($\rho = 0.5, \alpha = 0$); (C_4) weakly cross-sectionally correlated errors ($\rho = 0, \alpha = 0.4$); (C_5) weakly cross-sectionally correlated errors ($\rho = 0, \alpha = 0.6$); (C_6) both serially and cross-sectionally correlated errors ($\rho = 0.2, \alpha = 0.4$).

For the generation of factor loadings, we consider the following settings:

DGP 1: (No structural change) $\lambda_{0,i} = \lambda_{1,i} = \lambda_{2,i} \overset{i.i.d.}{\sim} N(b\iota_2, I_2)$;

DGP 2: (Single structural change) $\lambda_{0,i} \overset{i.i.d.}{\sim} N(0.5b\iota_2, I_2)$, $\lambda_{1,i} = \lambda_{2,i} \overset{i.i.d.}{\sim} N(b\iota_2, I_2)$, $\lambda_{0,i}$ are independent of $\lambda_{1,i}$, and we set the break date $t = 0.5T$;

DGP 3(i): (Multiple structural changes) $\lambda_{0,i} \overset{i.i.d.}{\sim} N(0.5b\iota_2, I_2)$, $\lambda_{1,i} \overset{i.i.d.}{\sim} N(b\iota_2, I_2)$, $\lambda_{2,i} \overset{i.i.d.}{\sim} N(1.5b\iota_2, I_2)$, $\lambda_{0,i}$, $\lambda_{1,i}$ and $\lambda_{2,i}$ are independent of each other. We set the two break dates as $t = 0.3T$ and $0.6T + \lfloor 0.3T/(J+1) \rfloor$;

DGP 3(ii): (Multiple structural changes) $\lambda_{0,i} \overset{i.i.d.}{\sim} N(0.5b\iota_2, I_2)$, $\lambda_{1,i} \overset{i.i.d.}{\sim} N(b\iota_2, I_2)$, and the first element of $\lambda_{2,i}$ is the same as that of $\lambda_{1,i}$, while the second element is zero, indicating that the number of factors decreases from 2 to 1. Similarly, $\lambda_{0,i}$ is dependent of $\lambda_{1,i}$. The breaks occur at $t = 0.3T$ and $0.7T$.

The simulation is conducted with a cross-sectional size of $N = 100$ and time dimensions $T = 200, 300, 400$. For $T = 200$ and $T = 300$, J is set to be 9, resulting in $J + 1 = 10$ subintervals along the time dimension, each containing approximately 10% of the observations. For $T = 400$, J is set to be 14, resulting in $J + 1 = 15$ subintervals along the time dimension, each containing approximately 6.7% of the observations. The choice of J follows the principle that each subinterval should contain 5%–25% of the information in the sample; see Ma and Su (2018) and Bai and Perron (1998) for details. In this simulation, the number of factors is estimated using the eigenvalue ratio method based

Table 1
Percentages of the correct estimation of the number of breaks.

(T, J)	$(200, 9)$			$(300, 9)$			$(400, 14)$		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
$b = 0$									
DGP1	99.5	98.8	98.5	100.0	100.0	100.0	100.0	99.8	99.8
DGP2	99.8	99.6	99.1	100.0	100.0	100.0	100.0	99.8	99.7
DGP3-(i)	99.2	99.5	99.7	100.0	100.0	100.0	99.9	99.9	99.9
DGP3-(ii)	99.8	99.7	99.7	100.0	100.0	100.0	99.8	100.0	99.9
	C_4	C_5	C_6	C_4	C_5	C_6	C_4	C_5	C_6
$b = 0$									
DGP1	98.9	96.7	97.6	99.8	99.6	100.0	99.8	99.3	99.0
DGP2	99.0	98.0	98.6	100.0	99.8	99.9	99.9	99.1	99.6
DGP3-(i)	99.2	97.9	99.5	99.8	99.9	99.9	99.7	99.4	99.6
DGP3-(ii)	99.8	99.7	99.8	100.0	100.0	100.0	99.9	98.9	99.9
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
$b = 1$									
DGP1	92.7	90.1	89.3	99.5	96.7	99.0	97.4	97.0	95.7
DGP2	97.3	95.8	92.9	99.7	99.8	99.8	98.8	98.2	99.0
DGP3-(i)	81.9	79.3	74.3	96.8	97.4	94.6	90.6	84.9	84.9
DGP3-(ii)	97.2	97.4	96.6	100.0	100.0	99.6	99.5	99.0	98.5
	C_4	C_5	C_6	C_4	C_5	C_6	C_4	C_5	C_6
$b = 1$									
DGP1	87.4	75.3	81.7	97.6	94.8	96.4	94.3	86.0	91.7
DGP2	93.9	88.3	92.2	99.1	97.7	99.5	97.8	93.9	95.9
DGP3-(i)	71.0	61.2	69.9	90.8	83.8	86.2	80.0	65.6	78.1
DGP3-(ii)	96.5	95.4	94.2	99.9	98.4	99.3	97.7	93.6	96.3

