Expectation Distance-Based Distributional Clustering for Noise-Robustness

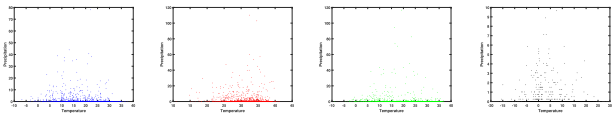
Rahmat Adesunkanmi, *Student Member, IEEE*, and Ratnesh Kumar, *Fellow, IEEE*

***Abstract*—This paper presents a clustering technique that reduces the susceptibility to data noise by learning and clustering the data-distribution and then assigning the data to the cluster of its distribution. In the process, it reduces the impact of noise on clustering results. This method involves introducing a new distance among distributions, namely the expectation distance (denoted, ED), that goes beyond the state-of-art distribution distance of optimal mass transport, also called 2-Wasserstein (denoted, W_2): The latter essentially depends only on the marginal distributions while the former also employs the information about the joint distributions, making it more powerful. Using the ED, the paper extends the classical K-means and K-medoids clustering to those over data-distributions (rather than raw-data) and further introduces K-medoids using W_2. The paper also presents the closed-form expressions of the W_2 and ED distance measures. The implementation results of the proposed ED and the W_2 distance measures to cluster real-world weather data as well as stock data are also presented, which involves efficiently extracting and using the underlying data distributions—Gaussians for weather data versus lognormals for stock data. The results show striking performance improvement over classical clustering of raw-data, with higher accuracy realized for ED. Also, not only does the distribution-based clustering offer higher accuracy, but it also lowers the computation time due to reduced time-complexity.**

***Index Terms*—Clustering algorithms, expectation distance, Wasserstein distance, uncertain data.**

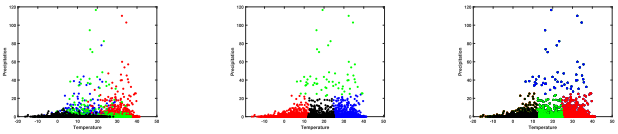
I. Introduction

CLUSTERING, a widely studied unsupervised learning technique, is commonly used in many fields for data analysis to make valuable inferences by observing what group each data point falls into. Classical clustering methods of K-means and K-medoids iteratively group raw-data into "similarity classes" depending on their relative distances or similarities. Classical K-means and K-medoids clusterings aim to group the data points into clusters so that the data's total distance to their assigned cluster centers is minimized [1]. The classical clustering algorithms work with raw-data and are not designed to be robust to uncertain/noisy data. However, data is naturally and inherently affected by the random nature of the physical generation process and measurement inaccuracies, sampling discrepancy, outdated data sources, or other errors, making it prone to noise/uncertainty [2], [3], [4]. As an application, consider a weather station that monitors and measures daily variables like temperature, humidity, and vapor pressure. The data is naturally noisy due to physical measurement equipment (that introduces thermal noise) and variations resulting from other natural sources such as lightning and solar flares. While the daily weather conditions are expected to be within specific predicted ranges for certain seasons, there can be days when those will vary. A tangible illustration can be drawn from weather data obtained from Colorado State University CoAgMET Raw Data Access, as plotted in Fig. 1 for the 4 respective seasons. Their union is shown in Fig. 2(a), which reveals significant overlap across the seasons. Applying classical K-means and K-medoids algorithms for a $4-$way clustering yields only 53.1250% and 53.0574% accuracy, respectively, for grouping weather data in seasons, as depicted in Fig. 2(b)–(c), highlighting the challenge of accurately clustering such noisy and overlapping data. Our analysis of distributions-based clustering reported in the paper shows striking improvement in clustering results, as can be seen in Section IV.



(a) Spring (b) Summer (c) Winter (d) Fall

Fig. 1. Data distribution of each of the four seasons for the Avondale station.



(a) Original clusters (b) K-means 53.1250 Accuracy (c) K-medoids 53.0574 Accuracy

Fig. 2. Data clustering into seasons for the Avondale station.

Clustering uncertain data has been well-recognized as a challenge in the data mining fields [5], having applications in diverse fields such as weather forecasting, medical diagnosis, image

Manuscript received 6 July 2022; revised 20 February 2024; accepted 28 March 2024. Date of publication 16 April 2024; date of current version 27 September 2024. This work was supported in part by U.S. National Science Foundation under Grant NSF-CSSI-2004766 and Grant NSF-PFI-2141084. Recommended for acceptance by L. Chen. *(Corresponding author: Rahmat Adesunkanmi.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50010 USA (e-mail: rahma@iastate.edu; rkumar@iastate.edu).

Digital Object Identifier 10.1109/TKDE.2024.3386401

1041-4347 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

processing, and many more. Addressing the uncertainties in data can significantly improve the accuracy and robustness of clustering results. In the presence of data uncertainty, accounting for noise distribution and its impact on data values toward data clustering is meaningful, and so, one viable way to reduce the impact of uncertainty in data clustering is to extract and utilize its probability distribution whenever feasible. Accordingly, [5], [6], [7], [8] cluster the data-distributions, estimated from datasets belonging to the same random variables, and assign raw-data to its distribution cluster. The motivation for extending classic K-means and K-medoids is to enhance clustering accuracy in the presence of noise. As demonstrated visually with real weather data from the Avondale, Colorado weather station across four seasons (Fig. 1), classical clustering exhibits limited accuracy due to the data overlap caused by random weather variability. Thus, motivating advanced modifications, such as clustering over probability distributions, are aimed at mitigating such limitations and improving clustering accuracy, as exemplified in Fig. 2. By clustering over the probability distributions of the data rather than the raw data, this approach avoids erroneous clustering of the data that appear as outliers due to the noise. Raw-data is assigned to the cluster of its distribution.

A. Our Contributions

This paper studies the clustering of uncertain/noisy data, making the following contributions:

- For clustering data distributions, we propose a new distance measure among distributions, namely Expectation Distance (ED), which extends the widely used 2-Wasserstein (W_2) distance by factoring in the correlation information ignored by W_2. We formally show that ED meets all the required criteria of being a metric.
- The proposed ED measure can also be utilized as a similarity measure in other machine learning applications (not just clustering), such as implementing a Self-organizing map (SOM) that optimizes a distributional measure distance as a dissimilarity measure to compare distributional data [9], learning invariant features in images by way of using a distributional distance to measure the dissimilarity among the feature distributions [10] and for anomaly detection by examining the change in the probability distribution of variables by measuring the dissimilarities between the baseline data and the observed data [11].
- We provide a closed-form formula for the W_2 and ED distances in terms of the means and covariances of the distributions and provide those parameters for Gaussian and lognormal distributions.
- Using the proposed ED distance, we extend the classical clustering techniques of K-means and K-medoids to cluster over the data-distributions. We denote the corresponding K-means and K-medoids as EKM and EKMd, respectively. The corresponding W_2-distance-based versions are termed WKM and WKMd, respectively.
- We show that the Barycenter is the cluster-center in both WKM and EKM. In contrast, we show that the same is not true for WKMd versus EKMd, which generally can have different cluster-center.
- We provide time-complexity for clustering based on raw-data versus the distributions. It is shown that the latter is of lower complexity and yet offers higher accuracy.
- We implement and compare the results of all six clustering algorithms: Classical K-means and K-medoids for raw-data versus W_2-based versus ED-based clustering of data distributions, by applying to real-world noisy weather data with Gaussian characteristic and real-world stocks data having lognormal distribution.

B. Related Works

Recent advancements have introduced variants of clustering methods to enhance their accuracy/robustness. Authors in [12] introduce an adaptive weighting to mitigate the influence of outliers, while researchers in [13], [14], [15] leveraged kernel methods to embed data into higher-dimensional space through nonlinear mapping, potentially facilitating better separation of clusters. Furthermore, FC (fixed-centered)-K-means [16], a method utilizing K-means clustering, weighted fuzzy logical relations, and probabilistic fuzzy sets, address limitations in handling non-spherical or non-convex data distributions. To tackle scalability, [17], [18] employ MapReduce to distribute computations, while [19] harnesses Apache Spark for parallelization for accelerated processing. Additionally, the classical K-means and K-medoids clustering methods have been integrated with various machine learning techniques to improve their performance further. These include integration with dimensionality reduction using feature selection results for improved computational speed [20], combination with Deep Convolutional Neural Networks (DCNN) for feature learning and subsequent clustering [21], and integration with Latent Dirichlet Allocation (LDA) to enhance identifying topics within text documents more effectively [22].

The task of distribution clustering requires distance metrics among the distributions. The Maximum Mean Discrepancy (MMD) distance [23] measures the difference between the mean of the probability distributions, while the Integral Probability Metrics (IPMs) [24] measures the distance based on the integrals of a discrepancy function. MMD is restrictive by being limited to only the mean values, and IPMs are generally computationally expensive. The KL divergence measure misses the important property of symmetry, while the Bhattacharyya distance violates the triangular inequality. Optimal mass transport (OMT) is a commonly used metric that seeks to find the least costly way of transforming one distribution of mass to another relative to a given transport cost [25]. The OMT has been increasingly used in recent years in various applied fields such as economics [26], image processing [27], machine learning [28], data science [29], among others. W_2-distance or 2-Wasserstein distance [30] uses the OMT concept where the cost of transportation is the expectation of the Euclidean distance. W_2-distance has been used in clustering algorithms, such as Wasserstein K-means [31], [32], [33], and also as a Wasserstein auto-encoder [34]. However, the

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

Wasserstein distance only considers the pairwise marginal distribution information and ignores the true correlation information. This paper proposes a new distance metric, the Expectation Distance (ED), that can account for the uncertainty and factor in the correlation information.

