






A Hausdorff Regression Paradigm for Interval Privacy

Xinlai Kang , Mengyu Li , Xuqiang Chen , *Graduate Student Member, IEEE*, Fangyu Li , *Member, IEEE*, and Cheng Meng 

Abstract—Data privacy has become a critical concern in today’s data-driven world. Interval privacy emerges as a promising safeguard, representing private values as intervals. Traditional interval analysis methods, however, often rely on critical assumptions that are questionable in practice. To address this gap, we propose a novel paradigm for analyzing interval-valued data generated by the interval privacy mechanism. Our contributions are two-fold: First, we innovatively model intervals as random objects in a metric space and use the Hausdorff distance to quantify their dissimilarity without imposing restrictive assumptions. Second, as an application of our paradigm, we develop an interval-to-interval regression method named Hausdorff distance-based regression (HDBR), extending multivariate linear regression to metric spaces. The HDBR method estimates regression coefficients by minimizing the Hausdorff distance between the observed and estimated intervals. Simulation studies demonstrate the effectiveness and robustness of our proposed approach compared to mainstream competitors. We also provide a real data example to illustrate how to perform regression analysis within the interval privacy framework, and the results further validate the superiority of the HDBR method.

Index Terms—Hausdorff distance, interval data, interval privacy, linear regression model, metric space.

I. INTRODUCTION

DATA privacy is a crucial aspect of data generation, storage, and processing within the context of increasing emphasis on data security. In the past decade, numerous methods have been developed to protect privacy. Traditional approaches rely on anonymization techniques [1]. Examples of such methods include HybrEx [2], k -anonymity [3], t -closeness [4], and l -diversity [5], aiming to render each released dataset indistinguishable concerning (w.r.t.) a minimum number of individuals in the population. These techniques safeguard published datasets from identity disclosure. However, the anonymization data remains a problem because it is hard to analyze and can’t maintain

Manuscript received 14 November 2023; accepted 5 December 2023. Date of publication 18 December 2023; date of current version 5 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 12101606. The work of Mengyu Li was supported by the Renmin University of China through Outstanding Innovative Talents Cultivation Funded Programs 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Guang Hua. (*Corresponding author: Cheng Meng.*)

Xinlai Kang, Mengyu Li, and Cheng Meng are with the Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China (e-mail: kangxinlai@ruc.edu.cn; limengyu516@ruc.edu.cn; chengmeng@ruc.edu.cn).

Xuqiang Chen and Fangyu Li are with the Faculty of Information Technology, Beijing University of Technology, Beijing 100022, China (e-mail: chenxuqiang-work@gmail.com; fangyu.li@bjut.edu.cn).

Digital Object Identifier 10.1109/LSP.2023.3344376

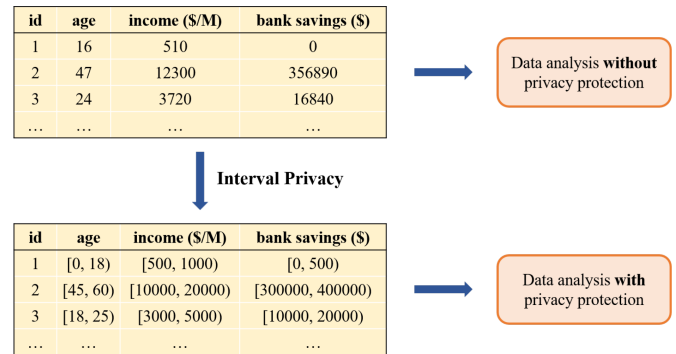


Fig. 1. Data processed by interval coverage can protect privacy during data analysis.

the relationship between the data. Another category of privacy protection methods that allow for analysis employs encryption techniques, including garbled circuits [6], homomorphic encryption [7], secret sharing [8], and others. More recent advancements in privacy protection involve noise-based algorithms, such as differential privacy [9], [10]. These techniques necessitate that the result of analyses conducted on a released dataset remains insensitive to the insertion or deletion of a tuple in the dataset. However, previous data privacy protection methods typically perturb the true value of the data, thus affecting its usability and accuracy.