on the D_K-distance proposed by Guo et al. (2025b). All simulation results are based on 1000 replications.

Table 1 shows that as the sample size increases, the overall accuracy of the estimates tends to improve. This confirms the theoretical result of Theorem 1. However, in the case of $(T, J) = (400, 14)$, the accuracy is slightly lower than that in $(T, J) = (300, 9)$. This difference may be attributed to the fact that each subinterval in the former case contains only about 6.7% of the observations, whereas the latter's subintervals contain 10%. In addition, the estimation procedure exhibits robustness to the correlation structure of the disturbances. Regardless of whether the disturbances have cross-sectional dependence, serial correlation, or cross-sectional heterogeneity, the estimator performs well in finite samples. Overall, the estimator demonstrates desired finite sample performance.

5. Application

To illustrate the practical relevance of the proposed method, we apply it to stock return data from the S&P 100. We construct a balanced panel consisting of 74 stocks with complete observations from January 3, 2008 to December 28, 2008. For each trading day, we construct interval-valued returns following Fischer et al. (2016). Two definitions are considered: the return-interval-from-closing-price (RICP), $R_t^{RICP} = [\ln(\frac{L_t}{C_{t-1}}), \ln(\frac{H_t}{C_{t-1}})]$, and the return-interval-from-price-interval (RIPI), $R_t^{RIPI} = [\ln(\frac{L_t}{H_{t-1}}), \ln(\frac{H_t}{L_{t-1}})]$, where L_t and H_t respectively denote the low and high prices on day t, and C_{t-1} is the closing price on day $t-1$. In addition, we compute conventional point-valued log returns based on closing prices, i.e., $\ln(\frac{C_t}{C_{t-1}})$. After constructing these returns, we obtain $T = 248$ daily observations for each stock. As discussed in Sun et al. (2020), interval-valued returns contain richer information on price fluctuations and have been shown to improve volatility modeling and prediction. This motivates our focus on interval-valued data in structural break analysis.

Similar to arguments in the simulation study, we select the tuning parameter $J = 9$ in the estimation procedure due to the sample size $T = 248$. Applying the proposed estimation method for both the RIPI- and RICP-based interval-valued data, we can detect one structural break. This finding is consistent with the result reported by Bai et al. (2020), who detected a single break when analyzing monthly return data for stocks traded on the NYSE, AMEX, and NASDAQ between January 2005 and December 2012. In contrast, when using point-valued returns based on closing prices, the result shows that three structural changes occur in the consider time range, which appear less economically plausible.

Overall, this application demonstrates two key points. First, the proposed estimator effectively detects economically meaningful structural changes in high-dimensional financial data. Second, the interval-based estimates are stable and align well with major economic events, confirming the practical usefulness of the proposed method.

6. Conclusion and discussion

This paper develops a high-dimensional factor model with potential structural changes for interval-valued data and proposes an estimator for the number of breaks. Building on the finding that the number of factors is usually overestimated in the presence of structural changes, we construct a consistent estimator by employing a sample-splitting strategy and comparing the estimated number of factors across subintervals. Monte Carlo simulation results show that the proposed procedure performs well in finite samples, and also remains valid in the case with no breaks. In addition, we conduct an empirical study to further demonstrate the practical usefulness of our method.