Classical clustering algorithms often require complete data. Several techniques have been proposed to handle missing data in clustering: Imputation-based methods [35], [36] fill the missing values with estimated values, while subspace-based methods [37], [38] identify subspaces with no missing values and cluster the data in those subspaces. A probabilistic method in [39] models the distribution of missing values and uses that model to generate the missing data.

C. Clustering Definition

Consider a set, S, that needs to be clustered into K clusters. The clustering problem requires finding a function, $C : S \to [1, K]$, to map elements of S to one of the K clusters in some optimal sense. Then for each $i \in [1, K]$, the ith cluster set under the clustering C is given by,

$$S_C(i) := \{s \in S | C(s) = i\}, \tag{1}$$

with its cluster-center being the minimizer of the distance to the cluster members

$$\bar{s}_C(i) := \arg\left\{\min_{s} \sum_{s' \in S_C(i)} \|s - s'\|_2\right\}, \forall i \in [1, K], \tag{2}$$

where the notation $|.|$ measures the size of its argument set. The cluster-center turns out to be the center-of-mass, also called the Barycenter, of the cluster members

$$\bar{s}_C(i) := \frac{\sum_{s \in S_C(i)} s}{|S_C(i)|}, \forall i \in [1, K]. \tag{3}$$

The goal of clustering is to find an optimal K-cluster that minimizes the aggregate distances of each of the data to their respective cluster-centers, i.e.,

$$\begin{aligned}\min_{C} \sum_{i=1}^{K} \sum_{s' \in S_C(i)} \|\bar{s}_C(i) - s'\|_2 \\ = \min_{C} \sum_{i=1}^{K} \sum_{s' \in S_C(i)} \left\| \frac{\sum_{s \in S_C(i)} s}{|S_C(i)|} - s' \right\|_2.\end{aligned} \tag{4}$$

The corresponding optimal cluster is called K-means.

For K-means, a cluster-center is a Barycenter and may not coincide with any of the data points. If we require the cluster-center be one of the data points, then the resulting clustering is called K-medoids, for which the objective function can be written as

$$\min_{C} \sum_{i=1}^{K} \left\{ \min_{s \in S_C(i)} \left(\sum_{s' \in S_C(i)} \|s - s'\|_2 \right) \right\}. \tag{5}$$

Here, the inner optimization minimizes the distance between one data point in a cluster to all other data points within the same cluster to determine a cluster-center

$$\hat{s}_C(i) := \arg\left\{\min_{s \in S_C(i)} \sum_{s' \in S_C(i)} \|s - s'\|_2\right\}, \forall i \in [1, K]. \tag{6}$$

One popular heuristic to find a locally optimal clustering involves starting with an arbitrary initial clustering, C_0, and iteratively finding a better clustering C_{n+1} from a prior clustering C_n, $(n \geq 0)$, until this process converges, i.e., until $C_{n+1} = C_n$. The heuristic finds the ith cluster of the $(n + 1)$th iteration as the set of those elements that are nearest to the ith cluster-center of the nth iteration. The exact iterative computation for K-means can be used to find K-medoids with the change that the cluster-center is restricted to a data point.

II. Clustering Using Data-Distributions for Noise-Robustness

One approach to extend the K-Means and K-medoids and make them robust to noise-led outliers is to perform clustering over the data-distributions and then assign each raw-data to the cluster of its distribution. This way, the effect of outliers is reduced, making the clustering more robust. Clustering over data-distributions requires measuring distances between distribution pairs. For this, we present a new "Expectation Distance" (ED) and also utilize the commonly used Optimal Mass Transport (OMT) distance, also called W_2 distance, for comparison.

A. Optimal Mass Transport/W_2-Distance

OMT computes the distance between two random variables X and Y having distributions f_X and f_Y, respectively, by associating cost to "transport" the probability mass from the starting distribution f_X to the destination distribution f_Y, while minimizing that cost among all possible transports. Letting $T : \mathbb{R}^n \to \mathbb{R}^n$ denote a transport map, OMT minimizes the associated cost of transport

$$\min_{T} \int_{\mathbb{R}^n} c(x, T(x)) f_X(dx), \tag{7}$$

where $c(\cdot, \cdot)$ is a user-specified cost function of transport. Kantorovich proposed the cost to be Euclidean distance and minimized the transport cost over the joint distributions f_{XY} so that the marginals along the two coordinate directions coincide with f_X and f_Y, respectively, resulting in the 2-Wasserstein or W_2-distance [30]

$$W_2^2(X, Y) := \inf_{\substack{f_{XY}: \\ E_X(Y|X) = f_Y, \\ E_Y(X|Y) = f_X}} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2^2 f_{XY}(x, y) dx dy. \tag{8}$$

1) Formula for W_2: For a random variable X, we let $\mu_X := \mathbb{E}(X)$ denote the mean of X, similarly for another random variable Y, $\mu_Y := \mathbb{E}(Y)$ is its mean, and their covariance is denoted $\Sigma_{XY} := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T]$. The variances of X and Y are denoted $\Sigma_X := \Sigma_{XX}$ and $\Sigma_Y := \Sigma_{YY}$ respectively. To compute $W_2(X, Y)$, consider the term in (8) that needs to be

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

minimized

$$\begin{aligned}&\int_{\mathbb{R}^n\times\mathbb{R}^n}\|x-y\|_2^2\, f_{XY}(x,y)dxdy\\&=\mathbb{E}[\|X-Y\|_2^2]\\&=\mathbb{E}[\|(X-\mu_X+\mu_X)-(Y-\mu_Y+\mu_Y)\|_2^2]\\&=\mathbb{E}[\|(X-\mu_X)-(Y-\mu_Y)\|_2^2]+\|\mu_X-\mu_Y\|_2^2\\&=\text{trace}(\Sigma_X+\Sigma_Y-2\Sigma_{XY})+\|\mu_X-\mu_Y\|_2^2. \qquad (9)\end{aligned}$$

Note that $\mathbb{E}[\|X-Y\|_2^2]$ only depends on the first two moments—This is because the 2-norm is used for measuring the distance. If instead p-norm, $p>2$, is used, higher-order moments will be required.

For computing $W_2(X,Y)$, we need to minimize (9) with respect to those joint distributions f_{XY} that possess the marginals f_X and f_Y. Fixing the marginals f_X and f_Y fixes $\mu_X,\mu_Y,\Sigma_X,\Sigma_Y$, leaving Σ_{XY} to be the only variable of optimization. Since (9) is a decreasing function of Σ_{XY}, it is then obvious that the minimization will be achieved when Σ_{XY} is the largest, i.e., X and Y are the most correlated. Mathematically, we need to solve the following semidefinite program:

$$\begin{aligned}&\min_{\Sigma_{XY}}\left[\text{trace}(\Sigma_X+\Sigma_Y-2\Sigma_{XY})+\|\mu_X-\mu_Y\|_2^2\right]\\&\text{s.t.}\begin{bmatrix}\Sigma_X & \Sigma_{XY}\\ \Sigma_{XY}^T & \Sigma_Y\end{bmatrix}\geq 0. \qquad (10)\end{aligned}$$

The minimum in (10) is achieved at

$$\Sigma_{XY}=\left(\Sigma_X^{1/2}\Sigma_Y\Sigma_X^{1/2}\right)^{1/2}. \qquad (11)$$

Thus, the W_2 distance has the closed-form formula

$$\begin{aligned}W_2^2(X,Y)=&\ \|\mu_X-\mu_Y\|_2^2\\&+\ \text{trace}\left[\Sigma_X+\Sigma_Y-2\left(\Sigma_X^{\frac{1}{2}}\Sigma_Y\Sigma_X^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]. \qquad (12)\end{aligned}$$

Several numerical methods have been developed to compute the Wasserstein distance efficiently, such as the Sinkhorn algorithm [40] and the Entropic Regularization of Optimal Transport (EROT) [41].

