Statistics and Computing (2025) 35:207
https://doi.org/10.1007/s11222-025-10743-2

ORIGINAL PAPER

Hausdorff correlation for interval-valued random objects

Xinlai Kang[1] · **Xiaxue Ouyang**[1] · **Haoxian Liang**[2] · **Cheng Meng**[3]

Received: 15 April 2025 / Accepted: 26 September 2025

Abstract
Analyzing the correlation between interval-valued data presents an essential yet challenging problem in modern statistical research due to the lack of basic geometric and algebraic structures. Existing methods are often limited by their reliance on algebraic formulations or assumptions about the underlying distribution of true values within intervals. Moreover, they primarily focus on simple midpoint-range interval representations, restricting their applicability to more complex interval structures, e.g., when the interval contains multiple segments. To address these limitations, we introduce the Fréchet framework into the interval metric space equipped with the Hausdorff distance, extending the notions of Fréchet mean and proposing a more general and straightforward interval dependency measure, called Hausdorff correlation. The proposed method offers a strong geometric interpretation, revealing the relationship between random intervals and their Hausdorff mean, while also accommodating a broader range of interval forms. From a theoretical perspective, we establish the foundational properties of the proposed framework, proving the existence and uniqueness of the Hausdorff mean. Empirical evaluations on both synthetic and real-world datasets demonstrate the distinctiveness and effectiveness of Hausdorff correlation and its superior performance in feature selection compared to existing methods. In particular, a real-world Wearable Watch Dataset analysis shows the Hausdorff correlation successfully captures the relationship between multi-segment sleep intervals and physiological indicators, where existing methods fail to provide meaningful estimates.

Keywords Hausdorff distance · Metric space · Interval-valued data · Multi-segment intervals

1 Introduction

Due to privacy restrictions and the increasing complexity of large datasets, accessing individual observations can be challenging, except for interval-valued data. Examples include price fluctuations over a given period of time, daily variations in systolic and diastolic blood pressure of a patient, personal information such as income and age, and aggregated transaction summaries per card, which are often recorded as intervals rather than precise values (Billard and Diday 2003; Gil et al. 2007; Sinova et al. 2012).

Analyzing such interval-valued data is a crucial, yet challenging problem in modern statistical research. Billard and Le-Rademacher (2012) and D'Urso and Giordani (2004) provided a detailed description and illustration of the principal component methodology for interval data, while El Ghaoui et al. (2003) addressed a binary linear classification problem using an interval matrix uncertainty model. Numerous studies have explored interval regression from various perspectives and within different frameworks. For example, Diamond (1990) performed least squares regression on the interval data based on interval arithmetic, while the other way is by characterizing the intervals by their midpoint-range tandem (de Carvalho et al. 2004; Neto and de Carvalho 2008, 2010). When dealing with interval data for statistical purposes, intervals are treated as random elements over a probability space with the expected value in Kudō-Aumann's sense, as explored by Aumann (1965) and Gil et al. (2007). Furthermore, Billard (2006) and Billard (2008) considered the interval-valued data as a form of symbolic data. However, existing methods are often constrained by algebraic definitions or rely on assumptions regarding the distribution of true values within intervals. They presuppose a specific data distribution or focus solely on compact convex intervals on

✉ Cheng Meng
chengmeng@ruc.edu.cn

1 Institute of Statistics and Big Data, Renmin University of China, Beijing, China

2 School of Mathematical Sciences, Beijing Normal University, Beijing, China

3 Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Published online: 05 October 2025

the real line, typically represented using the midpoint-range representation.

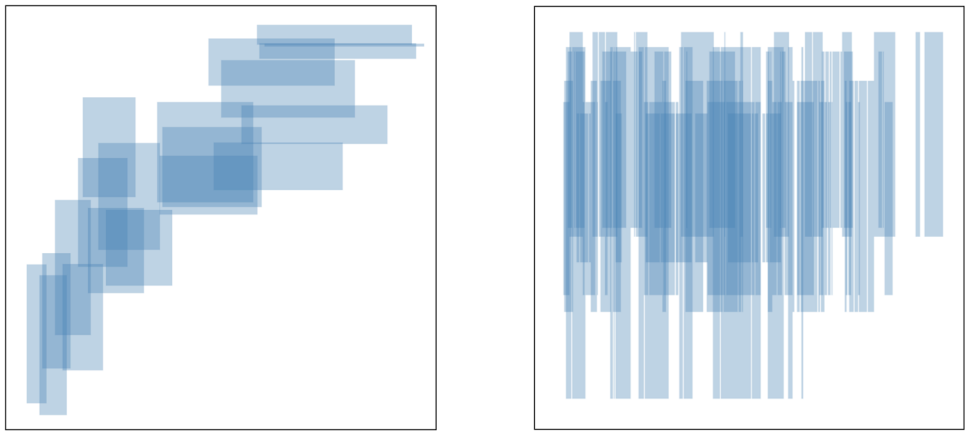
Thus, we aim to develop a more general and straightforward framework for analyzing interval data. Recently, Petersen and Müller (2019) considered modeling distributional data as random objects in the Wasserstein metric space, which has been extensively studied for its theoretical foundation and interpretability related to the optimal transport problem (Meng et al. 2019, 2020; Zhang et al. 2021; Li et al. 2023a; Zhang et al. 2023a, b; Li et al. 2024; Hu et al. 2025). In the space of one-dimensional probability densities equipped with the Wasserstein metric, the authors extended the notion of the Fréchet mean to define the Wasserstein-Fréchet mean (also referred to as the Wasserstein barycenter) in metric space. This extension led to the definition of a Wasserstein-based covariance measure, which admits a natural interpretation as the expected inner product of Centralized optimal transport maps in the Hilbert space L^2.

Motivated by this line of research, we focus on modeling interval-valued data as random objects in a metric space. In particular, we adopt the Hausdorff distance, a classical metric that quantifies the discrepancy between two subsets of a metric space. The intervals we consider are more general than those in previous studies, as they are only required to be nonempty and compact, whereas prior work additionally assumed convexity in a real line. We fully utilize the structure of the interval metric space endowed with the Hausdorff distance to explore the intrinsic nature of intervals, rather than characterizing them solely through their midpoint and length.

Our key contributions can be summarized as follows. First, by treating intervals as random objects, we extend the concepts of Fréchet mean and variance to an interval metric space equipped with the Hausdorff distance, redefining the corresponding mean and variance of intervals. These definitions offer an intuitive geometric interpretation as a natural generalization of the standard mean and variance. Second, we propose a novel covariance measure, termed Hausdorff covariance, to quantify the relationship between two random intervals, independent of interval type or data distribution. Analogous to the notion of a regular covariance for an appropriate inner product, Hausdorff covariance can be interpreted as measuring the degree of synchronization in deviations from their Fréchet means. Finally, we establish the basic theoretical foundation for our definitions and demonstrate the superior performance of the proposed Hausdorff correlation across various synthetic and real-world datasets. Unlike existing approaches, which are restricted to single-segment interval data, our method extends naturally to multi-segment interval data, thereby providing a broader and more general analytical framework.

Figure 1 highlights the key characteristics of our proposed correlation measure. As shown in Figure 1(a), The midpoints of the intervals exhibit a strong positive correlation (0.88), whereas the interval lengths demonstrate a negative correlation (-0.95), based on the two interval sequences $\{x_i\}_{i=1}^{100}$ and $\{y_i\}_{i=1}^{100}$. The midpoint pairs (x_i^c, y_i^c) and range pairs (x_i^r, y_i^r) are generated under the settings in Section 3.1, with y_i^c following Type 3 and y_i^r following Type 1. Please refer to Section 3.1 for further details of the experimental setup. To quantify the relationship between two interval sequences, we compute four types of correlations: midpoint-based, symbolic, arithmetic-based, and the proposed Hausdorff correlation. The corresponding correlation values are 0.88, 0.86, 0.80, and 0.62, respectively. Compared to other methods, our method effectively incorporates the influence of interval lengths in the overall correlation analysis, achieving a more balanced consideration of both midpoint and length correlations. In contrast, other methods primarily focus on the correlation of the midpoint while overlooking the role of interval lengths. Furthermore, our method extends beyond single-segment interval data, accommodating multi-segment interval structures. As a concrete example, we apply our method to the Wearable Watch Dataset to assess the correlation between multi-segment sleep intervals and single-segment physiological signal intervals. As shown in Figure 1(b), the intervals at the same position on the horizontal line represent a multi-segment interval sample. The Hausdorff correlation between the intensity of high-frequency heart rate and sleep durations is estimated at 0.299, aligning well with physiological expectations. This demonstrates the broader applicability of our method, whereas other interval data analysis frameworks may not be suitable for handling multi-segment intervals. Detailed experiments and discussion can be found in Section 3.