Beyond these findings, there are several interesting directions for future research. First, as the referee pointed out, although the proposed method is available for the case with no breaks (i.e., $\mathcal{L} = 0$), the theoretical result in Theorem 1 is still far from being designed as a test for the existence of breaks which is one of important direction in high-dimensional factor analysis. Second, while this paper focuses on estimating the number of breaks, it would also be of great interest to give the estimation of break locations and the corresponding theoretical properties in the proposed interval-valued factor model, following approaches developed for point-valued factor models. These issues are left for future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are deeply grateful to Professor Brendan Kline and an anonymous referee for valuable comments that led to substantial improvement of this paper. Wu gratefully acknowledges the financial support from the National Natural Science Foundation of China (Grant No. 72173086). All errors are the authors' sole responsibilities.

Appendix. Technical details

Denote $T_j = v_{j+1} - v_j$, $T_{j,1} = h_l - v_j$ and $T_{j,2} = v_{j+1} - h_l$. Let $\mathbb{S}_1$ denote the set of subintervals without any breaks. Let $\mathbb{S}_{21}$ denote the set of subintervals that contain one break and satisfy $\lim_{T\to\infty} \frac{T_{j,1}}{T_j} \in (0, 1)$. Let $\mathbb{S}_{22}$ denote the set of subintervals that contain one break and satisfy $\lim_{T\to\infty} \frac{T_{j,1}}{T_j} = 0$. Let $\mathbb{S}_{23}$ denote the set of subintervals that contain one break and satisfy $\lim_{T\to\infty} \frac{T_{j,1}}{T_j} = 1$. Finally, define $\mathbb{S}_2 = \mathbb{S}_{21} \cup \mathbb{S}_{22} \cup \mathbb{S}_{23}$. Thus, $\mathbb{S}_2$ represents the set of subintervals that contain a break.

Lemma 1. *Suppose that Assumptions 1–4 hold, we then have* $\lim_{N,T\to\infty} P(\hat{r}_{K,j} = r + q_j) = 1$, *where* $q_j = 0$ *if* $S_j \in \mathbb{S}_1$, $q_j > 0$ *if* $S_j \in \mathbb{S}_{21}$, *and* $q_j \geq 0$ *if* $S_j \in \mathbb{S}_{22} \cup \mathbb{S}_{23}$.

Proof. We consider the following four cases: (i) $S_j \in \mathbb{S}_1$, (ii) $S_j \in \mathbb{S}_{21}$, (iii) $S_j \in \mathbb{S}_{22}$, and (iv) $S_j \in \mathbb{S}_{23}$. Note that this discussion focuses on the case where only the factor loadings change; thus, the true number of factors within each subinterval $(h_l, h_{l+1}]$ is the same, i.e., $r_l = r, l = 0, 1, \ldots, \mathcal{L}$.

When $S_j \in \mathbb{S}_1$, there are no breaks within S_j. Therefore, there exists an l such that $S_j \subseteq (h_l, h_{l+1}]$. This implies that the factor model has stable factor loadings over the period S_j. Moreover, it is easy to verify from the assumptions in this paper that the model satisfies the assumptions for the interval-valued factor model in Guo et al. (2025b). Thus, according to Theorem 1, $\hat{r}_{K,j}$ is a consistent estimator of r.

When $S_j \in \mathbb{S}_{21}$, there is a break within S_j, and $\lim_{T\to\infty} \frac{T_{j,1}}{T_j} \in (0,1)$. In this case, there exists an $l \geq 1$, such that $S_j \subseteq (h_{l-1}, h_{l+1}]$ and $h_l \in S_j$. Therefore, the model in subinterval S_j can be represented in the following matrix form,

$$Y_j = \begin{bmatrix} F_{j,1}\Lambda_{l-1}' \\ F_{j,2}\Lambda_l' \end{bmatrix} + E_j = \begin{bmatrix} F_{j,1} & 0 \\ 0 & F_{j,2} \end{bmatrix} \begin{bmatrix} \Lambda_{l-1}' & \Lambda_l' \end{bmatrix} + E_j, \tag{A.1}$$

where $Y_j = (y_{v_j+1}, y_{v_j+2}, \ldots, y_{v_{j+1}})'$, $E_j = (e_{v_j+1}, e_{v_j+2}, \ldots, e_{v_{j+1}})'$, and $F_{j,1} = (f_{v_j+1}, f_{v_j+2}, \ldots, f_{h_l})'$, $F_{j,2} = (f_{h_l+1}, f_{h_l+2}, \ldots, f_{v_{j+1}})'$, Λ_{l-1} and Λ_l denote the true loading matrices before and after the lth structural change, respectively. Let Γ_l be the matrix consisting of the maximal linearly independent set of columns of $\begin{bmatrix}\Lambda_{l-1} & \Lambda_l\end{bmatrix}$, and note that $\text{rank}(\begin{bmatrix}\Lambda_{l-1} & \Lambda_l\end{bmatrix}) = r + p_l > r$. There exist $(r+p_l)\times r$ matrices A_l and B_l such that model (2.3) can be expressed in the following form,