2) Cluster-Center Under W_2: The cluster-center for a cluster set S of distributions in the case of W_2-based K-means, denoted WKM, is given by

$$\arg\min_{X'}\sum_{X\in S}W_2^2(X',X). \qquad (13)$$

The cluster-center turns out to be the Barycenter [31]

$$\frac{1}{|S|}\sum_{X\in S}X. \qquad (14)$$

In contrast, in the case of the W_2-based K-medoids, denoted WKMd, a cluster-center is restricted to be chosen from one of the data points and may differ from the Barycenter

$$\arg\min_{X'\in S}\sum_{X\in S}W_2^2(X',X). \qquad (15)$$

For distributions $\{X_i, 1\leq i\leq n\}$, with $\mathbb{E}(X_i)=\mu_i$, $Var(X_i)=\Sigma_i$, the Barycenter distribution's mean μ and covariance Σ are given by

$$\mu=\frac{1}{n}\sum_{i=1}^n\mu_i,\ \text{and}\ \Sigma=\frac{1}{n}\sum_{i=1}^n(\Sigma^{\frac{1}{2}}\Sigma_i\Sigma^{\frac{1}{2}})^{\frac{1}{2}}. \qquad (16)$$

(16) provides Σ in an implicit form, and its computation is a fixed point of the following iteration [42]:

$$\Sigma_{n+1}=\Sigma_n^{-\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^N(\Sigma_n^{\frac{1}{2}}\Sigma_i\Sigma_n^{\frac{1}{2}})^{\frac{1}{2}}\right)^2\Sigma_n^{-\frac{1}{2}}. \qquad (17)$$

B. Expectation Distance

While W_2-based distance measure is popular, it ignores the true correlation information: The minimization in (10) is achieved when the two given marginals are most correlated, which may not be the case. Recognizing this limitation of W_2 distance, we hereby propose a new and more general distance measure between any two probability distributions that also accounts for their joint distributions (and not just their marginals); it is simply the expectation distance (ED) of the given random variables X and Y

$$d_{X,Y}^2:=E[\|X-Y\|_2^2]=\int_{\mathbb{R}^n\times\mathbb{R}^n}\|x-y\|_2^2\, f_{XY}(x,y)dxdy. \qquad (18)$$

The following result establishes that the above definition provides a metric over the distributions.

Theorem 1: $d_{X,Y}=[E[\|X-Y\|_2^2]]^{\frac{1}{2}}$ in definition (18) meets all the required criteria of being a distance measure (namely, positivity, symmetry, zero if and only if equal, and triangular inequality).

Proof: The proof below takes into consideration some common properties of expected value and the fact that $\|x-y\|$ is a metric in itself (and satisfies the said four properties).

Positivity ($d_{X,Y}\geq 0$):
If a random variable p is non-negative, its expected value is also non-negative, i.e., $p\geq 0\iff E(p)\geq 0$. Then since $p:=\|x-y\|_2^2\geq 0$, it holds that $d_{X,Y}^2\geq 0\iff d_{X,Y}\geq 0$ ($d_{X,Y}$ being the positive square root of $d_{X,Y}^2$).

Symmetry ($d_{X,Y}=d_{Y,X}$):
We have $p_{x,y}:=\|x-y\|_2^2=\|y-x\|_2^2=:p_{y,x}$. Then $p_{x,y}=p_{y,x}\Rightarrow E[p_{x,y}]=E[p_{y,x}]\Rightarrow d_{X,Y}^2=d_{Y,X}^2\Rightarrow d_{X,Y}=d_{Y,X}$.

Zero iff equal ($d_{X,Y}=0\Rightarrow X=Y$):
The non-degeneracy property of an expected value asserts that $p=0\Leftrightarrow E[p]=0$ for equivalence classes of almost surely equal variables. Then $X=Y\Leftrightarrow X-Y=0\Leftrightarrow p:=\|x-y\|_2^2=0\Leftrightarrow E[p]=0\Leftrightarrow d_{X,Y}^2=0\Leftrightarrow d_{X,Y}=0$.

Triangle Inequality ($d_{X,Z}\leq d_{X,Y}+d_{Y,Z}$):
By Minkowski inequality in $L^{\mathbb{P}}$ spaces:
$d_{X,Z}=[E[\|X-Z\|_2^2]]^{\frac{1}{2}}$
$\leq[E[\|X-Y\|_2+\|Y-Z\|_2]^2]^{\frac{1}{2}}$
$\leq[E[\|X-Y\|_2]^2]^{\frac{1}{2}}+[E[\|Y-Z\|_2]^2]^{\frac{1}{2}}$
$=d_{X,Y}+d_{Y,Z}$.

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

These properties conclude that the ED proposed in (18) is a distance measure over distributions. □

1) Formula for ED: Given random variables X, Y, it follows from the definition (18) and equality (9) that their ED is given by

$$d^2_{X,Y} = \mathbb{E}\|X - Y\|_2^2 = \text{trace}(\Sigma_X + \Sigma_Y - 2\Sigma_{XY}) + \|\mu_X - \mu_Y\|_2^2. \quad (19)$$

It can be noted that whenever the correlation Σ_{XY} of the two random variables is the same as the one given in (11), the ED distance coincides with the W_2 distance. However, in general, ED is of a higher value: To attain the minimization of (10), which is a decreasing function of Σ_{XY}, its largest possible value (i.e., most correlated) gets picked, but in reality, Σ_{XY} may be smaller (i.e., less correlated), leading to ED being larger than W_2.

Remark 1: From (19), it is clear that ED depends on the 2^{nd} moment of the joint distribution but not on the higher moments. This is because the ED uses the 2-norm to define the distribution distance. It is possible to generalize ED to utilize the p-norm ($p \geq 1$) so that it also depends on the higher order moments

$$\hat{d}^p_{X,Y} := E[\|X - Y\|_p^p] = \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_p^p \, f_{XY}(x, y) dx dy. \quad (20)$$

2) Cluster-Center Under ED: Theorem 2: The cluster-center for EKM (ED-based K-means) is the Barycenter of the cluster-set (as in the case of W_2-based K-means).

Proof: Under the ED measure, the cluster-center in case of K-means for a cluster-set $S = \{X_1, \ldots, X_{|S|}\}$ of distributions is given by the following expression that optimizes the total distance between a candidate cluster-center distribution and each of the distributions in S, with respect to all possible choices for the cluster-center candidate distribution X, along with all possible choices for the joint distribution candidates between the candidate cluster-center and the elements of the cluster, $\{f_{X'X_i}, 1 \leq i \leq |S|\}$ with $\forall i : E_{X'}(X_i|X') = X_i, E_{X_i}(X'|X_i) = X'$

$$\arg \min_{X'} \inf_{\substack{f_{X'X_i}, 1 \leq i \leq |S| : \\ E_{X'}(X_i|X') = X_i, \\ E_{X_i}(X'|X_i) = X'}} \sum_{1 \leq i \leq |S|} \mathbb{E}\|X' - X_i\|_2^2. \quad (21)$$

Since the joint distributions ($f_{X'X_i}$ versus $f_{X'X_j}, 1 \leq i \neq j \leq |S|$) can be chosen independent of each other, and the minimization is of the sum of positive entries, the operations in (21) can be rearranged to obtain

$$\arg \min_{X'} \sum_{1 \leq i \leq |S|} \inf_{\substack{f_{X'X_i}, 1 \leq i \leq |S| : \\ E_{X'}(X_i|X') = X_i, \\ E_{X_i}(X'|X_i) = X'}} \mathbb{E}\|X' - X_i\|_2^2,$$

$$= \arg \min_{X'} \sum_{1 \leq i \leq |S|} W_2^2(X', X_i),$$

where the last equality follows from the definition of W_2-distance. It can then be seen that the last expression is the same as that of W_2-based distance in (13), and hence, the resulting cluster-center in the case of EKM is again the Barycenter of (14). □

In the case of K-medoids using ED distance, denoted EKMd, a cluster-center is chosen to be one of the data points, so their joint distribution, as already estimated from the dataset, is known and used to perform the optimization

$$\arg \min_{X' \in S} \sum_{X \in S} \mathbb{E}\|X' - X\|_2^2. \quad (22)$$

C. Covariances With Cluster-Centers

In the case of WKMd or EKMd, the cluster-center is one of the data points, so its joint distribution with any other data point, and hence the corresponding covariance, is already known. However, in the case of WKM or EKM, a cluster-center is the Barycenter of the cluster-set. We can compute its covariance with the other data points within its cluster-set, say, $\{X_1, X_2, \ldots, X_n\}$ as follows. The covariance between a data point X_i and a cluster-center $X = \frac{X_1 + X_2 + \ldots + X_n}{n}$ is

$$\Sigma_{X_i X} = Cov\left(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n}\sum_{j=1}^{n} \Sigma_{X_i X_j}. \quad (23)$$

Eq. (23) provides the closed-form expression to compute the covariance of the joint distribution between a data point and its cluster center in the case of WKM or EKM. In the case of EKM, since the pairwise joint distributions appearing in (23) are already known and fixed, those pairwise covariances are also known and fixed, and so for the case of EKM, (23) provides the final answer. However, in the case of WKM, (11) provides the optimum covariance between a pair of distributions, and hence, for the case of WKM, the covariance between a cluster-element and its cluster-center is given by

$$\Sigma_{X_i X} = \frac{1}{n}\sum_{j=1}^{n} \left(\Sigma_{X_i}^{1/2} \Sigma_{X_j} \Sigma_{X_i}^{1/2}\right)^{1/2}.$$

D. W_2 and ED for Lognormals

From the definition of the lognormal distribution, it is known that if $X \sim lognorm(\theta_X, \Delta_X)$ and $Y \sim lognorm(\theta_Y, \Delta_Y)$ are multivariate lognormal random variables with parameters (θ_X, Δ_X) and (θ_Y, Δ_Y), respectively, then $A = [A(i) := \ln(X(i))]_{n \times 1}$ and $B = [B(i) := \ln(Y(i))]_{n \times 1}$, are multivariate normal random variables with $A \sim \mathcal{N}(\theta_X, \Delta_X), B \sim \mathcal{N}(\theta_Y, \Delta_Y)$. Then, the means and covariances of the two lognormal random variables are as given

$$\mu_X = \mathbb{E}[X] = [\exp(\theta_X(i) + 0.5\Delta_X(ii))]_{n \times 1}$$

$$\mu_Y = \mathbb{E}[Y] = [\exp(\theta_Y(i) + 0.5\Delta_Y(ii))]_{n \times 1}$$

$$\mathbb{E}[XY^T] = \mathbb{E}[YX^T] = [\exp(\theta_X(i) + \theta_Y(j) + 0.5(\Delta_X(ii) + \Delta_Y(jj) + 2\Delta_{XY}(ij)))]_{n \times n}$$

$$\Sigma_{XY} = \mathbb{E}(XY^T) - \mathbb{E}(X)\mathbb{E}(Y)^T$$

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Distributional K-Means Using W_2 (WKM).