The remainder of this paper is organized as follows. In Section 2, we start by introducing the fundamental concepts of Hausdorff distance, Fréchet mean, and Fréchet variance. And then, we extend the notion of Fréchet mean to an interval metric space equipped with the Hausdorff distance and introduce a novel relationship measure, Hausdorff covariance for random intervals. In Section 3, we compare our proposed method with existing mainstream correlation measures for interval-valued data, evaluating their performance across various bivariate association patterns and model settings. We utilize the wearable watch dataset to demonstrate the capability of our method in processing multi-segment data, a task that other methods are unable to accomplish. Additionally, two real-world datasets are used to assess the effectiveness of our method against other interval correlations in feature selection. Finally, Section 4 presents our conclusions, summarizing the proposed framework and outlining potential directions for future research.



(a) Single-segment interval case. (b) Multi-segment interval case.

Fig. 1 (a) An example illustrating the correlation between single-segment intervals, where the interval midpoints exhibit a strong positive correlation, while the interval lengths show a negative correlation. The computed correlation values are 0.88 (midpoint-based), 0.86 (symbolic), 0.80 (arithmetic-based), and 0.62 (Hausdorff). (b) An example demonstrating the correlation between multi-segment intervals and single-segment intervals. The computed Hausdorff correlation for this case is 0.299, whereas existing methods fail to compute a valid correlation

2 Hausdorff Correlation

In this section, we begin by introducing the definition of the Hausdorff metric along with its variant expressions. Next, we incorporate the Fréchet mean and Fréchet variance framework into the interval space equipped with the Hausdorff distance and establish fundamental theoretical guarantees. Building upon this foundation, we propose the Hausdorff covariance, a more general measure of dependency for random intervals without imposing any assumptions on algebraic operations.

2.1 The Hausdorff Distance

Hausdorff distance is a measure of dissimilarity between two subsets of a metric space $(\mathcal{M}, d)$. It is widely applied in various research fields, including computer vision (Zhao et al. 2005; Gao et al. 2013), fractal geometry (Min et al. 2007; Chen et al. 2011), and the medical domain (Sun et al. 2018; Karimi and Salcudean 2019), particularly in applications such as evaluating image registration and medical segmentation.

Let $\mathcal{I}$ denote the collection of all closed and bounded subsets of the metric space $(\mathcal{M}, d)$. The Hausdorff distance is traditionally defined as the minimum expansion range required to ensure mutual coverage of A and B. Given two elements A and $B \in \mathcal{I}$, let $d_H(A, B)$ represent the Hausdorff distance between them. The original formulation of the Hausdorff distance is given by:

$$d_H(A, B) = \inf\{r \geq 0 : A \subseteq B_r \text{ and } B \subseteq A_r\}, \tag{1}$$

where $A_r = \{x \in \mathbb{R} : d(x, A) \leq r\}$ represents the r-envelope of A. An equivalent and more commonly used formulation expresses the Hausdorff distance as follows (Birsan and Tiba 2006):

$$d_H(A, B) = \max\left\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\right\},$$

where $d(a, B) = \inf_{b\in B} d(a, b)$ and $d(b, A) = \inf_{a\in A} d(b, a)$ represent the minimum distance from a point to the opposing set. Furthermore, Rockafellar and Wets (2009) provided another equivalent representation:

$$d_H(A, B) = \sup_{\omega\in A\cup B} \left|\inf_{a\in A} d(\omega, a) - \inf_{b\in B} d(\omega, b)\right|.$$

Let the metric space $\mathcal{M} = \mathbb{R}$ and $\mathcal{I}$ accordingly denote the set of all nonempty compact intervals on the real number line, and $\mathcal{I}_c \subset \mathcal{I}$ be the collection of all nonempty, compact, and convex intervals. It can be formally proved that the Hausdorff distance defines a valid metric on $\mathcal{I}$, satisfying the properties of the identity of indiscernibles (zero self-distance), positivity, symmetry, and triangle inequality (Conci and Kubrusly 2018).

Several formal variations of the Hausdorff distance have been proposed to address different limitations. For instance,

Dubuisson and Jain (1994) introduced the modified Hausdorff distance, which reduces the impact of outliers by replacing the minimum distance between pairs of points in the intervals with their mean value. When faced with unevenly sampled data, Lu et al. (2001) developed the weighted Hausdorff distance, which improves computational accuracy by assigning different weights to points within the interval. Additionally, the partial Hausdorff distance (Huttenlocher et al. 1991) has demonstrated effectiveness in handling impulse noise. For the sake of brevity, we do not delve into these specific variations here, and we reserve the exploration of the broader Hausdorff family for future research.

In the single-segment interval case, the Hausdorff distance corresponds to the greatest distance from any point in one nonempty compact set to the closest point in the other set. Consequently, it can be explicitly represented as the maximum distance between the respective endpoints of the two intervals, as illustrated in Figure 2.

In the multi-segment interval case, the Hausdorff distance no longer admits an explicit solution. Given any two intervals $A, B \in \mathcal{I}$, where one can be expressed as a union of multiple disjoint sub-intervals, such that $A = \bigcup_{i=1}^{m} [a_i, b_i]$, the Hausdorff distance between A and B can be determined using the traditional definition given in Equation (1). To compute this distance, we employ a binary search strategy to iteratively identify the minimal range r that satisfies the condition in Equation (1), subject to a predefined accuracy threshold. Algorithm 1 summarizes the computation process for the Hausdorff distance between arbitrary interval structures. A simple example illustrating the Hausdorff distance between two multi-segment intervals is provided in Figure 3. For more complex cases, Algorithm 1 can be applied iteratively to obtain the desired distance.

Algorithm 1 Hausdorff distance computation between two intervals

1: **Input:** Intervals A, B; threshold value $\epsilon > 0$
2: **Initialize:** $r_1 \leftarrow \max_{\omega \in A \cup B} \omega - \min_{\omega \in A \cup B} \omega$; $r_0 \leftarrow r_1$
Initialized range $r_0 = r_1$
3: **while** $r_0 > \epsilon$ **do**
4: $r_0 \leftarrow r_0/2$
5: $r_1 \leftarrow r_1 - r_0$ when $A \subseteq B_{r_1}$ and $B \subseteq A_{r_1}$, otherwise $r_1 \leftarrow r_1 + r_0$
6: **end while**
7: **Return:** Hausdorff distance r_1

2.2 The Hausdorff Mean and Variance

In mathematics and statistics, the Fréchet mean and variance (Fréchet 1948) extend the classical notions of the centroid and variance to general metric spaces. These concepts serve as fundamental tools for characterizing the first- and second-order properties of a random object within a given metric space. This framework naturally extends to the space of intervals equipped with the Hausdorff distance. Given observations $x_1, \ldots, x_n$ of the random interval X within $\mathcal{I}$, the Hausdorff-Fréchet, or simply Hausdorff mean and variance of X are defined as:

$$x_\oplus = \arg\min_{x \in \mathcal{I}} \frac{1}{n} \sum_{i=1}^{n} d_H^2(x, x_i),$$
$$\mathrm{var}_\oplus(X) = \frac{1}{n} \sum_{i=1}^{n} d_H^2(x_\oplus, x_i). \tag{2}$$

For simplicity, we restrict attention to the case where the Hausdorff mean $x_\oplus$ takes a single-segment form, characterized by its interval midpoint $m_\oplus$ and range $r_\oplus$. In other words, we limit the optimized domain in Equation (2) to $\mathcal{I}_c$, since extending it to the full space $\mathcal{I}$ will lead to non-uniqueness of the solution and introduce additional optimized parameters:

$$x_\oplus = \arg\min_{x \in \mathcal{I}_c} \frac{1}{n} \sum_{i=1}^{n} d_H^2(x, x_i),$$
$$\mathrm{var}_\oplus(X) = \frac{1}{n} \sum_{i=1}^{n} d_H^2(x_\oplus, x_i). \tag{3}$$

The following theorems provide a theoretical justification for our definitions in Equation (3) under both single- and multi-segment settings. The detailed proof is provided in the Supplementary Material.