In contrast, the interval privacy mechanism proposed by Ding and Ding [11] avoids the perturbation of true data; instead, it records each value as a random interval containing it as illustrated in Fig. 1. This mechanism has appealing characteristics such as conditional non-informativeness, information fidelity, privacy guarantee, and distributional identifiability [11]. Compared to other methods like differential privacy and encryption-based techniques, interval privacy is less demanding on data protocols, processing, and transmission conditions. While interval privacy is effective, analyzing interval-valued data requires different statistical methods than single-valued data, and the exploration of these methods remains limited in the literature.

In this context, some works have been conducted on analyzing the interval data. One natural analytical framework for interval data is modeling the midpoints and ranges of intervals [12], [13], [14]. For instance, Billard and Diday [15] computed the expectation and variance of interval-valued data based on interval midpoints and lengths. Neto and Carvalho [13], [16] conducted regression analysis on interval-valued data, and Le-Rademacher and Billard [17] performed principal component analysis on such data. Nevertheless, most of the existing methods rely on

critical assumptions, that is, (i) intervals lying in a vector space that allows calculation of summation and difference between intervals, and (ii) the true value obeying a specific distribution family on the interval. Such assumptions may not hold in practice, and misleading results may be generated when violated. Consequently, a more general paradigm for handling interval-valued data that requires fewer assumptions is still meager.

To bridge this gap, we propose a novel paradigm for analyzing interval-valued data in the context of interval privacy. We model an interval as a random object in a general metric space, without inherently possessing a vector space structure, and then quantify the dissimilarity between intervals using the Hausdorff distance, a popular metric for measuring the discrepancy between compact sets [18], [19], [20], [21]. Our major contributions are two-fold as follows:

- We develop an innovative framework that avoids imposing additional assumptions beyond defining the distance metric between intervals in the chosen metric space.
- We show the effectiveness of our framework by applying it to the regression analysis and numerical results on synthetic and real-world datasets show our framework outperforms or achieves comparable performance to mainstream competitors for interval-valued data analysis.

II. PROPOSED METHOD

In this section, we cover our interval-to-interval linear regression approach from the fundamental to the application. Firstly, we introduce the Hausdorff distance for measuring the discrepancy between intervals. Then we review the linear regression models for interval-valued data and lead to Hausdorff distance-based regression (HDBR).

A. Hausdorff Distance

The Hausdorff distance is a measure of the distance between compact sets, which can be naturally fitted into the setting of the interval set. We denote \mathcal{I} as the set of all closed intervals on the real number field. Let A and B be two elements in \mathcal{I} , and $d(A, B)$ be the distance between elements A and B in the set \mathcal{I} . The Hausdorff distance between the interval-valued data is defined as follows [18]:

$$\begin{aligned} d_H(A, B) &= \max \left\{ \sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A) \right\} \\ &= \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(b, a) \right\}, \quad (1) \end{aligned}$$

which is illustrated in Fig. 2. Intuitively, the Hausdorff distance between two non-empty sets is the maximum distance from any point in one set to its nearest point in the other set.

The Hausdorff distance has kinds of formal variants discussed in the literature [18], [20], [22], such as the normalization of (1) and the modified Hausdorff distance [23]. For brevity, we don't detail these specific variants in this paper and leave the investigation of the general Hausdorff family for further work.

B. Interval-to-Interval Regression

For analysis of the interval data based on Hausdorff distance, we establish an interval-to-interval regression analysis algorithm by transforming the metrics space of the traditional regression analysis. Traditional regression analysis aims to estimate the

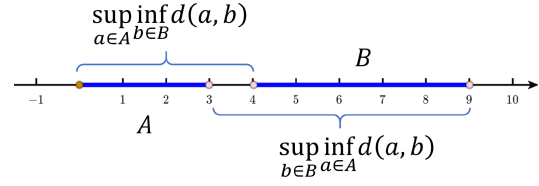
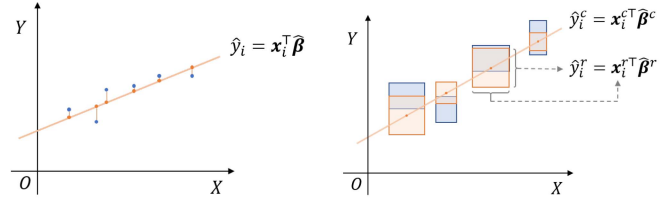


Fig. 2. Hausdorff distance between interval-valued data. A and B are two intervals on the real number axis, with a and b representing points on A and B , respectively. We calculate the sup-inf values of the two components separately and then take the maximum value as the Hausdorff distance between the two intervals.