Require: N distributions, $f_i \sim f_1, f_2, \ldots, f_N$

1: Choose K initial cluster-centers $f_{c_1}, f_{c_2}, \ldots, f_{c_k}$ from the given set of N distribution data.
2: **for** $i = 1$ to N(=total number of distributions) **do**
3: Solve
$$k_i = \arg\left\{\min_{1\le k\le K} W_2^2(f_i, f_{c_k})\right\}$$
4: Assign f_i to cluster k_i
5: **end for**
`This creates a disjoint partition of the data into subsets` $f_1, f_2, \ldots, f_K$.
6: **for** $k = 1$ to K(=total number of clusters) **do**
7: Update center $f_{c_k} = Barycenter(f_k)$
8: **end for**
`Repeat steps 2 to 8 using new` c_k`'s until convergence.`
9: Group data points using the final distribution groups in $f_1, f_2, \ldots, f_K$

Algorithm 2: Distributional K-Means Using ED (EKM).

Require: N distributions, $f_i \sim f_1, f_2, \ldots, f_N$
Steps same as Algorithm 1, with the following changed: step 3:
3: $$k_i = \arg\left\{\min_{1\le k\le K} d^2_{f_i, f_{c_k}}\right\}.$$

$$= [(\exp(\theta_X(i) + .5\Delta_X(ii)) \times ((\exp(\theta_Y(j) + .5\Delta_Y(jj)) \times (\exp(\Delta_{XY}(ij)) - 1)]_{n\times n}$$

$$\Sigma_X := \Sigma_{XX}; \quad \Sigma_Y := \Sigma_{YY}.$$

To compute the W_2 and ED measures for lognormal distributions, we can plug the above parameters for the lognormals into the formulas for W_2 (12) and ED (19) respectively, to get the two respective distances.

III. W_2- & ED-Based Distribution Clustering

Here, we extend the classical K-means and K-medoids-based clustering methods to clustering over the data-distributions (as opposed to raw-data). The distance measures considered in clustering over the data-distributions are the above-mentioned W_2 and ED distances. The corresponding WKM and EKM clustering algorithms are presented in Algorithms 1 and 2, respectively, and the corresponding WKMd and EKMd clustering algorithms are presented in Algorithms 3 and 4 respectively. Each algorithm starts with an initial guess of K cluster-centers, iteratively assigns data to its nearest cluster-center, then recomputes the cluster-centers and repeats until convergence.

A. Computational Complexity and Scalability

In general, the computational complexity of K-means clustering is $\mathcal{O}(nNKT)$ and that of K-medoids $\mathcal{O}(nN^2KT)$, where n is the data dimension, N is the number of data

Algorithm 3: Distributional K-Medoids Using W_2 (WKMd).

Require: N distributions, $f_i \sim f_1, f_2, \ldots, f_N$
Steps same as Algorithm 1, with the following, changed: step 7:
7: Update medoid
$$f_{c_k} = \arg\left\{\min_{f\in f_k} \sum_{f'\in f_k} W_2^2(f, f')\right\}$$

Algorithm 4: Distributional K-Medoids Using ED (EKMd).

Require: N distributions, $f_i \sim f_1, f_2, \ldots, f_N$
Steps same as Algorithm 1, with the following, changed: steps 3 and 7:
3: $$k_i = \arg\left\{\min_{1\le k\le K} d^2_{f_i, f_{c_k}}\right\}$$
7: $$f_{c_k} = \arg\left\{\min_{f\in f_k} \sum_{f'\in f_k} d^2_{f,f'}\right\}$$

elements to be clustered, K is the number of clusters, and T is the number of iterations employed. (T in the worst case can be exponential, leading to worst-case complexity of $n^{\mathcal{O}(nK)}$ [43].) Additionally, there is $O(nN^2)$ complexity of finding pairwise distances. In the case of distributional clustering, there is the added task of estimating the distribution parameters, whose complexity is $\mathcal{O}(nm^2\, M)$, where m is the number of data points per distribution, and M is the number of distributions (implying a total of $N = mM$ data points). Thus the complexities of K-means and K-medoids for raw-data clustering are $\mathcal{O}(nmMKT) + \mathcal{O}(nm^2\, M^2)$ and $\mathcal{O}(nm^2\, M^2KT) + \mathcal{O}(nm^2\, M^2)$ respectively, and those for distributional clustering are $\mathcal{O}(nMKT) + \mathcal{O}(n^2\, M) + \mathcal{O}(nM^2)$ and $\mathcal{O}(nM^2KT) + \mathcal{O}(nm^2\, M) + \mathcal{O}(nM^2)$ respectively. These quadratic computational complexities suggest their scalability. Also, since $\mathcal{O}(nMKT) + \mathcal{O}(nm^2\, M) + \mathcal{O}(nM^2) < \mathcal{O}(nmMKT) + \mathcal{O}(nm^2\, M^2)$ and similarly since $\mathcal{O}(nM^2KT) + \mathcal{O}(nm^2\, M) + \mathcal{O}(nM^2) < \mathcal{O}(nm^2\, M^2KT) + \mathcal{O}(nm^2\, M^2)$, it follows that the distributional clustering has a lower time-complexity for both K-means and K-medoids compared to the raw-data clustering. Nevertheless, we show below that the accuracy of distributional clustering is higher than that of raw-data clustering.

IV. Results and Discussion

We implemented the distributional clustering algorithms to cluster synthetic and real-world noisy data—weather and stock data. First, the data is cleaned by removing unwanted attributes and ensuring an equal number of remaining attributes with no missing attribute values. We then extract the underlying distributions by estimating from data from the same random variable its distributions parameters—means and covariances

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

for Gaussians (weather data) and the means of covariances of the natural logarithms for the lognormals (stock data). For the case of the real-world weather data, we treat each season of each year to be a Gaussian distribution, and accordingly, we have 4 distributions per year. For the case of the real-life stocks data, we model each stock to be a lognormal distribution, considering 77 total stocks picked from the Nasdaq top-100 for the years 2018-19. The performance of the six different clustering algorithms—the classical versions of K-means and K-medoids and their W_2 and ED-based extensions, namely, WKM, WKMd, EKM, EKMd—are compared using the measures of Accuracy, NMI, and ARI as described next. All simulations were executed using Matlab R2022 A on a Windows 10 operating system. The in-built K-medoids and K-means functions were used to run KM and KMd. Matlab's K-medoids function accepts user-defined distance functions for WKMd and EKMd, whereas we implemented WKM and EKM from scratch. The running times of our algorithms have been reported in tables for respective datasets presented in the following sections.

A. Performance Metrics

The accuracy of the six clustering techniques: classical K-means (KM), W_2 K-means (WKM), ED K-means (EKM), classical K-medoids (KMd), W_2 K-medoids (WKMd), and ED K-medoids (EKMd), are compared using the following defined three commonly used performance metrics of Accuracy, NMI (normalized mutual information), and ARI (adjusted rand index). They all assume the existence of the ground truth clustering, denoted C^*, to compare against the computed clustering, C, and compute a normalized score within the unit interval, with 1 being the maximum accuracy score. Given S, a set of N data points, and its two K-sized cluster partitions, the computed one C and the ground truth C^*

$$C = \{S_C(1), \ldots, S_C(K)\};\ C^* = \{S_{C^*}(1), \ldots, S_{C^*}(K^*)\},$$

define $n_{ij} := |S_C(i) \cap S_{C^*}(j)|$, $a_i := \sum_{j=1}^{K^*} n_{ij}$, $b_j = \sum_{i=1}^{K} n_{ij}$. Note n_{ij} denotes the number of data points common between the clusters $X_C(i)$ and $X_{C^*}(j)$.