Theorem 1 (Existence of the Hausdorff mean) *Given intervals $\{x_i\}_{i=1}^n \subset \mathcal{I}$, there exists an interval in $\mathcal{I}_c$ that minimizes the mean of squares of distances to the sequence $\{x_i\}_{i=1}^n$.*

Theorem 2 (Uniqueness of the Hausdorff mean for single-segment intervals) *Given intervals $\{x_i\}_{i=1}^n \subset \mathcal{I}_c$ with corresponding interval ranges $\{r_i\}_{i=1}^n$, if the midpoint or range of x in Equation (3) is fixed, the minimizer $x_\oplus \in \mathcal{I}_c$ is unique.*

Theorem 1 establishes that for any given sequence of intervals $\{x_i\}_{i=1}^n$ in $\mathcal{I}$, the Hausdorff mean of single-segment form can be obtained by minimizing the sum of squared distances. Furthermore, Theorem 2 ensures that, when all intervals are single-segment, the optimizer in Equation (3) is uniquely determined once either the midpoint or the interval length of the Hausdorff mean is fixed. In practical applications, this result suggests that one can set the predetermined length of the Hausdorff mean as the average length of the given intervals $\{x_i\}_{i=1}^n$ and then optimize the interval midpoint to minimize the sum of squared distances to $\{x_i\}_{i=1}^n$. Alternatively, one may fix the interval midpoint and optimize the length accordingly.

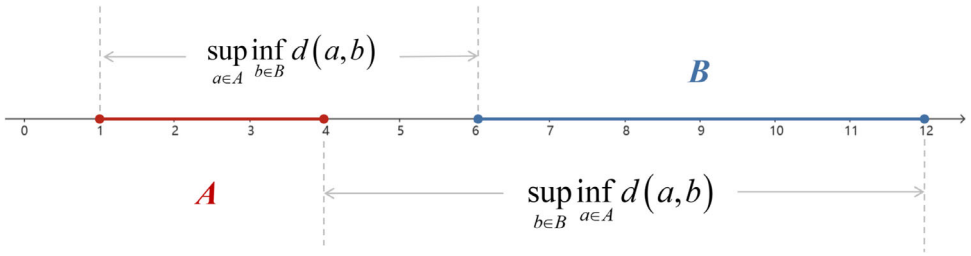


Fig. 2 An illustration of the Hausdorff distance between two single-segment intervals A and B. We first compute distances from each point in one interval to the closest point in the other, separately for both intervals, that is $\sup_{a\in A}\inf_{b\in B} d(a,b)$ and $\sup_{b\in B}\inf_{a\in A} d(a,b)$. The Hausdorff distance is then given by the maximum of these computed values

Theorem 3 (Uniqueness of the Hausdorff mean for multi-segment intervals) *Given intervals* $\{x_i\}_{i=1}^n \subset \mathcal{I}$ *with corresponding interval ranges* $\{r_i\}_{i=1}^n$*, if the range of* x *in Equation (3) is fixed to the average range across all intervals, the minimizer* $x_\oplus \in \mathcal{I}_c$ *is unique.*

Under the multi-segment setting, the uniqueness of the Hausdorff mean is guaranteed by Theorem 3 when the interval range of the Hausdorff mean is pre-specified. We remark that since $\mathcal{I}_c \subset \mathcal{I}$, Theorem 3 reduces to Theorem 2 with a fixed range when multi-segment intervals degenerate into single-segment intervals.

A heuristic procedure for computing a feasible Hausdorff mean for both single- and multi-segment intervals is summarized in Algorithm 2. This simplified strategy fixes the interval range (e.g., set to the average range across all intervals) and optimizes only the midpoint. An alternative algorithm based on a joint optimization strategy is provided in the Supplementary Material in the single-segment setting. The latter alternately updates both the midpoint and the range, and we show that it converges to the global optimum of Equation (3). However, it incurs substantially higher computational costs while yielding performance comparable to Algorithm 2. Further discussion of these two algorithms, along with their computational complexity and convergence analysis, is provided in the Supplementary Material.

Algorithm 2 Computation of the Hausdorff mean (simplified optimization strategy)

1: **Input:** Interval observations $\{x_i\}_{i=1}^n$ with corresponding interval ranges $\{r_i\}_{i=1}^n$
2: Compute the mean of the n ranges: $r_\oplus = \frac{1}{n}\sum_{i=1}^n r_i$
3: Determine the feasible domain of the midpoint: $L^m \le m \le U^m$, where $L^m = \min_{\omega\in\{x_i\}_{i=1}^n} \omega$ and $U^m = \max_{\omega\in\{x_i\}_{i=1}^n} \omega$
4: Construct the candidate single-segment interval: $x(m) = [m - \frac{r_\oplus}{2}, m + \frac{r_\oplus}{2}]$
5: Optimize the midpoint by solving the problem $m_\oplus = \arg\min_{L^m\le m\le U^m} \frac{1}{n}\sum_{i=1}^n d_H^2(x(m), x_i)$
6: **Return:** The Hausdorff mean $x_\oplus = [m_\oplus - \frac{r_\oplus}{2}, m_\oplus + \frac{r_\oplus}{2}]$

2.3 The Hausdorff Covariance and Correlation

To quantify the dependency between two random intervals in a more general setting, we introduce a novel correlation measure, termed Hausdorff correlation, which builds upon the extended concept of the Hausdorff mean. Consider a pair of bivariate random intervals (X_1, X_2) with observations $x_{11}, \ldots, x_{1n}$ and $x_{21}, \ldots, x_{2n}$, respectively. Let $x_{\oplus 1}$ and $x_{\oplus 2}$ denote their corresponding Hausdorff means. The Hausdorff covariance between X_1 and X_2 is then defined as:

$$\mathrm{Cov}(X_1, X_2) = \frac{1}{n}\sum_{i=1}^n \left\{ \mathrm{sgn}(A_{\omega_i}) \sup_{\omega\in x_{1i}\cup x_{2i}\cup x_{\oplus 1}\cup x_{\oplus 2}} |A_\omega| \right\},$$

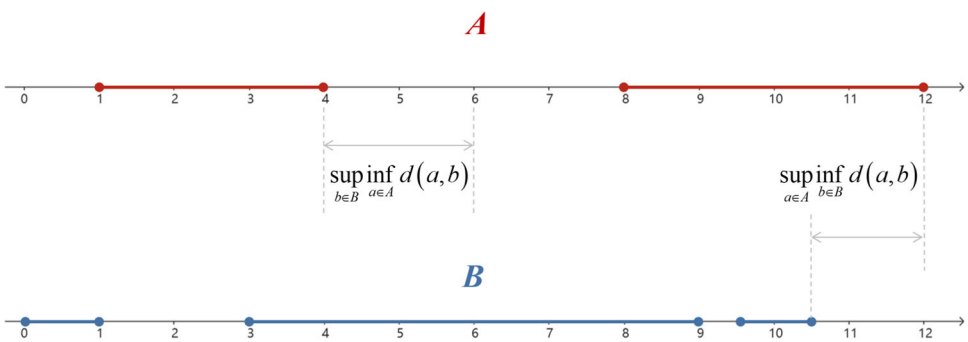


Fig. 3 A simple illustration of the Hausdorff distance between two multi-segment intervals A and B. The distances $\sup_{a\in A}\inf_{b\in B} d(a,b)$ and $\sup_{b\in B}\inf_{a\in A} d(a,b)$ are indicated in the figure. The Hausdorff distance is defined as the maximum of these two values. This distance can also be iteratively computed using Algorithm 1