(a) Linear regression model for single-valued variables. (b) Linear regression model for interval-valued variables.

Fig. 3. Illustration of linear regression models for single-valued and interval-valued variables. (a) The blue and orange points are observed and predicted values, respectively. (b) The blue and orange rectangles are observed and predicted intervals, respectively.

relationships between a response variable and one or more predictor variables, as illustrated in Fig. 3(a). As for the linear regression of interval-valued data, a popular method is finding the lines that most closely fit the midpoint and range of the data according to a specific mathematical criterion, as shown in Fig. 3(b).

Let $\Omega = \{s_1, \dots, s_n\}$ be a set of observations from $(p+1)$ interval-valued variables Y, X_1, \dots, X_p , where $s_i = (y_i, x_{i1}, \dots, x_{ip})$. The lower bound, upper bound, midpoint, and range of the interval y_i (or x_{ij}) are denoted by single-valued variables y_i^l, y_i^u, y_i^c , and y_i^r (or $x_{ij}^l, x_{ij}^u, x_{ij}^c, x_{ij}^r$), respectively. The quantities denoted above obey the following operational relationships: $y_i^c = (y_i^l + y_i^u)/2$, $y_i^r = (y_i^u - y_i^l)/2$, $x_{ij}^c = (x_{ij}^l + x_{ij}^u)/2$, and $x_{ij}^r = (x_{ij}^u - x_{ij}^l)/2$, for $i = 1, \dots, n$ and $j = 1, \dots, p$. Moreover, we use the notation $[y_i^c, y_i^r]$ (or $[x_{ij}^c, x_{ij}^r]$) to represent the interval y_i (or x_{ij}).

We aim to develop a regression model where the predictors and responses are intervals in a general metric space. In particular, we fit a multivariate linear regression to the interval-valued dataset Ω . According to the previous work [12], [13], [14], [24], a common approach is to model the midpoint and range of intervals respectively as follows:

$$\begin{cases} y_i^c = \mathbf{x}_i^{cT} \boldsymbol{\beta}^c + \varepsilon_i^c \\ y_i^r = \mathbf{x}_i^{rT} \boldsymbol{\beta}^r + \varepsilon_i^r \end{cases}, \text{ for } i = 1, \dots, n. \quad (2)$$

Here, $\mathbf{x}_i^c = (1, x_{i1}^c, \dots, x_{ip}^c)$ and $\mathbf{x}_i^r = (1, x_{i1}^r, \dots, x_{ip}^r)$ denote predictor vectors for the midpoint and range, respectively; ε_i^c and ε_i^r represent random noises; and $\boldsymbol{\beta}^c$ and $\boldsymbol{\beta}^r$ are unknown coefficient vectors.

To estimate the interval-to-interval linear model, Billard and Diday [25] proposed the center method (CM), which establishes two linear models with the same coefficients on the lower and

Algorithm 1: Hausdorff Distance-Based Regression.

-
- 1: **procedure** HDBR $\{y_i^c\}_i, \{y_i^r\}_i, \{x_i^c\}_i, \{x_i^r\}_i$
 - 2: $(\hat{\beta}^c, \hat{\beta}^r) = \arg \min_{\beta^c, \beta^r} L(\beta^c, \beta^r)$, where the loss function L is defined by (3)
 - 3: **end procedure**
-

upper bounds of the intervals. Subsequently, Neto and Carvalho [13] introduced the center and range method (CRM), fitting the two linear models in (2) independently. Further, several improvements have been proposed for CRM to ensure the non-negativity of interval ranges, including the constrained center and range method (CCRM) [16], lasso-based interval-valued regression model (Lasso-IR) [24], and the constrained center and range joint model (CCRM) [14]. There are also approaches focusing on robust regression techniques for interval-valued data.