1) *accuracy* is simply the ratio of the correctly clustered data points to the total number of data points

$$\textbf{Accuracy} = \frac{1}{N} \sum_{i=1}^{\min\{K,K^*\}} n_{ii}.$$

2) *Normalized Mutual Information (NMI)* [44] The mutual information $I(C^*;C)$ between the two clusterings is used to compute the two normalized indices $\frac{I(C^*;C)}{H(C^*)}$ and $\frac{I(C^*;C)}{H(C)}$, respectively, whose harmonic mean gives the desired index

$$\textbf{NMI} = 2\frac{I(C^*;C)}{H(C^*)+H(C)};$$

$$I(C^*;C) = H(C^*) - H(C^*|C)$$

$$= -\sum_{c^* \in C^*} p_{C^*}(c^*) \log_2 p_{C^*}(c^*)$$

$$- \sum_{c \in C} p_C(c) H(C^*|C=c)$$

$$= -\sum_{j=1}^{K^*} \frac{b_j}{n} \log_2 \frac{b_j}{n} - \sum_{i=1}^{K} \frac{a_i}{n} \sum_{j=1}^{K^*} \frac{n_{ij}}{a_i} \log_2 \frac{n_{ij}}{a_i}.$$

3) *Adjusted Rand Index (ARI)* [45] The Rand Index (RI) computes a similarity measure between two clustering by counting samples in all pairs of cells taken from the two clusters. The ARI score is then the adjusted version of RI, "corrected-for-chance," and normalized

$$\textbf{ARI} = \frac{\text{RI} - \text{Expected(RI)}}{\text{Max(RI)} - \text{Expected(RI)}}$$

$$= \frac{\sum_{ij} \binom{n_{ij}}{2} - \left(\frac{\sum_{i=1}^{K} \binom{a_i}{2} \sum_{j}^{K^*} \binom{b_j}{2}}{\binom{n}{2}} \right)}{\frac{1}{2}\left(\sum_i^K \binom{a_i}{2} + \sum_j^{K^*} \binom{b_j}{2} \right) - \left(\frac{\sum_i^K \binom{a_i}{2} \sum_j^{K^*} \binom{b_j}{2}}{\binom{n}{2}} \right)}.$$

B. Synthetic Gaussian Data With Unbalanced Clusters

Clusters are termed unbalanced if their sizes are disparate. Traditional clustering methods face difficulty handling unbalanced data effectively [46]. To demonstrate the robustness of our algorithm in clustering unbalanced data with noise, we generated a synthetic dataset comprising 3,000 2D-samples from 3 groups of normal distributions of sizes $n_1 = 100, n_2 = n_3 = 25$, totaling 150, by generating 20 2D-samples from each of the 150 2D-distributions. To start, we created three random matrices, $C_i, i = 1, 2, 3$, of sizes $2n_i \times 2n_i$, with each entry in C_i being a random number uniformly distributed between 0 and 1. The covariance matrices for the three groups of random variables were generated using $C_i, i = 1, 2, 3$ as follows:

$$\Sigma_i := \begin{bmatrix} \frac{C_i(1,1)}{sc(i)} & \ldots & C_i(2n_i,1) \\ \vdots & \ddots & \vdots \\ C_i(2n_i,1) & \ldots & \frac{C_i(2n_i,2n_i)}{sc_i} \end{bmatrix} \times \begin{bmatrix} \frac{C_i(1,1)}{co(i)} & \ldots & C_i(2n_i,1) \\ \vdots & \ddots & \vdots \\ C_i(2n_i,1) & \ldots & \frac{C_i(2n_i,2n_i}{sc_i} \end{bmatrix}^\top.$$

Essentially, three symmetric square matrices $\Sigma_i, i = 1, 2, 3$ were created where for each $i = 1, 2, 3$ and $k, j \le n_i$, $\Sigma_i(2 \times k - 1 : 2 \times k, 2 \times j - 1 : 2 \times j)$ represents the covariance of the kth and jth random-variables of the ith group. Similarly, the means μ_i of sizes $(2, n_i), i = 1, 2, 3$ were generated as random numbers between 0 and 1. The Matlab command $mvnrnd(\mu_i, \Sigma_i), i = 1, 2, 3$ was next used to generate data, 20 per group, yielding a total of $2n_1 \times 20 + 2n_2 \times 20 + 2n_3 \times 20 = 2 \times 150 \times 20 = 2 \times 3000$ samples. The covariance among the pairs of random variables from different groups was then numerically computed using these samples.

The synthetic 3000 2D-samples were next clustered into three groups using classical versus W_2-based versus ED-based

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

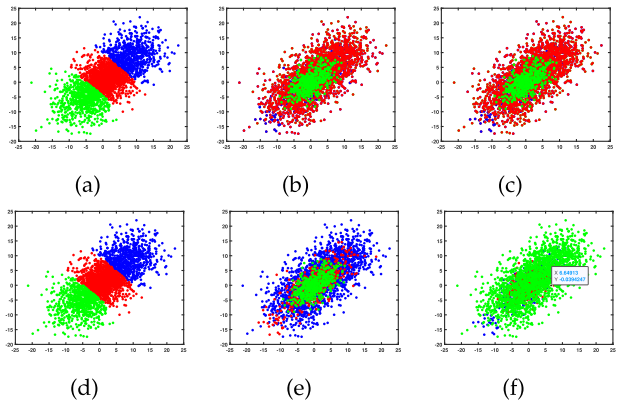


Fig. 3. Synthetic Gaussian data result: Plots of data clusters (dots) using (a) KM, (b) WKM, (c) EKM, (d) KMd, (e) WKMd, (f) EKMd.

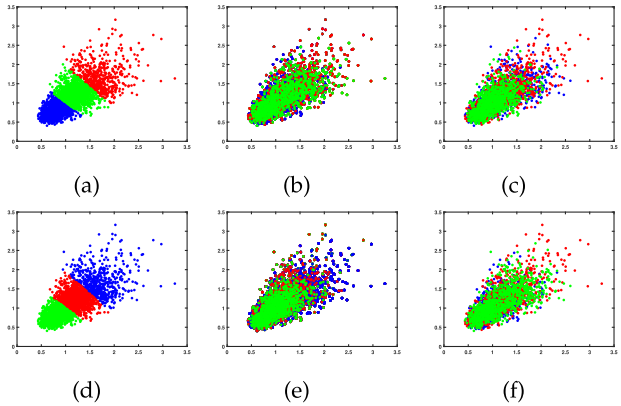


Fig. 4. Synthetic data result: Plots of data clusters (dots) using (a) KM, (b) WKM, (c) EKM, (d) KMd, (e) WKMd, (f) EKMd.

TABLE I
TABLE OF EVALUATION MEASURES FOR SYNTHETIC GAUSSIAN DATA

	Accuracy	NMI	ARI	Compute time(s)	Centers
KM	0.3847	0.0656	0.4936	0.1691	
WKM	0.9800	0.9130	0.9675	0.6324	0.6932 0.5635 1.8465 2.0326 0.0072 0.0395
EKM	1.0000	1.0000	1.0000	0.3187	0.6475 0.5510 1.8234 1.9916 0.0072 0.0395
KMd	0.3837	0.0659	0.4933	0.2437	
WKMd	0.9667	0.8751	0.9466	0.5925	(28, 108, 140)
EKMd	1.0000	1.0000	1.0000	0.2214	(124, 136, 70)

TABLE II
TABLE OF EVALUATION MEASURES FOR SYNTHETIC LOGNORMAL DATA

	Accuracy	NMI	ARI	Compute time(s)	Centers
KM	0.3843	0.0234	0.5521	0.2048	
WKM	0.8467	0.6407	0.8388	1.1201	1.0303 1.0365 1.1985 1.2310 1.1720 1.1496
EKM	1.0000	1.0000	1.0000	0.4710	1.0270 1.0355 1.1708 1.1546 1.1985 1.2310
KMd	0.3860	0.0242	0.5536	0.3124	
WKMd	0.7200	0.5241	0.7699	0.6424	(150, 19, 92)
EKMd	1.0000	1.0000	1.0000	0.2788	(87, 127, 43)

methods. The clustering results from six algorithms are visually depicted in Fig. 3 and evaluated in Table I. Classical K-means (KM) and K-medoids (KMd) failed to produce accurate clusters (accuracy of 38.47% for KM and 38.37% for KMd). In contrast, the accuracy of W_2-based methods were 98% and 96.67%, while the ED-based method yielded an accuracy of 100%, showcasing the robustness of ED-based clustering for correlated unbalanced clusters.

C. Synthetic Lognormal Data

To demonstrate the algorithm's versatility across various distributions, we generated synthetic lognormal distributions using parameters constructed similarly to the Gaussian distributions described in Section IV-B. In this setup, the randomly generated data entries in $C_i, i = 1, 2, 3$ ranged from 0 to 0.06, and the three groups were of balanced sizes with $n_1 = n_2 = n_3 = 150$. The lognormal parameters were estimated following the procedure outlined in Section II-D. Generating 20 samples per distribution, we obtained 3000 2D-samples. These were then clustered into $K = 3$ groups. The results obtained from the six clustering algorithms are visually shown in Fig. 4, while the performance is reported in Table II, along with the run-times. The accuracies of KM, WKM, and EKM were 0.3843, 0.8467, and 1.0000 respectively, whereas the accuracies of the corresponding K-medoids versions were 0.3860, 0.7200, and 1.0000 respectively. The NMIs were: 0.0234, 0.6407, 1.0000 and 0.0242, 0.5241, 1.0000, whereas the ARIs were: 0.5521, 0.8388, 1.0000 and 0.5536, 0.7699, 1.0000. Like the synthetic Gaussian distribution clustering, it follows that the distribution-based clustering also outperforms the classical ones in the case of synthetic lognormal data.