(4)

where

$$A_\omega = \left(\inf_{a \in x_{1i}} d(\omega, a) - \inf_{a \in x_{\oplus 1}} d(\omega, a) \right) \left(\inf_{a \in x_{2i}} d(\omega, a) - \inf_{a \in x_{\oplus 2}} d(\omega, a) \right), \quad (5)$$

$$\omega_i = \arg\sup_{\omega \in x_{1i} \cup x_{2i} \cup x_{\oplus 1} \cup x_{\oplus 2}} |A_\omega|. \quad (6)$$

The proposed Hausdorff covariance is motivated by the classical definition of covariance for real-valued random variables. It measures the consistency of deviations between two sets of interval-valued data and their respective means. Below, we offer some intuitive explanations to facilitate understanding of this construction. First, the mathematical form of the Hausdorff covariance closely resembles that of classical covariance. In particular, $\sup |A_\omega|$ quantifies the deviation of interval-valued observations from their respective Fréchet means, and $\text{sgn}(A_{\omega_i})$ reflects the directionality (i.e., sign) of dependence in the interval setting. Second, the Hausdorff covariance numerically reduces to the classical covariance when lengths of the interval sequences $\{x_{1i}\}_{i=1}^n$ and $\{x_{2i}\}_{i=1}^n$ gradually shrink to zero under certain regularity conditions. Detailed simulation is provided in the Supplemental Material.

Accordingly, the Hausdorff correlation is obtained by normalizing the Hausdorff covariance:

$$\text{Cor}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{var}_\oplus(X_1)\text{var}_\oplus(X_2)}}.$$

The correlation proposed in our work exhibits several advantageous properties. Analogously to the standard correlation of random variables, it can be easily proved that when the interval midpoint and interval range of X_1 can be linearly represented by the interval midpoint and interval range of X_2, respectively, $\text{Cor}(X_1, X_2) = \pm 1$. Moreover, for any random interval pair (X_1, X_2), the following inequality holds, which is further proved in the Supplementary Material:

$$|\text{Cor}(X_1, X_2)| = \left| \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{var}_\oplus(X_1)\text{var}_\oplus(X_1)}} \right| \leq 1. \quad (7)$$

Using such definitions in Equation (4) has two major advantages when extending the Fréchet mean-variance notions to random intervals. First, the definitions of the Hausdorff mean, variance, and covariance are independent of any specific algebraic operations on intervals or assumptions regarding their underlying distributions. This generality eliminates the need to impose additional algebraic structures on the set of intervals, making the approach more flexible and widely applicable. In particular, the Hausdorff covariance serves as a novel correlation capable of measuring dependencies between multi-segment intervals. Second, by treating intervals as random objects, the proposed notions of mean, variance, and covariance have natural geometric interpretations within the corresponding interval metric space. The Hausdorff mean is defined as the minimizer of the mean of squared Hausdorff distances to the given interval $\{x_i\}_{i=1}^n$, effectively acting as their centroid. The Hausdorff variance quantifies the average squared distance from each interval to the Hausdorff mean, providing a measure of dispersion. The Hausdorff covariance, in turn, can be interpreted as an "inner product" between two sequences of intervals, aligning with the conventional formulation of covariance for random variables. Specifically, as shown in Figure 4, if the relative position of x_1 with respect to $x_{\oplus 1}$ aligns with that of x_2 relative to $x_{\oplus 2}$, then the value of A_ω in Equation (5) is positive; conversely, if their relative positions diverge, A_ω takes a negative value.

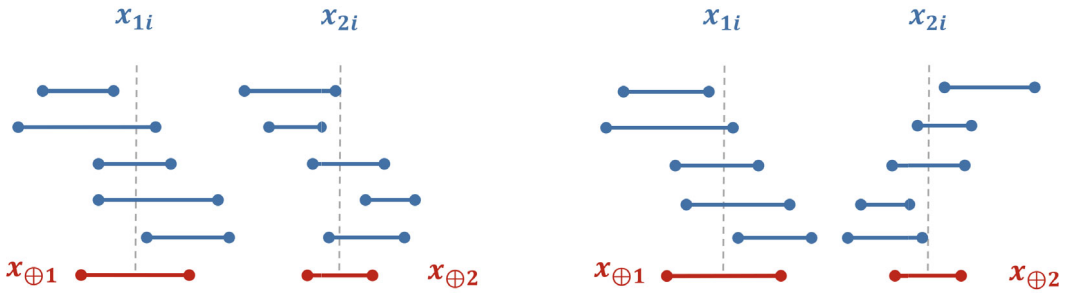
In terms of computational complexity, the computation of $\sup |A_\omega|$ requires four grid-based searches, each repeated n times, yielding an overall complexity of $O(n)$. Therefore, the computational complexity of the Hausdorff covariance for both single-segment interval data and multi-segment interval data is $O(n)$. The details of this derivation are provided in the supplementary materials.

3 Experimental Results

In this section, we conduct experiments on both synthetic and real-world datasets to evaluate the performance of conventional interval correlation measures alongside the proposed Hausdorff correlation. The benchmark methods include the midpoint correlation, symbolic correlation, and arithmetic-based correlation. The midpoint correlation method (Bock and Diday 2000; Birsan and Tiba 2006; Billard 2008) represents an interval by its midpoint and then calculates the corresponding Pearson correlation coefficient. The symbolic correlation method (Billard 2006, 2008) models intervals using empirical distribution functions and derives covariance from a functional perspective. The arithmetic-based correlation method (Gil et al. 2007; Sinova et al. 2012) combines the correlations of the midpoints and the lengths of the intervals, weighing them by their respective standard deviations to calculate an aggregated measure.

3.1 Synthetic Data

To illustrate the performance of Hausdorff correlation in capturing the interval-valued data structures, we compare correlation estimates obtained from four interval correlations: midpoint-based, symbolic, arithmetic-based, and Hausdorff



(a) Positive Hausdorff correlation case.

(b) Negative Hausdorff correlation case.

Fig. 4 (a) An example illustrating a positive Hausdorff correlation between intervals $\{x_{1i}\}$ and $\{x_{2i}\}$. The intervals in $\{x_{1i}\}$ and the intervals in $\{x_{2i}\}$ have the same positional relationship with respect to their own Hausdorff Means. In this case, the value of A_ω in Equation (5) takes more positive numbers, making the final correlation sign positive. (b) An example illustrating a negative Hausdorff correlation between intervals $\{x_{1i}\}$ and $\{x_{2i}\}$. The intervals in $\{x_{1i}\}$ and the intervals in $\{x_{2i}\}$ have the opposite positional relationship with respect to their own Hausdorff Means. In this case, the value of A_ω in Equation (5) takes more negative numbers, making the final correlation sign negative

correlation. This evaluation is conducted across nine different association scenarios between two interval-valued samples.

Consider the synthetic dataset $\Omega = \{(y_i, x_i)\}_{i=1}^n$ of observations from two interval-valued variables (Y, X). For each observation y_i (or x_i), its midpoint and range are represented by the single-valued variables y_i^c, and y_i^r (or x_i^c, x_i^r), respectively. To construct this dataset, we consider the following distributions to generate interval samples $\{(y_i, x_i)\}_{i=1}^n$ of size $n = 100$:

- x_i^c: x_i^c is sampled evenly spaced from 1 to 100;
- x_i^r: x_i^r is sampled evenly spaced from 5 to 50;
- y_i^c Type 1 (negatively correlated with x_i^c): $y_i^c = a_i + \epsilon_i$, where a_i are 100 equally spaced points from 100 to 1 and $\epsilon_i \sim N(0, 16)$;
- y_i^c Type 2 (not correlated with x_i^c): $y_i^c \sim \text{Uniform}(1, 100)$;
- y_i^c Type 3 (positively correlated with x_i^c): $y_i^c = a_i + \epsilon_i$, where a_i are 100 equally spaced points from 1 to 100 and $\epsilon_i \sim N(0, 16)$;
- y_i^r Type 1 (negatively correlated with x_i^r): $y_i^r = |a_i + \epsilon_i|$, where a_i are 100 equally spaced points from 50 to 5 and $\epsilon_i \sim N(0, 5)$;
- y_i^r Type 2 (not correlated with x_i^r): $y_i^r \sim \text{Uniform}(5, 50)$;
- y_i^r Type 3 (positively correlated with x_i^r): $y_i^r = |a_i + \epsilon_i|$, where a_i are 100 equally spaced points from 5 to 50 and $\epsilon_i \sim N(0, 5)$;

With the above settings, the correlation between x_i^c and y_i^c of type 1, 2, and 3 are -0.88, -0.04, and 0.88, respectively. The correlation between x_i^r and y_i^r of type 1, 2, and 3 are -0.95, 0.01, and 0.92, respectively. As depicted in Figure 5, this 3×3 graph illustrates that the first, second, and third columns have y_i^r of type 1, 2, and 3, while the first, second, and third rows have y_i^c of type 1, 2, and 3, respectively. Thus, we get nine types of $\{(y_i, x_i)\}_{i=1}^n$ samples, each with a sample size $n = 100$.