However, most of these approaches have two main limitations. First, they rely on questionable assumptions, as discussed in Section I. Second, they separately estimate the two models in (2) or combine them in some manner (e.g., using a weighted average of their respective losses). Instead, a more natural and reasonable alternative is to directly minimize the “distance” between the observed and predicted intervals. For example, Gil et al. [26] regarded intervals as support functions in a Hilbert space and calculated their ℓ_2 distance. Considering our focus on interval-valued variables in a metric space, it is natural to propose using the Hausdorff distance as the loss function to fit the regression model in (2). The resulting loss function is defined as

$$L(\beta^c, \beta^r) := \sum_{i=1}^n d_H([\mathbf{y}_i^c, \mathbf{y}_i^r], [\mathbf{x}_i^{c\top} \beta^c, \mathbf{x}_i^{r\top} \beta^r]), \quad (3)$$

that is, the cumulative Hausdorff distance across all data points, calculated as the summation of individual Hausdorff distances between each observed interval $[\mathbf{y}_i^c, \mathbf{y}_i^r]$ and its corresponding predicted interval $[\mathbf{x}_i^{c\top} \beta^c, \mathbf{x}_i^{r\top} \beta^r]$. Compared to the references mentioned above, the proposed criterion (3) incorporates the information of the entire intervals rather than solely relying on midpoints and ranges. Such a property indicates its improved robustness against outliers, which will be further demonstrated in Section III.

C. HDBR Algorithm

The procedure for estimating the coefficients β^c and β^r in the model (2), referred to as the Hausdorff Distance-Based Regression (HDBR), is summarized in Algorithm 1.

The objective of Algorithm 1 is to find the values of $\hat{\beta}^c$ and $\hat{\beta}^r$ that minimize the loss function L , serving as a measure of model accuracy. Since the optimization problem in Algorithm 1 does not have an analytical solution, we commence the process with all-zero vectors as the initial values and utilize the Nelder – Mead method [27], a widely-used optimization technique, to iteratively solve the problem. The solutions $\hat{\beta}^c$ and $\hat{\beta}^r$ correspond to the regression coefficient estimates of the model (2). Consequently, the predicted interval for the i th observation can be expressed as $[\mathbf{x}_i^{c\top} \hat{\beta}^c, \mathbf{x}_i^{r\top} \hat{\beta}^r]$.

III. NUMERICAL RESULTS

In this section, we evaluate the HDBR algorithm for the regression analysis of interval data. We compare HDBR with state-of-the-art competitors, including CM [25], CRM [13], CCRM [16], and the DK method [26], on both synthetic and real-world datasets. Regarding the proposed HDBR algorithm, we take the distance d in the (2) as the Euclidean distance.

Datasets: For the synthetic dataset, we randomly generate the interval data and introduce four types of outliers [28], [29]. The first three types may arise due to errors in the data storage or collection process, while the fourth type may be attributed to an inherent abnormality in the data distribution; see Section III-A for more details. For the real-world dataset, we consider the maternal health risk dataset from UCI Machine Learning Repository [30] with 1014 observations and 5 variables, collected from various healthcare facilities (e.g., hospitals, community clinics, and maternal health centers) in rural areas of Bangladesh using an IoT-based risk monitoring system. The variables in the dataset contain age, systolic and diastolic blood pressure, blood glucose levels, heart rate, and risk level during pregnancy.

Metrics: To evaluate the performance of various regression approaches, we employed a combination of the root mean squared errors (RMSE) of the midpoint, range, lower bound, and upper bound of intervals, denoted as RMSE_C , RMSE_R , RMSE_L , and RMSE_U , based on the testing set \mathcal{I}_{test} comprising 200 observations. In particular, the RMSE criterion is given by

$$\text{RMSE} = \frac{\text{RMSE}_C + \text{RMSE}_R + \text{RMSE}_L + \text{RMSE}_U}{4}, \quad (4)$$

where $\text{RMSE}_\alpha = \sqrt{\sum_{i \in \mathcal{I}_{test}} (y_i^\alpha - \hat{y}_i^\alpha)^2 / 200}$, with α representing each component, i.e., $\alpha \in \{C, R, L, U\}$. Here, $y_i = [\mathbf{y}_i^c, \mathbf{y}_i^r]$ and $\hat{y}_i = [\hat{\mathbf{y}}_i^c, \hat{\mathbf{y}}_i^r]$ represent the observed intervals and corresponding predicted ones, respectively; $\hat{y}_i^c = \hat{\mathbf{y}}_i^c - \hat{\mathbf{y}}_i^r$ and $\hat{y}_i^u = \hat{\mathbf{y}}_i^c + \hat{\mathbf{y}}_i^r$ are predictions of y_i^l and y_i^u , respectively.