TABLE III
WEATHER DATA DETAILS

	Station	Number of years obtained	Data Length
A	Avondale	22	7392
B	Ault	19	6384
C	Dove Creek	21	7056
D	Fort Collins	23	7728
E	Kirk	20	6720
Total		105	35280

D. Real-World Weather Data

To demonstrate the performance of the explained clustering algorithms and show that the distribution-based algorithms work better in the case of uncertain data, we applied them to real-life weather data sourced from Colorado State University CoAgMET Raw Data Access, that were collected from five weather stations listed in Table III. For an even computation of data-distributions, we kept only 28 entries from each month, and accordingly, the data lengths, based on the number of years, are as listed in Table III.

Data in the same meteorological seasons in the USA were considered to be in the same cluster: Spring: 03/01 - 05-31; Summer: 06/01 - 08/31; Fall: 09/01 - 11/30; Winter: 12/01- 02/28. Thus, we have 4 clusters in total, and the ground truth cluster for each measurement was acquired based on the date and the

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

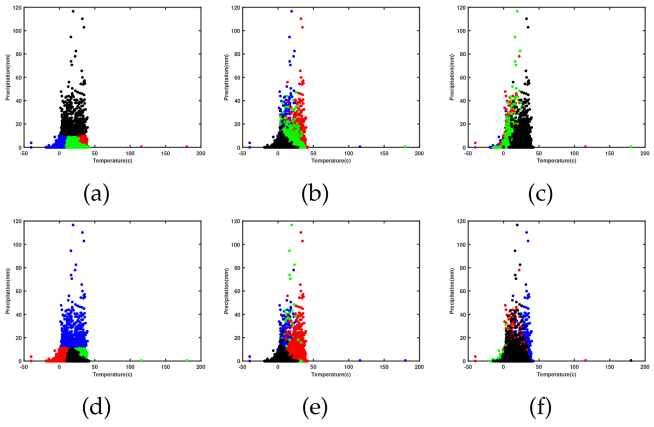


Fig. 5. 3-D Weather data result: Plots of data clusters (dots) using (a) KM, (b) WKM, (c) EKM, (d) KMd, (e) WKMd, (f) EKMd.

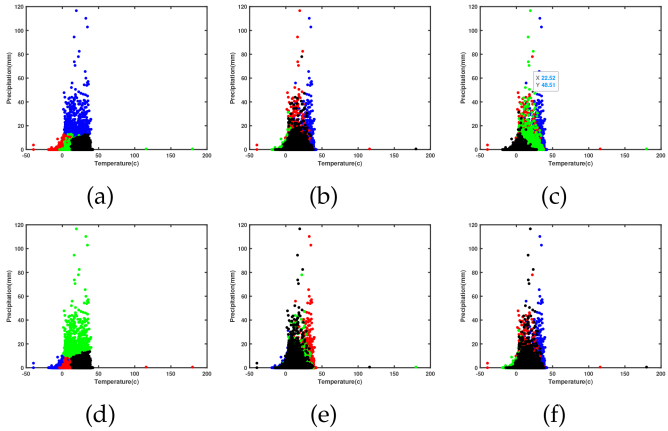


Fig. 6. 7-d weather data result: Plots of data clusters (dots) using (a) KM, (b) WKM, (c) EKM, (d) KMd, (e) WKMd, (f) EKMd.

TABLE IV
EVALUATION MEASURES FOR 3-D WEATHER DATA

	Accuracy	NMI	ARI	Compute Time (s)
KM	0.5649	0.3148	0.7052	8.3068
WKM	0.8429	0.7755	0.8922	7.6164
EKM	1.0000	1.0000	1.0000	6.9747
KMd	0.5637	0.3141	0.7047	5.3981
WKMd	0.8500	0.7799	0.8950	9.9327
EKMd	0.9976	0.9903	0.9976	5.0629

TABLE V
EVALUATION MEASURES FOR 7-D WEATHER DATA

	Accuracy	NMI	ARI	Compute time(s)
KM	0.5596	0.2941	0.7014	12.6322
WKM	0.8595	0.7778	0.8977	9.6157
EDKM	1.0000	1.0000	1.0000	8.4273
KMd	0.5578	0.2911	0.7005	5.7702
WKMd	0.8548	0.7745	0.8956	12.6525
EKMd	1.0000	1.0000	1.0000	5.7022

corresponding season for each entry. For each weather station, the data for the same year's season, with each season containing 84 days (28 days/month for 3 months in a season) worth of data, was treated as a random variable, thereby producing $4\times$ #Station $\times$ #Years $= 4 \times 5 \times 21 = 420$ of total random variables to be clustered into 4 seasons, with each random variable supported by 84 days of data. Thus, there were a total of $420 \times 84 = 35820$ data points. It was reasonably assumed that the weather within a season of a year at a location follows Gaussian distribution. This is shown in Fig. 1 where the data-distribution of each of the four seasons for the Avondale station for temperature and precipitation are plotted. The outliers can also be visualized.

1) 3-Dimensional Weather Data Analysis: In this part of the study, we analyzed 3-dimensional weather data consisting of three features: *maximum daily temperature (°C), precipitation (mm), and vapor pressure (kPa)*, extracted from the weather data from five stations: Avondale, Ault, Dove Creek, Fort Collins, and Kirk, between 1992 and 2021. As noted above, there are 420 random variables, with each random variable having 84 days/season of 3D-data, implying a data set of $35,280 \times 3$ entries. In the preprocessing stage, we computed 420 numbers of 3D means and 3×3 variances, and 420^2 3×3 covariances, which took 1.9194 sec.

We used distributional clustering algorithms to analyze the data, and the performances and compute times of the clustering algorithms are shown in Fig. 5 and Table IV. The accuracies of KM, WKM, and EKM were 0.5649, 0.8429, and 1.0000 respectively, whereas the accuracies of the corresponding K-medoids versions were 0.5637, 0.8500, and 0.9976 respectively. The NMIs were: 0.3148, 0.7755, 1.0000 and 0.3141, 0.7799, 0.9903, whereas the ARIs were: 0.7052, 0.8922, 1.0000 and 0.7047, 0.8950, 0.9976.

The results show that the distribution-based K-means and K-medoids algorithms outperform the corresponding classical versions. The ED-based distance measurement offers higher accuracy, NMI, and ARI values over the W_2-based ones, which performs better than the classical ones. These indicate that the algorithms are more robust to real-life noisy weather data, and the overall *computational time of distributional clustering is less compared to the classical ones that operate on raw-data, yet the accuracy of our algorithms remains higher.*

In summary, this study provides an efficient and accurate approach to analyzing three-dimensional weather data, which can be useful in various weather-related applications.

2) Clustering of 7-Dimensional Weather Data and Comparison of Computation Times: To compare the time complexity as the data size grows, we increased the measurement parameters of daily weather data from 3 to 7 to include the following features: *mean, maximum, and minimum temperature (° C), vapor pressure (kPa), maximum and minimum relative humidity (Fraction), and precipitation (mm).* The resulting data entries are larger compared to our previous experiment, having increased 7-dimensions for the means, variances, and covariances. During preprocessing, computing the means, variance, and covariance for the data of 7-dimensions took 2.1818 seconds; that is only a 13% increase, although the data size has increased by 133%.

The clustering results for 7-dimensional weather data with their compute-time are depicted in Fig. 6 and Table V. The

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

summary results show that distribution-based K-means and K-medoids under both distance measures significantly outperform the corresponding classical versions. The accuracies of KM, WKM, and EKM are 0.5596, 0.8595, and 1.0000, respectively, while the accuracies of the corresponding K-medoids versions are 0.5578, 0.8548, and 1.0000, respectively. The corresponding NMIs are 0.7014, 0.7778, 1.0000, 0.2911, 0.7745, and 1.0000, and the corresponding ARIs are 0.6836, 0.8977, 1.0000, 0.7005, 0.8956, 1.0000. The performance progression (classical $< W_2 <$ ED) is consistent across all clustering methods. The ED-based clustering results offer higher accuracy, NMI, and ARI values than the W_2 based ones, plus the compute-time is smaller. As shown in Table V, the overall computational time of our algorithm is less compared to the classical ones that operate on raw data as well as the W_2-based ones, yet, the accuracy of our algorithms remains higher. This significant performance gain results from the use of distributional clustering, where ED outperforms W_2, and both outperform the classical methods in terms of accuracy and computation time.