In the 3×3 matrix of graphs, the diagonal subgraphs illustrate that the correlations of interval midpoints and lengths are largely aligned, with all four correlation estimates being quite similar. In contrast, in the off-diagonal subgraphs, particularly the lower left and upper right, the correlations of interval midpoints and lengths differ. While the midpoint, symbolic, and arithmetic correlations primarily emphasize the correlation of interval midpoints, the Hausdorff correlation more effectively balances the correlations of both interval midpoints and lengths.

3.2 Analysis of Wearable Watch Dataset

The widespread adoption of smartwatches in the past decade has provided a convenient and effective means of collecting physiological time-series data during sleep. These devices continuously monitor time-series data such as sleep stages, heart rate, body temperature, and blood oxygen levels. The Wearable Watch Dataset (Walch 2019), collected in the work of sleep stage prediction (Walch et al. 2019), provides minute-level time series data for four sleep stages: wake, light sleep, deep sleep, and rapid eye movement (REM) sleep. In addition, heart rate, heart rate variability (HRV), oxygen saturation, and temperature are recorded per minute. HRV is characterized by three main characteristics: the root mean square of successive differences (RMSSD), the intensity of high-frequency heart rate f_{high}, and the intensity of low-frequency heart rate f_{low}. RMSSD quantifies short-term fluctuations between consecutive heartbeat intervals, providing insights into the rhythmicity of heart rate. Daily sleep data are aggregated into multi-segment interval representations based on the four stages, as illustrated in Figure 6.

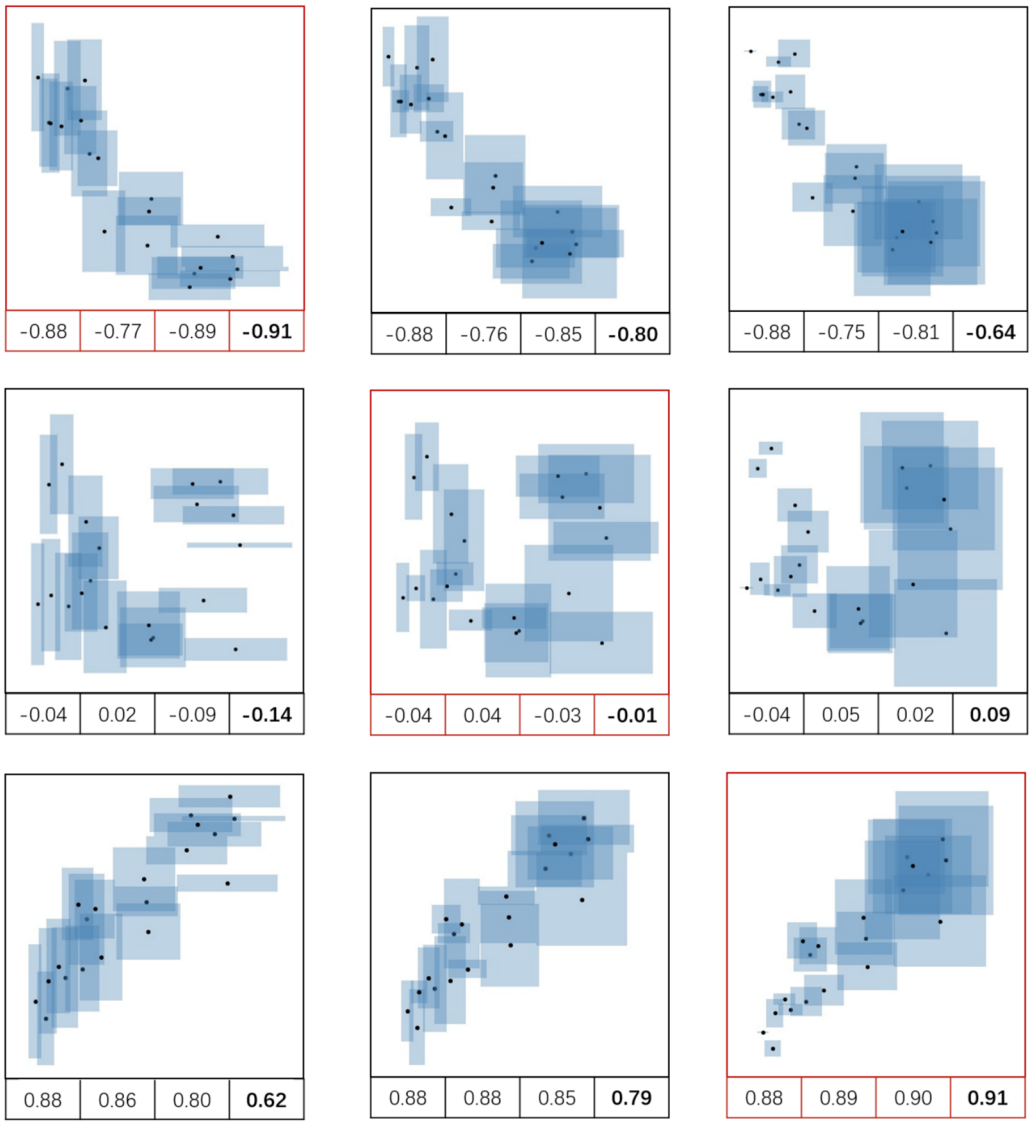


Fig. 5 Comparison between midpoint, symbolic, arithmetic-based, and Hausdorff correlation (The numbers in the table above are in this order) for several bivariate association patterns. The black dot indicates the midpoint of the interval, and the blue area indicates the area formed by a pair of interval samples. The fourth number in the table of each sub-graph represents the correlation value calculated by the Hausdorff correlation. The sub-graph borders on the diagonal are marked in red

For data preprocessing, we select the light sleep stage, the most predominant sleep phase in segmented sleep, as a representative of the overall sleep process. To construct data based on the single segment interval, we extract values within the 5th and 95th percentile range of HRV, oxygen saturation, and temperature for each day. The Hausdorff correlations between segmented sleep and other physiological features are presented in the following Table 1.

From Table 1, we observe a strong correlation of 0.903 between RMSSD and f_{high} in HRV. Regarding Segmented Sleep, it exhibits a higher correlation with f_{high} of 0.299 and a much lower correlation with f_{low} of 0.006, which aligns with physiological expectations. Specifically, f_{high} reflects the activity of the parasympathetic nervous system of the heart, which is known to be more active during sleep (Burgess et al. 1997; Zoccoli and Amici 2020).

