
Density based Spatial Clustering of Lines via Probabilistic Generation of Neighbourhood

Akanksha Das

Indian Statistical Institute
akankshadas2020@gmail.com

Abstract

Density based spatial clustering of points in \mathbb{R}^n has a myriad of applications in a variety of industries. We generalise this problem to the density based clustering of lines in high-dimensional spaces, keeping in mind there exists no valid distance measure that follows the triangle inequality for lines. In this paper, we design a clustering algorithm that generates a customised neighbourhood for a line of a fixed volume (given as a parameter), based on an optional parameter as a continuous probability density function. This algorithm is not sensitive to the outliers and can effectively identify the noise in the data using a cardinality parameter. One of the pivotal applications of this algorithm is clustering data points in \mathbb{R}^n with missing entries, while utilising the domain knowledge of the respective data. In particular, the proposed algorithm is able to cluster n -dimensional data points that contain at least $(n - 1)$ -dimensional information. We illustrate the neighbourhoods for the standard probability distributions with continuous probability density functions and demonstrate the effectiveness of our algorithm on various synthetic and real-world datasets (e.g., rail and road networks). The experimental results also highlight its application in clustering incomplete data.

1 Introduction

To understand the underlying structure of data, it is customary to explore groups within it [1]. The approach of identifying groups in data through unsupervised learning techniques is popularly known as clustering. Clustering of data points has received enormous attention in the last few decades. However recently, there is an emerging interest to generalize the clustering of data points to other shapes, particularly lines [2, 3]. The state-of-the-art clustering methods for points can not be directly extended to other shapes. The major challenges behind addressing this new class of problems are defining an appropriate distance measure and interpretation of clustering results.

There are many different approaches of clustering that perform well only for a specific type of datasets. For example, centroid-based clustering methods work well only for convex datasets [4]. The clustering of straight lines has become an appealing problem in recent years [2, 3, 5]. Unfortunately, this problem has been addressed with centroid-based approaches only. These methods suffer from data in which varying density of clusters is a concern. In density based clustering, groups are formed in a data space covering a continuous area with a high point density, divided from other groups by sparse regions. Density based spatial clustering of points in \mathbb{R}^n has a myriad of applications in a variety of industries. In this paper, we propose a density based spatial clustering approach (termed as DeLi) of lines via probabilistic generation of neighbourhood aiming to cluster lines. The proposed approach is inspired from a density-based spatial clustering approach that effectively deals with noise (popularly known as DBSCAN) [6].

We address the problem of clustering of lines (or line segments) in a generic form. Unlike the conventional scenarios of clustering of lines, where every point on the same line get an equal

preference, we make it flexible by introducing the concept of probability. We consider datasets consisting of both points and lines to demonstrate the effectiveness of our approach. Instead of considering a distance measure between a pair of lines, we adopt a novel concept of probabilistic neighbourhood generation to measure the proximity of those lines. In this way, our contributions in this area are indeed novel.

The rest of the paper is organized as follows. Section 2 discusses the major motivations behind the current work. Section 3 covers the state-of-the-art. Section 4 introduces the precursory notations and terminologies and formulates the problem. Section 5 presents the theoretical insights and details on the proposed approach of clustering of lines. Section 6 gives an overview of the datasets and presents the detailed experimental results. Finally, Section 6 concludes the paper by highlighting its advantages, disadvantages, and future prospects.

2 Motivation

There are several applications highlighted in the literature related to clustering of lines [2, 3]. However, in this paper, we bring the concept of density around a line envisioning its broader scope. This helps us to look at the problem of clustering in a generic way to foresee additional applications. Grouping of incomplete data is the major motivation behind the current work [7]. Let there be a set of points in \mathbb{R}^n that we have to cluster, even though some of them might possess incomplete entries (i.e., values are not available for all the n -given dimensions). Mathematically speaking, say a data item $D \in \mathbb{R}^n$ have missing values on k of its entries such that $k \ll n$. To address this version of clustering problem [8], we may consider the set of possible values of D as a $(n - k)$ dimensional object in \mathbb{R}^n . In this study, we particularly deal with the scenarios where $k = 1$, i.e., we visualise such incomplete information as lines or line segments, and cluster points and line segments together under the same approach.

3 Related Work

The very first attempt to formulate clustering of lines was made by Peretz in the form of finding k -median for lines back in 2011 [5]. Given a set of lines L , he aimed to find k points (facilities) such that the sum of distance from each line in L to its closest point is minimized. Peretz proposed both exact and approximate algorithms for the said problem. It was also shown that k -median for lines problem can be solved efficiently for a constant k , however is NP-complete if k is non-constant. Unfortunately, this approach works only well for convex datasets and for the cases where density is not varying. Moreover, it cannot identify the outliers in the data. Later Moram et al. have addressed the problem of clustering of lines with another centroid-based approach, precisely k -means in 2019 [3]. This also suffers from similar kind of limitations. We propose a density-based approach to overcome these limitations. Moreover, in recent times, grouping of a set of points with varying amount of reliability has been addressed by probabilistic clustering approaches for various applications [9, 10]. In this paper, we adopt a probabilistic approach in this paper to set up the neighbourhood of a dense cluster.