E. Real-World Stocks Data

To demonstrate the effectiveness of our clustering algorithms for also the non-Gaussian distributions, we applied our methods to the stocks market price data, which are commonly modeled as lognormal distributions [47]: Letting X_t denote the price of a stock X on day t, the data X_t/X_{t-1} is assumed to follow a lognormal distribution with fixed parameters (that are the mean and variance of $\ln(X_t/X_{t-1})$, that is taken to be normally distributed). Manually analyzing and grouping large numbers of stocks with copious data is nearly impossible, yet that is necessary for stock analysis. Automating clustering methods to group stocks based on their returns is helpful in this regard. In this study, we used data from 77 of the top 100 Nasdaq stocks from 2018 to 2019, consisting of 504 daily adjusted closing prices over the said period of 2 years for each stock.

We considered each stock as its own random variable, with 504 days worth of adjusted closing prices as the supporting dataset. The data size thus equals $77 \times 504 = 38808$ entries, each of one dimension. For each stock X, we examined its $\ln(X_t/X_{t-1})$ values spanning 504 days, and estimated the lognormal parameters described in Section II-D by finding the mean and variance of $\{\ln(X_t/X_{t-1}); 1 \le t \le 504\}$. We also estimated the covariance of each pair of stocks X, Y by using the data $\{\ln(X_t/X_{t-1}), \ln(Y_t/Y_{t-1}); 1 \le t \le 504\}$.

To be able to evaluate the performance (accuracy, NMI, ARI), we created ground truth cluster labels by grouping stocks based on their yearly rate of returns: Low return (LR), Moderate return (MR), High return (HR), Low volatility (LV), Moderate volatility (MV), and High volatility (HV), as documented in Table VI and clustered in Fig. 7. We used six clustering algorithms, including the classical K-means and K-medoids and their distribution-based algorithms employing W_2 and ED measures. We evaluated the clustering performance using the accuracy, normalized mutual information (NMI), and adjusted Rand index (ARI) metrics.

TABLE VI
GROUND TRUTH STOCK CLUSTERS FROM 2018-19 RETURNS

Class	Stock Label
LR/LV	AEP, AMGN, BKNG, CHTR, CMCSA, CPT, CSCO, CSX, CTSH, EBAY, EXC, GILD, GOOG, GOOG, HON, MAR,MDLZ, PAYX, PCAR, PEP, TMUS, WBA, XEL
LR/MV1 (Very-low)	KHC
LR/MV2	ADI, AMAT, ATVI, AVGO, BIIB, DLTR, EA, META, INTC, MCHP, MNST, NVDA, NXPI, REGN, SWKS, TXN
MR/LV	AAPL, ADP, ANSS, CPRT, CTAS, FISV, IDXX, INTU, MSFT, ODFL, ORLY, ROST, SBUX, SNPS, VRSK, VRSN
MR/MV	ADBE, ADSK, ALGN, AMZN, CDNS, FAST, FTNT, ILMN, ISRG, KLAC, LRCX, MU, NFLX, PYPL, QCOM, TSLA, VRTX
HR/MV	AMD, DXCM, MTCH
HR/HV	ENPH

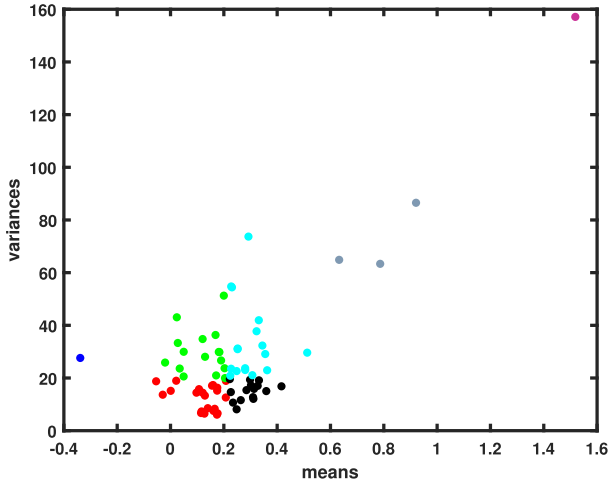


Fig. 7. Stock data ground clusters.

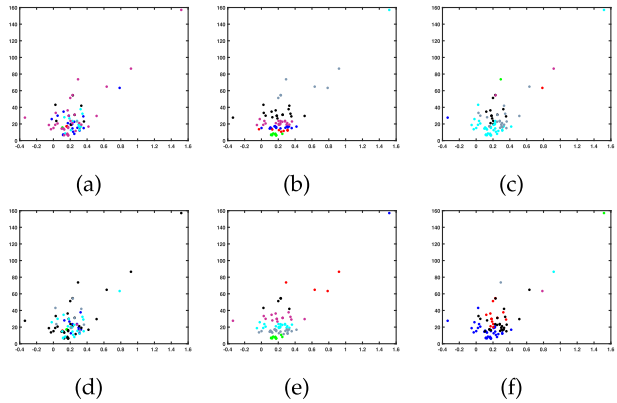


Fig. 8. Stock data result: Plots of data clusters (dots) using (a) KM, (b) WKM, (c) EKM, (d) KMd, (e) WKMd, and (f) EKMd.

The clustering results for stock datasets based on their lognormal distributions are shown in Fig. 8 and Table VII. We found the same progression of performance (classical $< W_2 <$ ED) of the clustering algorithms in terms of accuracy, NMI, ARI, and compute-time. The accuracies of KM, WKM, and EKM are 0.3766, 0.4675, and 0.5195 respectively, while the accuracies

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
EVALUATION MEASURES FOR STOCKS DATA

	Accuracy	NMI	ARI	Compute time(s)
KM	0.3766	0.1399	0.6391	0.523
WKM	0.4675	0.4336	0.7468	0.0459
EKM	0.5195	0.4384	0.7488	0.1533
KMd	0.3766	0.1362	0.6343	0.1015
WKMd	0.4935	0.4571	0.7478	0.0420
EKMd	0.5325	0.4651	0.7869	0.0358

of the corresponding K-medoids versions are 0.3766, 0.4935, and 0.5325 respectively. The corresponding NMIs for K-means are: 0.1399, 0.4336, 0.4384, and those for K-medoids: 0.1362, 0.4571, 0.4651; and the corresponding ARIs for K-means are: 0.6391, 0.7468, 0.7488, and those for K-medoids are 0.6343, 0.7478, 0.7869. These demonstrate that ED-based clustering results offer the highest accuracy, NMI, and ARI values while taking less time to compute as compared to the classical deterministic ones (.1533 sec versus. 523 sec for K-means and. 0358 sec versus 1015 sec for K-medoids).

In summary, we demonstrated the effectiveness of the clustering algorithms for non-Gaussian distributions using the stock market price data, which is commonly modeled as a lognormal distribution.

V. CONCLUSION

The paper introduced a new distance measure over distributions, called Expectation Distance (ED), and used it to develop noise-robust clustering algorithms, K-means, and K-medoids. A mathematical derivation proved that the proposed distance is a metric, satisfying the required properties of positivity, symmetry, zero if and only if equal, and triangle inequality. The presented distribution-based K-means and K-medoids methods cluster the data distributions first and then assign to each raw-data the cluster of its distribution. The ED-based K-means and K-medoids and W_2-distance-based K-medoids clustering were introduced for the first time. For both W_2 and ED, closed-form expressions for distribution distances were derived in terms of means and covariances, and those values were provided for the case of Gaussian and lognormal distributions. The paper also highlighted that the W_2-distance depends only on the marginal distributions, ignoring the correlation information. In contrast, the proposed ED overcomes this limitation by factoring in the correlation information and, in the process, yields higher noise-robust results. We also noted that while the cluster-centers of the distribution-based K-means are independent of the distance measure used, the same is not true of K-medoids. We implemented these noise-robust distance-based clustering algorithms and applied them to cluster noisy real-world weather and stock data by efficiently extracting and using the underlying uncertainty information (in terms of parameters of the distributions—Gaussian in case of weather data and lognormal in case of the stocks data). The real-life weather data results showed striking performance improvement for W_2-distance and ED-based K-means and K-medoids. Higher accuracy was observed for ED in both K-means and K-medoids: For a 35,280 entries of 3-D weather data spanning 4 seasons over 21 years and 5 stations, the accuracies of classical K-means, W_2 K-means, and ED K-means were found to be 0.5649, 0.8429, and 1.0000 respectively, whereas the accuracies of the corresponding K-medoids versions were 0.5637, 0.8500, and 0.9976 respectively. A similar performance progression was also obtained for stock data, demonstrating the method's effectiveness for non-Gaussian distributions. This performance validates the noise-robustness of the distribution-data-based clustering schemes and the benefits of factoring in the marginal distributions along with the joint distributions. It was also shown that while the distribution-implied clustering offers higher accuracy than the direct clustering of raw-data, strikingly, the former also has a lower time-complexity. Future research can explore applications to other distribution types, such as Gaussian mixtures.

ACKNOWLEDGMENT

The authors would like to thank CoAgMET (Colorado's Mesonet) for access to many years of real-life weather data.