A. Synthetic Data Example

The synthetic dataset consists of a response interval-valued variable Y and a predictor interval-valued variable X . To generate the dataset, we follow the procedure described below.

First, we generate a sample $\{(y_i, x_i)\}_{i=1}^n$ of size $n = 700$ according to the following distributions:

$$\begin{aligned} x_i^c &\sim \text{Uniform}(0, 20), & x_i^r &= |\beta^* x_i^c + \varepsilon_i^r|, \\ y_i^c &= \beta_0^c + \beta_1^c x_i^c + \varepsilon_i^c, & y_i^r &= |\beta^* y_i^c + \varepsilon_i^r|, \end{aligned} \quad (5)$$

where $\beta_0^c \sim \text{Uniform}(-10, -5)$, $\beta_1^c \sim \text{Uniform}(0, 2)$, $\beta^* \sim \text{Uniform}(0.2, 0.5)$, and $\varepsilon_i^c, \varepsilon_i^r \sim \text{Uniform}(0, 5)$.

Next, we randomly partition the generated sample into training and testing sets according to the ratio of 5:2. In the training set, we replace 10% of the observations with the following four types of outliers evenly [28], [29]:

- Type 1 (extremely large): $y_i^c = 1000$, $y_i^r = 250$;
- Type 2 (extremely small): $y_i^c = -1000$, $y_i^r = 0.1$;
- Type 3 (empty data): y_i^c and y_i^r are generated from a Bernoulli distribution with a probability of 0.5 to take 0 or 1;
- Type 4 (heavy-tailed noise): $y_i^c = \beta_0^c + \beta_1^c x_i^c + t_i^c$, $y_i^r = |\beta^* y_i^c + t_i^r|$, where t_i^c and t_i^r are generated from the Student's t-distribution with 2 degrees of freedom.

TABLE I
AVERAGE RMSE (WITH STANDARD DEVIATIONS PRESENTED IN PARENTHESIS)
FOR SYNTHETIC DATASETS UNDER DIFFERENT TYPES OF OUTLIERS

Method	Type 1	Type 2	Type 3	Type 4
CM	51.32 (7.87)	101.75 (10.44)	6.89 (4.11)	7.62 (11.24)
CRM	39.28 (6.16)	81.47 (6.70)	6.57 (4.07)	8.36 (33.85)
CCRM	39.29 (6.15)	81.45 (6.67)	6.59 (4.05)	7.75 (16.91)
DK	39.29 (6.20)	81.47 (7.68)	6.57 (4.07)	8.37 (33.84)
HDBR	6.83 (4.30)	6.81 (4.31)	6.77 (4.26)	6.82 (4.30)

* The best result of each dataset is in bold.

When introducing the outliers, no outliers were introduced in the testing set, ensuring its integrity for evaluating the models' performance. According to the seminal work of Ding and Ding [11], privacy coverage can be employed to quantify privacy levels, with higher values indicating enhanced privacy. In the above four cases, the privacy coverage for X is 0.519, and those for Y are 0.247, 0.207, 0.208, and 0.230, respectively.

We calculate the mean and standard deviation of RMSE based on 1000 replications, and the results under four types of outliers are reported in Table I. Note that the coefficients β_0 , β_1 , and β^* are randomly generated only once in each regression analysis, while the random noises and observations are regenerated for each replication.

We observe the proposed HDBR method yields significantly smaller errors than its competitors in most cases as shown in Table I. Such an observation indicates the superior robustness of our method against data corruption compared to the traditional methods. We also observe a positive relationship between RMSE and privacy coverage, which aligns with our expectation that increased privacy levels decrease the information contained within the data, leading to a decline in prediction accuracy.

B. Real Data Example

In the real-world data, the task is to predict the risk level during pregnancy using the remaining variables. The relationships between the risk level and other variables are shown in Fig. 4.

The blood pressure variable is naturally interval-valued and can be directly utilized in the analysis. We adopt the canonical interval mechanism proposed in [11] to handle other variables. For the population age variable, we construct coverage intervals following the conventions commonly used in population surveys [31], [32], which include categories such as <18, 18-24, and so on. Regarding the value-at-risk variable, the boundaries of the coverage intervals are generated following uniform distributions, and different risk levels are associated with distinct uniform distributions.