$$Y_j = G_j\Gamma_l' + E_j, \tag{A.2}$$

where the stable factor loadings $G_j = \begin{bmatrix} F_{j,1}A_l' \\ F_{j,2}B_l' \end{bmatrix} = (g_{v_j+1}, g_{v_j+2}, \ldots, g_{v_{j+2}})'$, $g_t = A_l f_t$, $v_j < t \leq h_l, g_t = B_l f_t, h_l < t \leq v_{j+1}$. In other words, the model can be expressed as an interval-valued pseudo-factor model (A.2) with $r + p_l$ stable factor loadings. Next, we will show that the equivalent model (A.2) satisfies all the assumptions for the factor model established in Guo et al. (2025b). This will establish that the factor number estimator introduced in Guo et al. (2025b) is consistent for this equivalent model.

By Assumption 1, it follows that $\langle s'_{F_{j,1}}, s_{F_{j,1}}\rangle_K$ and $\langle s'_{F_{j,2}}, s_{F_{j,2}}\rangle_K$ are positive definite. Therefore, by the definitions of A_l, B_l, we have $A_l\langle s'_{F_{j,1}}, s_{F_{j,1}}\rangle_K A_l' > 0$ and $B_l\langle s'_{F_{j,2}}, s_{F_{j,2}}\rangle_K B_l' > 0$. Moreover, $\zeta' A_l\langle s'_{F_{j,1}}, s_{F_{j,1}}\rangle_K A_l'\zeta + \zeta' B_l\langle s'_{F_{j,2}}, s_{F_{j,2}}\rangle_K B_l'\zeta = 0$ if and only if $\zeta = 0$. In other words, $\langle s'_{G_j}, s_{G_j}\rangle_K = A_l\langle s'_{F_{j,1}}, s_{F_{j,1}}\rangle_K A_l' + B_l\langle s'_{F_{j,2}}, s_{F_{j,2}}\rangle_K B_l'$ is positive definite and $\text{rank}(\langle s'_{G_j}, s_{G_j}\rangle_K) = r + p_l$.

Because $\text{rank}(\Gamma_l'\Gamma_l\langle s'_{G_j}, s_{G_j}\rangle_K) \geq \text{rank}(\Gamma_l'\Gamma_l) + \text{rank}(\langle s'_{G_j}, s_{G_j}\rangle_K) - (r + p_l) = r + p_l$, it follows that $\text{rank}(\frac{\Gamma_l'\Gamma_l\langle s'_{G_j}, s_{G_j}\rangle_K}{NT_j}) = r + p_l$. Moreover, since the two matrices $\frac{\Gamma_l'\Gamma_l\langle s'_{G_j}, s_{G_j}\rangle_K}{NT_j}$ and $\frac{\Gamma_l\langle s'_{G_j}, s_{G_j}\rangle_K\Gamma_l'}{NT_j}$ share the same non-zero eigenvalues, and the latter's eigenvalues are all non-negative, this confirms that Assumption 1 in Guo et al. (2025b) is satisfied.

Let $\Gamma_{l,i}$ denote the ith column of Γ_l. From Assumption 2, we have $\|\Gamma_{l,i}\| \leq \sqrt{\|\lambda_{l,i}\|^2 + \|\lambda_{l-1,i}\|^2} \leq 2c_1$. Note that $\sum_{t=v_j+1}^{v_{j+1}}\langle s_{e_t}, s'_{g_t}\rangle_K = \sum_{t=v_j+1}^{h_l}\langle s_{e_t}, s'_{f_t}\rangle_K A_l' + \sum_{t=h_l+1}^{v_{j+1}}\langle s_{e_t}, s'_{f_t}\rangle_K B_l'$. Therefore, by Assumption 2, we have $\mathbb{E}(\frac{\|\sum_{t=v_j}^{v_{j+1}}\langle s_{e_t}, s'_{g_t}\rangle_K\|^2}{NT_j}) < c_1$. As a result, Assumption 2 in Guo et al. (2025b) is verified. Meanwhile, the assumptions about the disturbances in this section directly verify Assumption 3 in Guo et al. (2025b). Therefore, when $S_j \in \mathbb{S}_{21}$, we have $\lim_{N,T\to\infty} P(\hat{r}_{K,j} = r + q_j) = 1$ with $q_j > 0$.