REFERENCES

[1] B. Kao, S. D. Lee, F. K. Lee, D. W. Cheung, and W.-S. Ho, "Clustering uncertain data using voronoi diagrams and R-tree index," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 9, pp. 1219–1233, Sep. 2010.
[2] M. Chau et al., "Uncertain data mining: A new research direction," in *Proc. Workshop Sci. Artif.*, Hualien, Taiwan, Citeseer, 2005, pp. 199–204.
[3] C. Aggarwal and P. Yu, "A survey of uncertain data algorithms and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 5, pp. 609–623, May 2009.
[4] C. C. Aggarwal, *Managing and Mining Uncertain Data*, vol. 35, New York, NY, USA: Springer-Verlag, 2009.
[5] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 751–763, Apr. 2013.
[6] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Evaluating probabilistic queries over imprecise data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2003, pp. 551–562, doi: 10.1145/872757.872823.
[7] J. Pei, B. Jiang, X. Lin, and Y. Yuan, "Probabilistic skylines on uncertain data," in *Proc. 33rd Int. Conf. Very Large Data Bases*, 2007, pp. 15–26.
[8] Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. Kao, and S. Prabhakar, "Indexing multi-dimensional uncertain data with arbitrary probability density functions," in *Proc. 31st Int. Conf. Very Large Data Bases*, 2005, pp. 922–933.
[9] F. D. A. de Carvalho, A. Irpino, R. Verde, and A. Balzanella, "Batch self-organizing maps for distributional data with an automatic weighting of variables and components," *J. Classification*, vol. 39, no. 2, pp. 343–375, 2022.
[10] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning invariant features for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019.
[11] K. Budhathoki, D. Janzing, P. Bloebaum, and H. Ng, "Why did the distribution change?," in *Proc. Int. Conf. Artif. Intell. Statist.*, PMLR, 2021, pp. 1666–1674.
[12] Z. Wu, B. Wang, and C. Li, "A new robust fuzzy clustering framework considering different data weights in different clusters," *Expert Syst. Appl.*, vol. 206, 2022, Art. no. 117728.
[13] D. Paul, S. Chakraborty, S. Das, and J. Xu, "Kernel k-means, by all means: Algorithms and strong consistency," 2020, *arXiv: 2011.06461*.
[14] Z. Rustam and A. Talita, "Fuzzy kernel k-medoids algorithm for anomaly detection problems," in *Proc. AIP Conf.*, AIP Publishing, 2017, pp. 030154-1–030154-7.
[15] B. Tavakkol and Y. Son, "Fuzzy kernel k-medoids clustering algorithm for uncertain data objects," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 1287–1302, 2021.

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.

[16] M. Ay, L. Özbakir, S. Kulluk, B. Gülmez, G. Öztürk, and S. Özer, "FC-kmeans: Fixed-centered k-means algorithm," *Expert Syst. Appl.*, vol. 211, 2023, Art. no. 118656.
[17] J. Yuan and Y. Tian, "Practical privacy-preserving MapReduce based k-means clustering over large-scale dataset," *IEEE Trans. Cloud Comput.*, vol. 7, no. 2, pp. 568–579, Second Quarter 2019.
[18] M. O. Shafiq and E. Torunski, "A parallel k-medoids algorithm for clustering based on MapReduce," in *Proc. IEEE 15th Int. Conf. Mach. Learn. Appl.*, 2016, pp. 502–507.
[19] L. Wan, G. Zhang, H. Li, and C. Li, "A novel bearing fault diagnosis method using spark-based parallel ACO-k-means clustering algorithm," *IEEE Access*, vol. 9, pp. 28753–28768, 2021.
[20] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 56–70, 2020.
[21] E. Ahn, A. Kumar, D. Feng, M. Fulham, and J. Kim, "Unsupervised feature learning with k-means and an ensemble of deep convolutional neural networks for medical image classification," 2019, *arXiv: 1906.03359*.
[22] J. A. Bakar et al., "Tiktok video cluster analysis based on trending topic," in *Proc. Int. Conf. Comput. Informat.*, Springer, 2023, pp. 193–205.
[23] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.
[24] A. Müller, "Integral probability metrics and their generating classes of functions," *Adv. Appl. Probab.*, vol. 29, no. 2, pp. 429–443, 1997.
[25] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," *IEEE Signal Process. Mag.*, vol. 7, no. 2, pp. 568–579, Second Quarter 2019.
[26] A. Galichon, P. Henry-Labordere, and N. Touzi, "A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options," *Ann. Appl. Probab.*, vol. 24, no. 1, pp. 312–336, 2014.
[27] J. Rabin, G. Peyré, J. Delon, and M. Bernot, "Wasserstein barycenter and its application to texture mixing," in *Proc. Int. Conf. Scale Space Variational Methods Comput. Vis.*, Springer, 2011, pp. 435–446.
[28] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *J. Appl. Probab.*, vol. 56, no. 3, pp. 830–857, 2019.
[29] G. Peyré et al., "Computational optimal transport: With applications to data science," *Found. Trends Mach. Learn.*, vol. 11, no. 5/6, pp. 355–607, 2019.
[30] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," *Michigan Math. J.*, vol. 31, no. 2, pp. 231–240, 1984.
[31] G. Domazakis, D. Drivaliaris, S. Koukoulas, G. Papayiannis, A. Tsekrekos, and A. Yannacopoulos, "Clustering measure-valued data with Wasserstein barycenters," 2019. [Online]. Available: https://arxiv.org/abs/1912.11801
[32] J. Ye, P. Wu, J. Z. Wang, and J. Li, "Fast discrete distribution clustering using Wasserstein barycenter with sparse support," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2317–2332, May 2017.
[33] I. Verdinelli and L. Wasserman, "Hybrid Wasserstein distance and fast distribution clustering," *Electron. J. Statist.*, vol. 13, no. 2, pp. 5088–5119, 2019.
[34] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," 2017, *arXiv: 1711.01558*.
[35] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2018, pp. 5689–5698.
[36] G. E. Batista et al., "A study of K-nearest neighbour as an imputation method," *His*, vol. 87, no. 251/260, pp. 251–260, 2002.
[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B. (Methodol.)*, vol. 39, no. 1, pp. 1–22, 1977.
[38] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 90–105, 2004.
[39] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.
[40] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2292–2300.
[41] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a Wasserstein loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2053–2061.
[42] P. C. Álvarez Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán, "A fixed-point approach to barycenters in Wasserstein space," *J. Math. Anal. Appl.*, vol. 441, no. 2, pp. 744–762, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022247X16300907
[43] A. Vattani, "K-means requires exponentially many iterations even in the plane," in *Proc. 25th Annu. Symp. Comput. Geometry*, 2009, pp. 324–332.
[44] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *J. Stat. Mechanics: Theory Exp.*, vol. 2005, no. 09, 2005, Art. no. P09008.
[45] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.
[46] J. Liang, L. Bai, C. Dang, and F. Cao, "The *k*-means-type algorithms versus imbalanced data distributions," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 4, pp. 728–745, Aug. 2012.
[47] I. Antoniou, V. Ivanov, V. Ivanov, and P. Zrelov, "On the log-normal distribution of stock market data," *Physica A: Stat. Mechanics Appl.*, vol. 331, no. 3, pp. 617–638, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437103008987

Rahmat Adesunkanmi (Student Member, IEEE) received the BS degree from the University of Ibadan, Nigeria, in 2015. She is cirrently working toward the PhD degree with the Department of Electrical and Computer Engineering, Iowa State University, Ames. Her research interest includes noise-robust machine learning and different clustering techniques. She is an active member of the Graduate students Society for Women in Engineering.

Ratnesh Kumar (Fellow, IEEE) received the BTech degree in electrical engineering from IIT Kanpur, India, in 1987, and the MS and PhD degrees in electrical and computer engineering from the University of Texas at Austin, in 1989 and 1991, respectively. He is a Palmer professor with the Department of Electrical and Computer Engineering, Iowa State University, where he directs the Embedded Software, Sensors, Networks, Cyberphysical, and Energy (ESSeNCE) Lab. Previously, he held a faculty position with the University of Kentucky and various visiting positions with the University of Maryland (College Park), the Applied Research Laboratory with the Pennsylvania State University (State College), NASA Ames, the Idaho National Laboratory, the United Technologies Research Center, and the Air Force Research Laboratory. He is a fellow of AAAS, and was a distinguished lecturer of the IEEE Control Systems Society. He is a recipient of *D. R. Boylan Eminent Faculty Award for Research* and *Award for Outstanding Achievement in Research* from Iowa State University, and also the *Distinguished Alumni Award* from IIT Kanpur. He received Gold Medals for the Best EE Undergrad, the Best All Rounder, and the Best EE Project from IIT Kanpur, and the Best Dissertation Award from UT Austin, the Best Paper Award from *IEEE Transactions on Automation Science and Engineering*, and has been Keynote Speaker and paper award recipient from multiple conferences. He is or has been an editor of several journals (including of IEEE, SIAM, ACM, Springer, IET, MDPI).

Authorized licensed use limited to: Iowa State University. Downloaded on February 14,2025 at 17:48:01 UTC from IEEE Xplore. Restrictions apply.