4 Proposed Problem

We first introduce the notations that will be used throughout the paper followed by the statement of the problem that we address.

4.1 Basic Notations

Let us define \mathcal{U} as the set of all lines or line segments in \mathbb{R}^n that we want to cluster. A line $l \in \mathbb{R}^n$ is denoted by the coordinates of any two distinct points on the line. On the other hand, a line segment is denoted by the coordinates of its end points, i.e., $l \in \mathbb{R}^n$ is represented by (x, y) , where $x, y \in \mathbb{R}^n$. Let us assume that a point P on l is denoted as $P \in l$. We consider ℓ^2 -norm as the distance measure between any pair of points, i.e., $\mathcal{D}(x, y) = \|x - y\|_2$, where $x, y \in \mathbb{R}^n$. Let $d_{P,l}$ denote the shortest distance from P to l , i.e., $d_{P,l} = \inf_{Q \in l} \mathcal{D}(Q, P)$. Let us also define $P_l = Q \in l$ such that $\mathcal{D}(Q, P) = d_{P,l}$.

Theorem 1 P_l exists and is unique with respect to the line or line segment l .

proof : We consider the two cases based on whether $l = (x, y)$ is a line or line segment. \square

Case 1: If l is a line, $d_{P,l}$ is the perpendicular distance from P to l hence, P_l is the foot of the perpendicular from P to l . Proof is immediate.

Case 2: If l is a line segment, and foot of the perpendicular (say F) lies on l then same as in Case 1. If not, then both x and y lie on the same side of F and x, y and F are co linear. Wlog, let x lie closer to F than y then $d_{P,l} = \mathcal{D}(x, P)$ and $P_l = x$. This is because $\mathbf{D}(P, z)$ is a strictly increasing function as z varies from x to y .

Let $V(A)$ denote the volume of region $A \subseteq \mathbb{R}^n$ if it exists. The other notations are standard unless specified otherwise.

4.2 Problem Formulation

Let us denote the continuous probability density function as $f_l : L \rightarrow \mathbb{R}^+ \cup \{0\}$, where $L = \{P : P \in l\}$ is taken as a parameter towards formulating the problem in a generic way. Let S_l denote the closure of $\{P \in l : f_l(P) > 0\}$. Based on this, we now define the neighbourhood of a line as follows.

Definition 1 (f -neighbourhood of a line) The f -neighbourhood of l is defined as $N_{f,l} = \{P \in \mathbb{R}^n : d_{P,l} < f(P_l)\}$, where $f : L \rightarrow \mathbb{R}^+ \cup \{0\}$ is any continuous function and $L = \{P : P \in l\}$.

We now introduce a scaling factor for the normalization of the f -neighbourhood. This scaling factor can either be learned from the data (precisely from the volume parameter) or taken as a free parameter (in the absence of volume parameter). For the former case, the scaling factor is defined as follows.

Theorem 2 Volume of f -neighbourhood of a line/line segment l always exists where $f : L \rightarrow \mathbb{R}^+ \cup \{0\}$ is any continuous probability density function and $L = \{P : P \in l\}$.

proof: Now f is continuous, bounded and integrable over L and f -neighbourhood of l is the region bounded by rotating f 360° about l , and l . Hence, $V(N_{f,l})$ exists. \square

Definition 2 (Scaling factor of f -neighbourhood of a line) The scaling factor of f -neighbourhood of l is defined as $\alpha_l = \frac{V}{V(N_{f_l,l})}$, where V is the volume parameter.

Using this scaling factor, we finally define the neighbourhood of a line as follows.

Definition 3 (Neighbourhood of a line) The neighbourhood of l is defined as $N_l = N_{(\alpha_l * f_l),l}$, i.e., the $(\alpha_l * f_l)$ -neighbourhood of a line.

Now, we define the neighbourhood relation \mathcal{R} over \mathcal{U} , to determine whether l_1 is in the neighbourhood of l_2 . Neighbourhood relation \mathcal{R} is defined using neighbourhood of a line or using a suitable distance metric in the absence of f_l parameter.

Definition 4 (Neighbourhood Relation) Neighbourhood relation is defined as

$$l_1 \mathcal{R} l_2 = \begin{cases} \exists P \in S_{l_2} \text{ such that } P \in N_{l_1}, & \text{if } f_l \text{ is available as a parameter.} \\ \inf_{p \in l_1, q \in l_2} \mathcal{D}(p, q) < \alpha_{l_1}, & \text{otherwise.} \end{cases}$$

Note that \mathcal{R} is reflexive but not symmetric and transitive. We define the l -cardinality to verify whether a sufficient number of lines (a lower bound) is included within a cluster. This is defined as follows.