For the heart rate and blood glucose level variables, the boundaries of the coverage intervals are generated based on normal distributions, with the standard deviation of the normal distribution set to equal the respective standard deviations of these variables. We assume that recording errors in this dataset cause no outliers. Considering the complexity and volatility of real-world scenarios, we randomly replace 10% of the data with Type 4 outliers introduced in Section III-A.

This experiment employs the 10-fold cross-validation, and the results are presented in Table II. The proposed HDBR method outperforms other approaches w.r.t. both the mean and standard deviation of RMSE. This observation supports the effectiveness

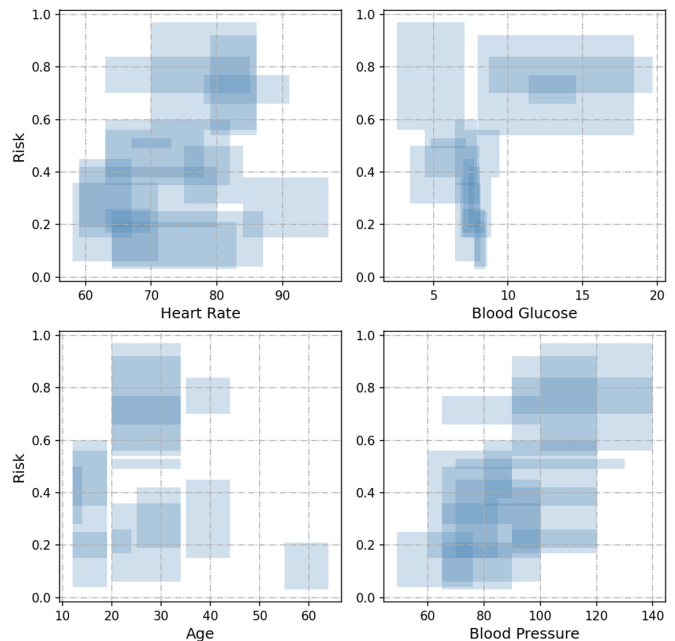


Fig. 4. Relationships between the risk level and predictor variables (i.e., heart rate, blood glucose, age, and blood pressure). For visual clarity, each subfigure shows 15 observations randomly drawn from the data.

TABLE II
AVERAGE RMSE $\times 10$ (WITH STANDARD DEVIATIONS PRESENTED IN
PARENTHESIS) FOR THE MATERNAL HEALTH RISK DATASET

Method	RMSE $\times 10$
CM	3.35 (1.66)
CRM	3.05 (0.59)
CCRM	2.98 (0.53)
DK	2.96 (0.55)
HDBR	1.55 (0.09)

* The best result is in bold.

of our interval-valued data analysis framework, indicating its applicability to complex real-life scenarios.

IV. CONCLUSION

We present a novel interval-valued data analysis paradigm for interval privacy. Specifically, we leverage the Hausdorff distance to measure the dissimilarity between intervals in the metric space, offering a more natural and reasonable interval-to-interval regression method. While our experimental results demonstrate the effectiveness and robustness of the proposed approach, we acknowledge the absence of theoretical guarantees, such as asymptotic properties, in our current study. This aspect remains an open avenue for future research. We also plan to extend the proposed interval regression paradigm to a broad range of data analysis techniques, including clustering and dimensionality reduction. These directions are also left to future works.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and Action Editor for their constructive comments that improved the quality of this letter.