Figure 7 presents the correlation plots for three pairs of variables. Figure 7(a) presents the relationship between RMSSD and f_{high}, where both midpoints and interval lengths exhibit a strong positive correlation. In Figure 7(b), the relationship between Segmented Sleep and f_{high} is depicted. As the positional range of f_{high} increases and extends upward, the total duration of sleep tends to be longer, leading to a general expansion in the time span. Finally, Figure 7(c) illus-

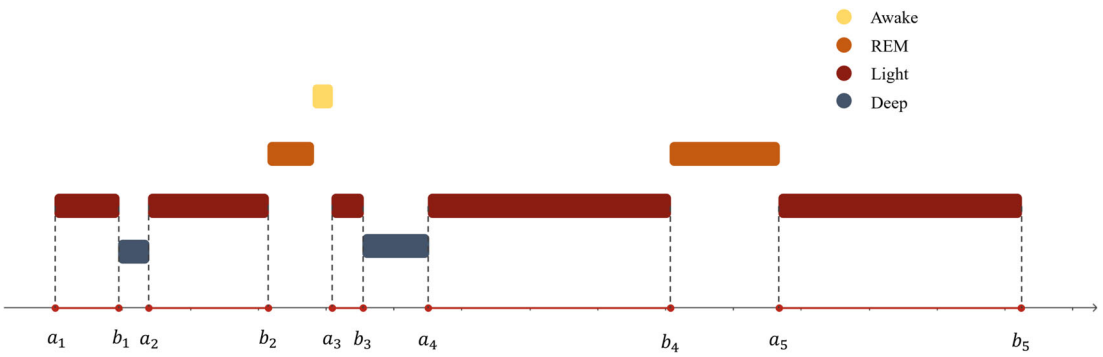


Fig. 6 Visualization of daily sleep patterns. Total sleep duration is segmented into four stages: wake, light sleep, deep sleep, and rapid eye movement (REM) sleep

Table 1 Hausdorff correlations for variables in the Wearable Watch Dataset. Correlations between Segmented Sleep, Heart Rate, RMSSD, and other physiological features are reported

	Segmented Sleep	Heart Rate	RMSSD	f_{high}	f_{low}	Temperature	Oxygen
Segmented Sleep	1	−0.259	0.193	0.299	0.006	0.298	−0.124
Heart Rate	−0.259	1	−0.290	−0.195	0.072	−0.038	−0.160
RMSSD	0.193	−0.290	1	0.903	0.025	0.097	0.060
f_{high}	0.299	−0.195	0.903	1	0.171	0.015	−0.002
f_{low}	0.006	0.072	0.025	0.171	1	−0.003	−0.091
Temperature	0.298	−0.038	0.097	0.015	−0.003	1	−0.020
Oxygen	−0.124	−0.160	0.060	−0.002	−0.091	−0.020	1

trates the relationship between Segmented Sleep and f_{low}, which shows little evident correlation.

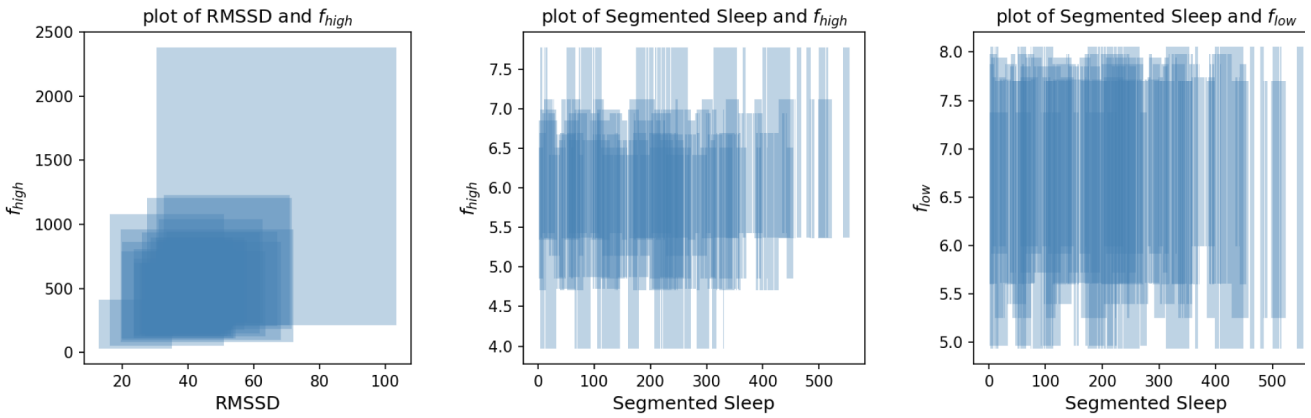
3.3 Feature Selection in Real-World Datasets

In this section, we implement a greedy feature selection strategy focused on maximizing dependence with respect to the response variable (Song et al. 2012; Lopez-Paz et al. 2013) using the Life Expectancy Dataset and the Sea level Dataset (Caldwell et al. 2015; Maharaj et al. 2019). Consistent with conventional procedures in the feature screening literature (Zhu et al. 2011; Li et al. 2023b), we begin by applying existing interval correlation measures to select important variables, and then evaluate their performance through a given regression modeling.

The Life Expectancy Dataset comprises life expectancy and various health-related factors for 193 countries from 2000 to 2015. It includes 19 variables that can potentially influence life expectancy, including factors related to immunization, mortality, economy, society, and other aspects of health. To construct interval-valued representations, after removing three highly autocorrelated variables, we extract the maximum and minimum values of these 16 variables, along with life expectancy statistics, across each country over the 15 years. After filtering out incomplete data, we obtain a final dataset consisting of 133 samples, where both the response and predictor variables are single-segment intervals.

Figure 8 illustrates the top three features that exhibit the strongest correlation with life expectancy, as identified by four different correlation methods: midpoint, symbolic, arithmetic-based, and Hausdorff correlation. The midpoint, symbolic, and arithmetic-based methods consistently emphasize *the Adult Mortality Rate (the probability of dying between ages 15 and 60 per 1000 population)*, *the Human Development Index (HDI)*, reflecting income-based resource composition, and *years of schooling* as the most influential features. In contrast, the Hausdorff correlation identifies *Gross Domestic Product (GDP) per capita* as the primary determinant of life expectancy, which aligns well with established theoretical insights and empirical findings (Acemoglu and Johnson 2007; Zaman et al. 2017). Additionally, the Hausdorff correlation uniquely highlights *HIV/AIDS* as an important feature, which is not recognized by the other three methods.

Next, we employ the Constrained Center and Range Method (CCRM) (Neto and de Carvalho 2010) regression model to investigate the relationship between life expectancy and the features selected by four different correlation methods. To assess the performance of the model, we randomly split the dataset, distributing 80% for training and 20% for



(a) RMSSD vs. f_{high}. (b) Segmented Sleep vs. f_{high}. (c) Segmented Sleep vs. f_{low}.

Fig. 7 Correlation plots between sleep-related variables. (a) Relationship between RMSSD and f_{high} with a Hausdorff correlation 0.903. (b) Relationship between Segmented Sleep and f_{high} with a Hausdorff correlation 0.299. (c) Relationship between Segmented Sleep and f_{low} with a Hausdorff correlation 0.006