When $S_j \in \mathbb{S}_{22}$, from the preceding proof it can be seen that $\langle s'_{G_j}, s_{G_j}\rangle_K$ is not necessarily full rank. Therefore, the above proof method is no longer applicable. To address this, we rewrite the model as the following equivalent form,

$$Y_j = F_j\Lambda_l' + \Delta_j + E_j,$$

where $\Delta_j = ((\Lambda_l - \Lambda_{l-1})F_{v_j+1}, \ldots, (\Lambda_l - \Lambda_{l-1})F_{h_l}, 0, \ldots, 0)'$. The change in factor loadings can be treated as an additional error term, meaning that the main properties of the model are determined by the subinterval $(h_l, v_{j+1}]$. By Lemma 3(i), we have $\lim_{N,T\to\infty} P(\hat{r}_{K,j} \geq r) = 1$. Therefore, when $S_j \in \mathbb{S}_{22}$, $\lim_{N,T\to\infty} P(\hat{r}_{K,j} = r + q_j) = 1$ with $q_j \geq 0$. □

Lemma 2. *Suppose that Assumptions 1–4 hold, we then have* $\lim_{N,T\to\infty} P(\hat{r}^*_{K,j} = r + q^*_j) = 1$, *where* $q^*_j = 0$ *if* $S_j, S_{j+1} \in \mathbb{S}_1$, $q^*_j > 0$ *if* $S_j \in \mathbb{S}_{21} \cup \mathbb{S}_{23}$ *or* $S_{j+1} \in \mathbb{S}_{21} \cup \mathbb{S}_{22}$, *and* $q^*_j \geq 0$ *if* $S_j \in \mathbb{S}_1, S_{j+1} \in \mathbb{S}_{23}$ *or* $S_j \in \mathbb{S}_{22}, S_{j+1} \in \mathbb{S}_1$.

Proof. The proof is similar to Lemma 1 and thus omitted. □

Lemma 3. *Suppose that Assumptions 1–4 hold, we have (i) if* $S_j \in \mathbb{S}_{22} \cup \mathbb{S}_{23}$, *then* $\lim_{N,T\to\infty} P(\hat{r}_{K,j} \geq r) = 1$. *(ii) if* $S_j \in \mathbb{S}_{22}$, $\frac{T_{j,1}}{T_j} = O_P(\eta^{-2}_{N,T_j})$ *with* $\eta_{N,T_j} = \min\{\sqrt{N}, \sqrt{T_j}\}$, *then* $\lim_{N,T\to\infty} P(\hat{r}_{K,j} = r) = 1$. *(iii) if* $S_j \in \mathbb{S}_{23}$, $\frac{T_{j,2}}{T_j} = O_P(\eta^{-2}_{N,T_j})$ *with* $\eta_{N,T_j} = \min\{\sqrt{N}, \sqrt{T_j}\}$, *then* $\lim_{N,T\to\infty} P(\hat{r}_{K,j} = r) = 1$.

Proof. When $S_j \in \mathbb{S}_{22} \cup \mathbb{S}_{23}$ by the proof of Lemma 1, the model can be decomposed into a factor structure and an additional error term. The eigenvalues associated with the additional error term are of order $O_p(\frac{T_{j,1}}{T_j})$ or $O_p(\frac{T_{j,2}}{T_j})$. In the absence of this additional error term, the eigenvalues dominated by the factor structure are of order $O_p(1)$, while those dominated by the error component are of order $O_p(\eta^{-2}_{N,T_j})$. Therefore, the asymptotic properties of the factor number estimation depend on the convergence rate of $\frac{T_{j,1}}{T_j}$ and $\frac{T_{j,2}}{T_j}$. When these convergence rates are unknown, we have $\lim_{N,T\to\infty} P(\hat{r}_{K,j} \geq r) = 1$. That is, when $S_j \in \mathbb{S}_{22}$ and $\frac{T_{j,1}}{T_j} = O_P(\eta^{-2}_{N,T_j})$ or when $S_j \in \mathbb{S}_{23}$ and $\frac{T_{j,2}}{T_j} = O_P(\eta^{-2}_{N,T_j})$, the additional error term does not affect the essential nature of the model. Consequently, we have $\lim_{N,T\to\infty} P(\hat{r}_{K,j} = r) = 1$. □