Definition 5 (l -cardinality) We define l -cardinality as the cardinality of the set $\{l_1 : l_1 \in \mathcal{U} \text{ and } l \mathcal{R} l_1\}$

Given all the above precursory notations and definitions, we formulate the target problem as follows.

Problem Statement: Given \mathcal{U} (a set of lines / line segments in \mathbb{R}^n), and a set of parameters as listed below:

- Compulsory parameter: c (cardinality),
- Substitute Parameters : V (volume) or α_l (scaling factor), and
- Optional parameter : f_l ,

we need to find out density based clusters of the lines / line segments in \mathcal{U} .

5 Proposed Method

We first put forward some theoretical aspects of the function f_l , to elaborate on the computation of the neighbourhood relation \mathcal{R} and then present the algorithm for density based spatial clustering of lines / line segments.

5.1 Theoretical Insights

Clustering a set of lines by the existing approaches assumes all the points on a particular line / line segment to have equal importance [9]. However, in some real-world applications, it may be useful to assume that they have varying amounts of reliability. Hence, we assign a probability density to each of the points on a line / line segment and generate a neighbourhood accordingly. For example, if p_1 and p_2 are points lying on l such that $f_l(p_1) \gg f_l(p_2)$, then the span of the neighbourhood of l at p_1 will be much larger than the span of the neighbourhood of l at p_2 . Assigning a probability density, gives us the autonomy to quantify the importance or reliability of the points on the line. This plays a vital role in incorporating domain knowledge of the variable with missing values during the clustering of incomplete data.

The neighbourhoods for the standard probability distributions are shown in Table 1. As can be seen from Table 1, the neighbourhoods in \mathbb{R}^3 are simple extensions of what we have as neighbourhoods in \mathbb{R}^2 . To illustrate this in depth, it is beneficial to transform the domain of f_l from $\{P : P \in l\}$ to a subset of \mathbb{R} . This makes it easier to define f_l , thereby making the implementation simpler too. For a line segment $l = (x, y)$, where $x, y \in \mathbb{R}^n$, the set of points belonging to $l = (x, y)$ can be written as $\{x + (y - x)t : t \in [0, 1]\}$. Similarly, the set of points belonging to a line can be denoted as $\{x + (y - x)t : t \in \mathbb{R}\}$. As there is a bijection, more precisely a linear function, say g_l such that $g_l(0) = x, g_l(1) = y$ between the value t and the points on the line or line segment, f_l may be re-written as a function of $t \in [0, 1]$ (for line segments) or \mathbb{R} (for lines) respectively. In Table 1, we define f_l as a function of t for the standard probability distributions, and show the region N_l in \mathbb{R}^2 and \mathbb{R}^3 . For each of the cases in Table 1, f_l is defined over \mathbb{R} .

5.2 Algorithm

The proposed density based clustering algorithm of lines (termed as DeLi) has three versions differing only in the definition of the neighbourhood relation \mathcal{R} . Hence, the definitions of \mathcal{R} along with their associated parameters is concisely presented in Table 2.

We formally present the approach in Algorithm 1. The input to the algorithm is a set of lines / line segments in \mathcal{U} and a set of appropriate parameters in accordance with the desired version (refer to Table 2). The algorithm closely mirrors the DBSCAN algorithm [6] for points with the neighbourhood being defined with the help of a probability density function f_l instead of a metric. All lines / line segments are declared unvisited. If at least one line / line segment is unvisited, choose any one of them, say l_U and run a loop over all lines / line segments to check their inclusions in the neighbourhood. If $l_U \mathcal{R} l_i$ holds (refer to Table 2), then add l_i to the list N_U . To verify whether $l_U \mathcal{R} l_i$ holds, we choose one of the following versions based on the input parameters.

Version 1: The minimum distance between $l_U = (x_U, y_U)$ and $l_i = (x_i, y_i)$ is strictly less than α_U i.e. $\min_{x \in l_i, y \in l_U} \|x - y\|_2^2 < \alpha_U^2$ that is equivalent to $\min_{t_1 \in D_1, t_2 \in D_2} \|g_{l_i}(t_1) - g_{l_U}(t_2)\|_2^2 < \alpha_U^2$,

Table 1: Generation of neighbourhoods in \mathbb{R}^2 and \mathbb{R}^3 using different standard probability distributions around a line. The scaling factor is considered to be 1.