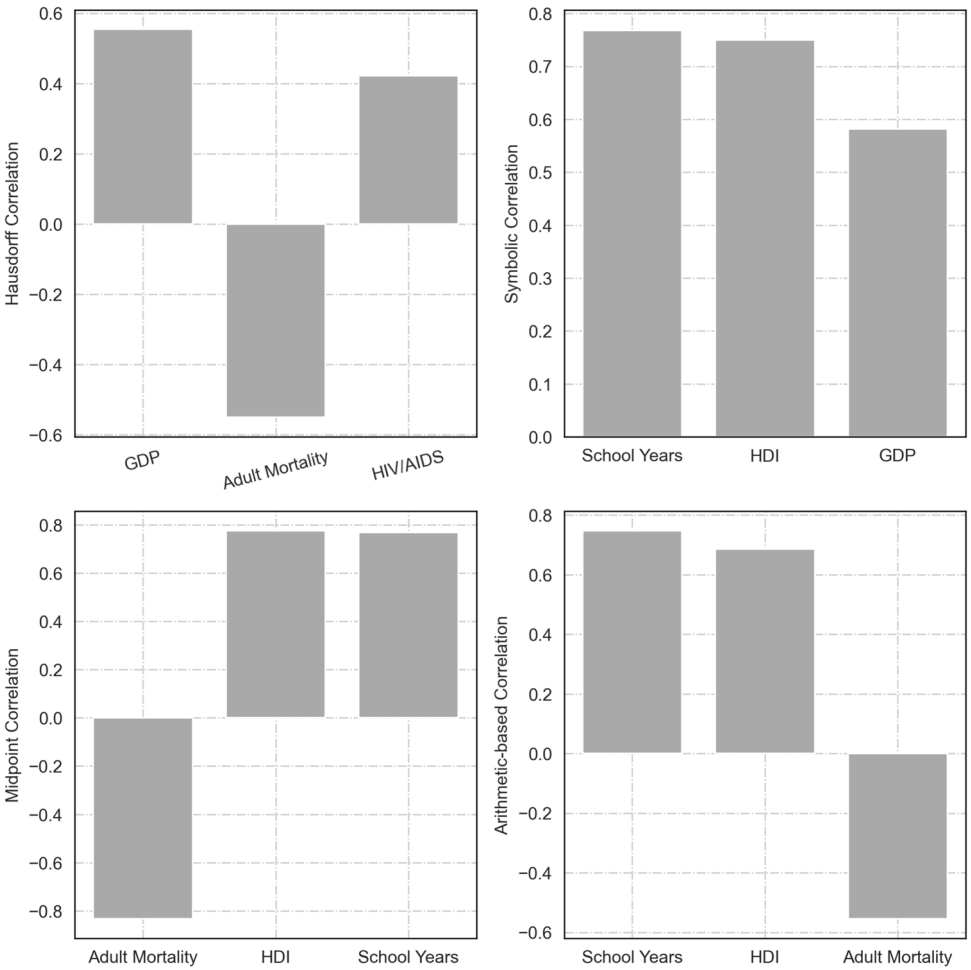
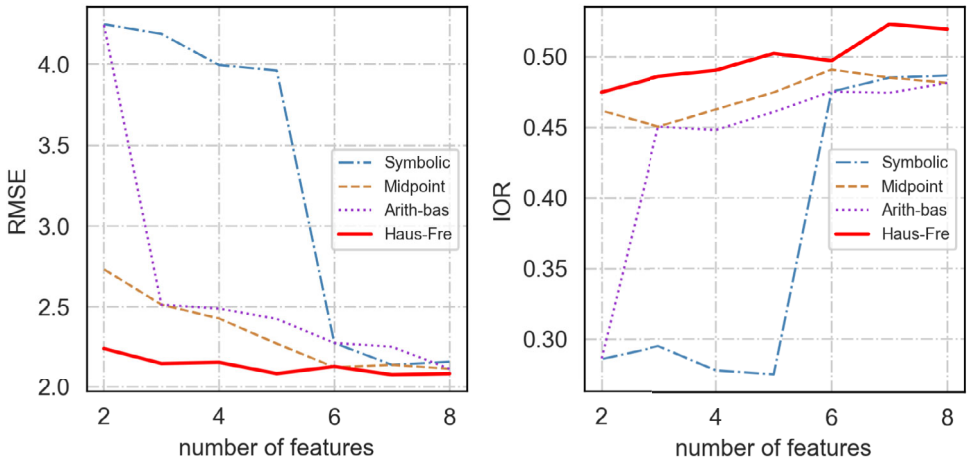
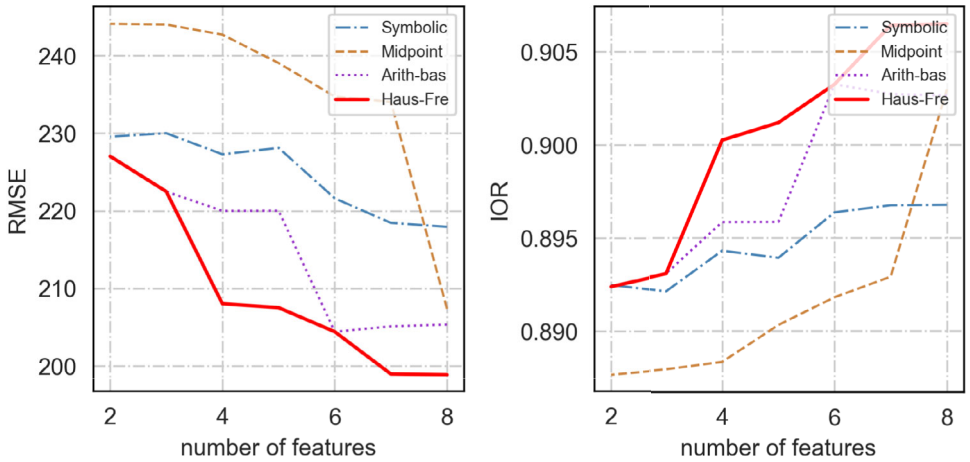


Fig. 8 Top three most important features that affect life expectancy, identified by midpoint, symbolic, arithmetic-based, and Hausdorff correlation



(a) Life Expectancy Dataset with selected features increasing from 2 to 8.



(b) Sea Level Dataset with selected features increasing from 2 to 12.

Fig. 9 Tendency on RMSE and IOR as the number of selected features increases in Life Expectancy Dataset and Sea Level Dataset. Hausdorff correlation performs better in most cases

testing. Evaluation is carried out using two key criteria on the test set:

- Root Mean Square Error (RMSE): The average RMSE of the interval midpoint and interval length.
- Interval Overlap Ratio (IOR): The overlap ratio between the predicted intervals and the true intervals. It is a widely used measure to quantify the differences between intervals (Wheeler et al. 2006; Kabir et al. 2017).

Lower RMSE and higher IOR indicate superior performance, reflecting a more precise interval estimation.

Figure 9(a) presents the results on the test set as the number of selected features increases from 2 to 8. As shown in the figure, increasing the number of selected features improves model performance across all correlation methods. However, the features identified by Hausdorff correlation consistently yield superior regression performance, underscoring their effectiveness in capturing key predictors of life expectancy.

The Sea Level Dataset consists of 500 days of observations collected from 16 ocean monitoring stations around Australia, recording the daily maximum and minimum sea levels. These values are used to construct an interval representation for each day. In this study, sea level measurements from the Booby Island observation station are designated as the dependent variable, while the measurements from the remaining 15 stations serve as independent variables. Similarly, after removing three highly autocorrelated variables, the dataset undergoes the same feature selection process, followed by CCRM regression. Figure 9(b) illustrates the regression performance as the number of selected features varies from 2 to 8. The results demonstrate that feature selec-

tion based on Hausdorff correlation consistently outperforms alternative methods, highlighting its effectiveness in identifying relevant predictors for sea level estimation.

More detailed descriptions of the data experiments are provided in the Supplementary Material 5.9.

4 Conclusion and Discussion

In this work, we propose Hausdorff correlation as a novel measure of dependence between two random intervals by incorporating the Fréchet framework into an interval metric space equipped with the Hausdorff distance, without traditional geometric and algebraic structures. By studying the Fréchet structure of random intervals, the proposed Hausdorff correlation offers distinct advantages over existing alternatives such as symbolic and arithmetic-based correlations. First, Hausdorff covariance is similar to the concept of regular covariance under an appropriate inner product and provides a strong geometric interpretation. It quantifies the degree of synchronization between random intervals and their Hausdorff means. Second, Hausdorff covariance is a more general dependency measure that imposes no additional algebraic definitions or distributional assumptions on interval-valued data. Notably, it applies to arbitrary interval forms, including complex multi-segment intervals, where existing methods fail to apply. We validated our method across various datasets, demonstrating that Hausdorff correlation achieves a superior balance in capturing dependencies related to both midpoints and interval lengths and greatly outperforms existing correlation measures in real-world feature selection tasks.

Several promising directions remain for future research. First, a more detailed investigation of the properties of the objective function is needed to better understand the conditions ensuring uniqueness of the Hausdorff mean, as well as to develop more refined optimization procedures. Second, the framework may be extended to multidimensional interval-valued data or to Hausdorff means with multi-segment forms, potentially through computationally efficient variants of the Hausdorff distance or by incorporating additional optimization constraints. Beyond these, the Hausdorff metric space framework opens new possibilities, such as utilizing Hausdorff distance as a loss function in regression models for interval-valued data, and extending concepts like Hausdorff mean and covariance to non-parametric interval fitting. These directions offer exciting potential for advancing interval data analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-025-10743-2.