REFERENCES

- [1] S. De Capitani Di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data privacy: Definitions and techniques," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 20, no. 06, pp. 793–817, 2012.
- [2] S. Y. Ko, K. Jeon, and R. Morales, "The HybrEx model for confidentiality and privacy in cloud computing," in *Proc. 3rd USENIX Workshop Hot Topics Cloud Comput.*, 2011, pp. 2–4.
- [3] L. Sweeney, " k -anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.
- [4] N. Li, T. Li, and S. Venkatasubramanian, " t -closeness: Privacy beyond k -anonymity and l -diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, 2006, pp. 106–115.
- [5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, " l -diversity: Privacy beyond k -anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 3–es, 2007.
- [6] A. C.-C. Yao, "How to generate and exchange secrets," in *Proc. 27th Annu. Symp. Foundations Comput. Sci.*, 1986, pp. 162–167.
- [7] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. 41st Annu. ACM Symp. Theory Comput.*, 2009, pp. 169–178.
- [8] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, 1979.
- [9] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang., Program.*, 2006, pp. 1–12.
- [10] L. Wasserman and S. Zhou, "A statistical framework for differential privacy," *J. Amer. Stat. Assoc.*, vol. 105, no. 489, pp. 375–389, 2010.
- [11] J. Ding and B. Ding, "Interval privacy: A framework for privacy-preserving data collection," *IEEE Trans. Signal Process.*, vol. 70, pp. 2443–2459, 2022.
- [12] P. Bertrand and F. Goupil, "Descriptive statistics for symbolic data," in *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin, Germany: Springer, 2000, pp. 106–124.
- [13] E. d. A. L. Neto and F. D. A. De Carvalho, "Centre and range method for fitting a linear regression model to symbolic interval data," *Comput. Statist. Data Anal.*, vol. 52, no. 3, pp. 1500–1515, 2008.
- [14] P. Hao and J. Guo, "Constrained center and range joint model for interval-valued symbolic data regression," *Comput. Statist. Data Anal.*, vol. 116, pp. 106–138, 2017.
- [15] L. Billard and E. Diday, "From the statistics of data to the statistics of knowledge: Symbolic data analysis," *J. Amer. Stat. Assoc.*, vol. 98, no. 462, pp. 470–487, 2003.
- [16] E. d. A. L. Neto and F. D. A. De Carvalho, "Constrained linear regression models for symbolic interval-valued variables," *Comput. Statist. Data Anal.*, vol. 54, no. 2, pp. 333–347, 2010.
- [17] J. Le-Rademacher and L. Billard, "Principal component histograms from interval-valued observations," *Comput. Statist.*, vol. 28, pp. 2117–2138, 2013.
- [18] T. Birsan and D. Tiba, "One hundred years since the introduction of the set distance by Dimitrie Pompeiu," in *Proc. IFIP Conf. Syst. Model. Optim.*, 2006, pp. 35–39.
- [19] J. H. Park, K. M. Lim, J. S. Park, and Y. C. Kwun, "Distances between interval-valued intuitionistic fuzzy sets," in *Proc. J. Phys.: Conf. Ser.*, vol. 96, 2008, Art. no. 012089.
- [20] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317. Berlin, Germany: Springer, 2009.
- [21] A. Conci and C. Kubrusly, "Distance between sets—A survey," 2018, [arXiv:1808.02574](https://arxiv.org/abs/1808.02574).
- [22] E. Deza, M. M. Deza, M. M. Deza, and E. Deza, *Encyclopedia of Distances*. Berlin, Germany: Springer, 2009.
- [23] M.-P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proc. 12th Int. Conf. Pattern Recognit.*, 1994, pp. 566–568.
- [24] P. Giordani, "Lasso-constrained regression analysis for interval-valued data," *Adv. Data Anal. Classification*, vol. 9, no. 1, pp. 5–19, 2015.
- [25] L. Billard and E. Diday, "Regression analysis for interval-valued data," in *Data Analysis, Classification, and Related Methods*. Berlin, Germany: Springer, 2000, pp. 369–374.
- [26] M. A. Gil, M. A. Lubiano, M. Montenegro, and M. T. López, "Least squares fitting of an affine function and strength of association for interval-valued data," *Metrika*, vol. 56, no. 2, pp. 97–111, 2002.
- [27] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.
- [28] R. A. Fagundes, R. M. De Souza, and F. J. A. Cysneiros, "Robust regression with application to symbolic interval data," *Eng. Appl. Artif. Intell.*, vol. 26, no. 1, pp. 564–573, 2013.
- [29] G. Lecué and M. Lerasle, "Robust machine learning by median-of-means: Theory and practice," *Ann. Statist.*, vol. 48, no. 2, pp. 906–931, 2020.
- [30] M. Ahmed, "Maternal health risk," UCI Machine Learning Repository, 2023, doi: [10.24432/C5DP5D](https://doi.org/10.24432/C5DP5D).
- [31] E. M. Badley and A. Tennant, "Changing profile of joint disorders with age: Findings from a postal survey of the population of calderdale, West Yorkshire, United Kingdom," *Ann. Rheumatic Dis.*, vol. 51, no. 3, pp. 366–371, 1992.
- [32] A. Bowling and J. Windsor, "Towards the good life: A population survey of dimensions of quality of life," *J. Happiness Stud.*, vol. 2, no. 1, pp. 55–82, 2001.