Proof of Theorem 1. Lemmas 1 and 2 respectively establish the consistency of the eigenvalue ratio estimator based on the D_K-distance from Guo et al. (2025b) for different scenarios of subintervals. These two lemmas together lead to Theorem 1. The proof is similar to that of Wang and Wu (2022) and is therefore omitted here to save space. □

Data availability

The dataset used in this study is available from the authors upon reasonable request.

References

Ahn, S.C., Horenstein, A.R., 2013. Eigenvalue ratio test for the number of factors. Econometrica 81 (3), 1203–1227.

Bai, J., Han, X., Shi, Y., 2020. Estimation and inference of change points in high-dimensional factor models. J. Econometrics 219 (1), 66–100.

Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66 (1), 47–78.

Billard, L., Diday, E., 2000. Regression analysis for interval-valued data. In: Data Analysis, Classification, and Related Methods. Springer Berlin Heidelberg, pp. 369–374.

Billard, L., Diday, E., 2002. Symbolic regression analysis. In: Classification, Clustering, and Data Analysis. Springer Berlin Heidelberg, pp. 281–288.

Billard, L., Diday, E., 2003. From the statistics of data to the statistics of knowledge: symbolic data analysis. J. Amer. Statist. Assoc. 98 (462), 470–487.

De Carvalho, F.d.A., Lima Neto, E.d.A., Tenorio, C.P., 2004. A new method to fit a linear regression model for interval-valued data. In: Annual Conference on Artificial Intelligence. Springer, pp. 295–306.

Fischer, H., Blanco-Fernández, Á., Winker, P., 2016. Predicting stock return volatility: Can we benefit from regression models for return intervals? J. Forecast. 35 (2), 113–146.

González-Rivera, G., Lin, W., 2013. Constrained regression for interval-valued data. J. Bus. Econom. Statist. 31 (4), 473–490.

Guo, Y., Zou, G., Wu, J., 2025a. Factor modeling for high-dimensional interval-valued data. Stud. Nonlinear Dyn. Econ. Published online. https://doi.org/10.1515/snde-2024-0019.

Guo, Y., Zou, G., Wu, J., 2025b. Factor modeling for high-dimensional interval-valued data: determining the number of factors. J. Syst. Sci. Complex. Published online. https://doi.org/10.1007/s11424-025-4442-7.

Han, A., Hong, Y., Wang, S., 2012. Autoregressive conditional models for interval-valued time series data. Manuscript, Department of Economics, Cornell University.

Han, A., Hong, Y., Wang, S., Yun, X., 2016. A vector autoregressive moving average model for interval-valued time series data. In: Essays in Honor of Aman Ullah. Emerald Group Publishing Limited, pp. 417-460.

Hansen, B.E., 2001. The new econometrics of structural change: Dating breaks in US labor productivity. J. Econ. Perspect. 15 (4), 117–128.

Hukuhara, M., 1967. Integration des applications mesurables dont la valeur est un compact convexe. Funkcial. Ekvac. 10 (3), 205–223.

Kaucher, E., 1980. Interval analysis in the extended interval space IR. Computing Suppl 2, 33–49.

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

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

Ma, S., Su, L., 2018. Estimation of large dimensional factor models with an unknown number of breaks. J. Econometrics 207 (1), 1–29.

Sun, Y., 2017. Asymptotic tests for interval-valued means. Statist. Probab. Lett. 121, 70–77.

Sun, Y., Han, A., Hong, Y., Wang, S., 2018. Threshold autoregressive models for interval-valued time series data. J. Econometrics 206 (2), 414–446.

Sun, Y., Lian, G., Lu, Z., Loveland, J., Blackhurst, I., 2020. Modeling the variance of return intervals toward volatility prediction. J. Time Series Anal. 41 (4), 492–519.

Wang, L., Wu, J., 2022. Estimation of high-dimensional factor models with multiple structural changes. Econ. Model. 108, 105743.