Author Contributions Xinlai Kang wrote the main manuscript text. Cheng Meng and Xiaxue Ouyang contributed to the writing and data experiments. All authors reviewed the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Acemoglu, D., Johnson, S.: Disease and development: the effect of life expectancy on economic growth. J. Polit. Econ. **115**(6), 925–985 (2007)

Aumann, R.J.: Integrals of set-valued functions. J. Math. Anal. Appl. **12**(1), 1–12 (1965)

Billard, L.: Symbolic data analysis: what is it? In Compstat 2006-Proceedings in Computational Statistics: 17th Symposium Held in Rome, Italy, 2006, pages 261–269. Springer, (2006)

Billard, L.: Sample covariance functions for complex quantitative data. In Proceedings of World IASC Conference, Yokohama, Japan, pages 157–163, (2008)

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

Billard, L., Le-Rademacher, J.: Principal component analysis for interval data. Wiley Interdisciplinary Rev. Comput. Stat. **4**(6), 535–540 (2012)

Birsan, T., Tiba, D.: One hundred years since the introduction of the set distance by Dimitrie Pompeiu. In System Modeling and Optimization, pages 35–39. Springer, (2006)

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

Burgess, H.J., Trinder, J., Kim, Y., Luke, D.: Sleep and circadian influences on cardiac autonomic nervous system activity. Am. J. Physiology-Heart Circulatory Phys. **273**(4), H1761–H1768 (1997)

Caldwell, P., Merrifield, M., Thompson, P.: Sea level measured by tide gauges from global oceans as part of the joint archive for sea level since 1846. Centers Environ. Information, Dataset, 10:V5V40S47W, (2015)

Chen, J., Wang, R., Liu, L., Song, J.: Clustering of trajectories based on Hausdorff distance. In 2011 International Conference on Electronics, Communications and Control (ICECC), pages 1940–1944. IEEE, (2011)

Conci, A., Kubrusly, C.S.: Distance between sets-a survey. arXiv preprint arXiv:1808.02574 (2018)

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

Diamond, P.: Least squares fitting of compact set-valued data. J. Math. Anal. Appl. **147**(2), 351–362 (1990)

Dubuisson, M.-P., Jain, A.K.: A modified Hausdorff distance for object matching. In Proceedings of 12th International Conference on Pattern Recognition, volume 1, pages 566–568. IEEE, (1994)

D'Urso, P., Giordani, P.: A least squares approach to principal component analysis for interval valued data. Chemom. Intell. Lab. Syst. **70**(2), 179–192 (2004)

El Ghaoui, L., Lanckriet, G.R.G., Natsoulis, G., et al.: Robust classification with interval data. Computer Science, (2003)

Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. In Annales De l'institut Henri Poincaré **10**, 215–310 (1948)

Gao, Y., Wang, M., Ji, R., Wu, X., Dai, Q.: 3-d object retrieval with hausdorff distance learning. IEEE Trans. Industr. Electron. **61**(4), 2088–2098 (2013)

Gil, M.Á., González-Rodríguez, G., Colubi, A., Montenegro, M.: Testing linear independence in linear models with interval-valued data. Comput. Stat. Data Anal. **51**(6), 3002–3015 (2007)

Hu, Y., Li, M., Liu, X., Meng, C.: Sampling-based methods for multi-block optimization problems over transport polytopes. Math. Comput. **94**(353), 1281–1322 (2025)

Huttenlocher, D.P., Rucklidge, W.J., Klanderman, G.A.: Comparing images using the Hausdorff distance under translation. In Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 654–656. IEEE, (1991)

Kabir, S., Wagner, C., Havens, T.C., Anderson, D.T., Aickelin, U.: Novel similarity measure for interval-valued data based on overlapping ratio. In 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pages 1–6. IEEE, (2017)

Karimi, D., Salcudean, S.E.: Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Trans. Med. Imaging **39**(2), 499–513 (2019)

Li, M., Yu, J., Li, T., Meng, C.: Importance sparsification for sinkhorn algorithm. J. Mach. Learn. Res. **24**(247), 1–44 (2023a)

Li, T., Meng, C., Xu, H., Yu, J.: Hilbert curve projection distance for distribution comparison. IEEE Trans. Pattern Anal. Mach. Intell. **46**(7), 4993–5007 (2024)

Li, T., Yu, J., Meng, C.: Scalable model-free feature screening via sliced-wasserstein dependency. J. Comput. Graph. Stat. **32**(4), 1501–1511 (2023b)

Lopez-Paz, D., Hennig, P., Schölkopf, B.: The randomized dependence coefficient. Advances in Neural Information Processing Systems, 26, (2013)

Lu, Y., Tan, C.L., Huang, W., Fan, L.: An approach to word image matching based on weighted Hausdorff distance. In Proceedings of Sixth International Conference on Document Analysis and Recognition, pages 921–925. IEEE, (2001)

Maharaj, E.A., Teles, P., Brito, P.: Clustering of interval time series. Stat. Comput. **29**, 1011–1034 (2019)

Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W., Ma, P.: Large-scale optimal transport map estimation using projection pursuit. Adv. Neural Information Process. Syst., 32, (2019)

Meng, C., Yu, J., Zhang, J., Ma, P., Zhong, W.: Sufficient dimension reduction for classification using principal optimal transport direction. Adv. Neural. Inf. Process. Syst. **33**, 4015–4028 (2020)

Min, D., Zhilin, L., Xiaoyong, C.: Extended hausdorff distance for spatial objects in gis. Int. J. Geogr. Inf. Sci. **21**(4), 459–475 (2007)

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

Neto, E.D.A.L., Carvalho, F.D.A.: Constrained linear regression models for symbolic interval-valued variables. Comput. Stat. Data Analysis **54**(2), 333–347 (2010)

Petersen, A., Müller, H.-G.: Wasserstein covariance for multiple random densities. Biometrika **106**(2), 339–351 (2019)

Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, vol. 317. Springer Science & Business Media, Newyork (2009)

Sinova, B., Colubi, A., González-Rodrı, G., et al.: Interval arithmetic-based simple linear regression between interval data: discussion and sensitivity analysis on the choice of the metric. Inf. Sci. **199**, 109–124 (2012)

Song, L., Smola, A., Gretton, A., Bedo, J., Borgwardt, K.: Feature selection via dependence maximization. J. Mach. Learn. Research **13**, 1393–1434 (2012)

Sun, X., Zhang, N., Wu, H., Yu, X., Wu, X., Yu, S.: Medical image retrieval approach by texture features fusion based on hausdorff distance. Math. Probl. Eng. **2018**(1), 7308328 (2018)

Walch, O.: Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography. PhysioNet, 101, (2019)

Walch, O., Huang, Y., Forger, D., Goldstein, C.: Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep **42**(12), zsz180 (2019)

Wheeler, M.W., Park, R.M., Bailer, A.J.: Comparing median lethal concentration values using confidence interval overlap or ratio tests. Environmental Toxicology Chem. Int. J. **25**(5), 1441–1444 (2006)

Zaman, S.B., Hossain, N., Mehta, V., Sharmin, S., Mahmood, S.A.I.: An association of total health expenditure with GDP and life expectancy. J. Med. Research Innovation **1**(2), AU7–AU12 (2017)

Zhang, J., Ma, P., Zhong, W., Meng, C.: Projection-based techniques for high-dimensional optimal transport problems. Wiley Interdisciplinary Rev. Comput. Stat. **15**(2), e1587 (2023a)

Zhang, J., Meng, C., Yu, J., Zhang, M., Zhong, W., Ma, P.: An optimal transport approach for selecting a representative subsample with application in efficient kernel density estimation. J. Comput. Graph. Stat. **32**(1), 329–339 (2023b)

Zhang, J., Zhong, W., Ma, P.: A review on modern computational optimal transport methods with applications in biomedical research. Modern Stat. Methods Health Res., pages 279–300, (2021)

Zhao, C., Shi, W., Deng, Y.: A new hausdorff distance for image matching. Pattern Recogn. Lett. **26**(5), 581–586 (2005)

Zhu, L.-P., Li, L., Li, R., Zhu, L.-X.: Model-free feature screening for ultrahigh-dimensional data. J. Am. Stat. Assoc. **106**(496), 1464–1475 (2011)

Zoccoli, G., Amici, R.: Sleep and autonomic nervous system. Curr. Opin. Physio. **15**, 128–133 (2020)