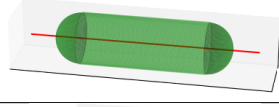
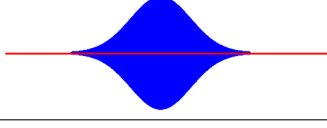
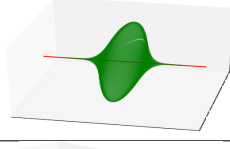
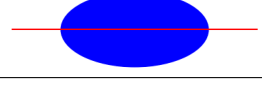
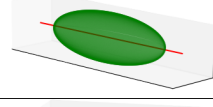
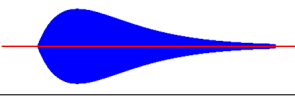
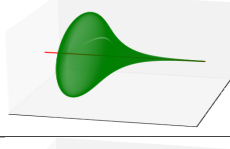
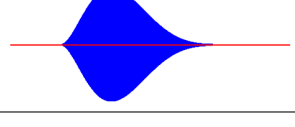
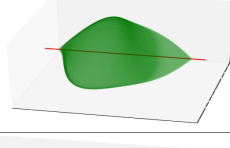
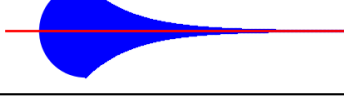
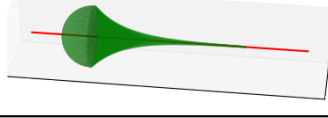
Distribution	Parameters	2-D Neighbourhood	3-D Neighbourhood
Uniform	(a, b)		
Normal	(μ, σ^2)		
Ellipsoidal	(a, b)		
Gamma	(α, λ)		
Beta	(α_1, α_2)		
Exponential	(λ)		

Table 2: Different versions of the algorithm for the computation of $l_1 \mathcal{R} l_2$ and the respective output based on the input parameters.

Version	Parameters			Computation of $l_1 \mathcal{R} l_2$ (see Def. 4)
	Mandatory	Substitute	Optional	
1	c	α_l	-	$\inf_{p \in l_1, q \in l_2} \mathcal{D}(p, q) < \alpha_{l_1}$
2	c	V	f_l	$\exists P \in S_{l_2}$ such that $P \in N_{l_1}$
3	c	α_l	f_l	$\exists P \in S_{l_2}$ such that $P \in N_{l_1}$

where $D_1, D_2 = \mathbb{R}$ or $[0, 1]$. 5.1. Now g_{l_i}, g_{l_U} are linear functions hence the problem reduces to finding out the minimum value of a quadratic equation in two variables.

Version 2: Calculate the $V(N_{l_U})$ and subsequently $\alpha_{l_U} = \frac{V}{V(N_{l_U})}$. Now, if any point in l_i lies in the $(\alpha_{l_U} * f_{l_U})$ -neighbourhood of l_U , then $l_U \mathcal{R} l_i$.

Version 3: If any point in l_i lies in the $(\alpha_{l_U} * f_{l_U})$ -neighbourhood of l_U , then $l_U \mathcal{R} l_i$.

If cardinality of N_U is at least c , the l_U and the elements of N_U are labeled as visited and are included in a new cluster C_j where j is auto-incremented. Otherwise, l_U is labeled as noise i.e an outlier. Repeat until no line/line segment is left unvisited. Finally the clusters C_1, C_2, \dots, C_k are returned as the result.

5.3 Complexity Analysis

The time and space requirements of the proposed algorithm are discussed below.

Algorithm 1: DeLi: Density based spatial clustering of lines via probabilistic generation of neighbourhood

Input: A set of lines or line segments $\mathcal{U} = \{l_1, l_2, \dots, l_n\}$, a cardinality parameter c .

Output: The clusters C_1, C_2, \dots, C_k .

Steps of the algorithm:

```

for  $l_i \in \mathcal{U}$  do
  | Label  $l_i$  as UNVISITED
end
 $j \leftarrow 1$ 
while At least one  $l_i \in \mathcal{U}$  is UNVISITED do
  |  $l_U \leftarrow$  A randomly chosen  $l_i$  that is labeled as UNVISITED
  |  $N_U \leftarrow \{l_i \in \mathcal{U} : l_U \mathcal{R} l_i\}$  /* Computation of  $l_U \mathcal{R} l_i$  is elaborated in Table 2 */
  if  $|N_U| \geq c$  then
    | Label  $l_U$  and the lines or line segments in  $N_U$  as VISITED
    | Include  $l_U$  and the lines or line segments in  $N_U$  in a new cluster  $C_j$ 
    |  $j \leftarrow j + 1$ 
  else
    | Label  $l_U$  as NOISE
  end
end
return  $C_1, C_2, \dots, C_k$ 

```

Time complexity: The initial labelling in the proposed algorithm requires $O(n)$ time, for n lines / line segments. Thereafter, the lines / line segments are processed one at a time and for each of them relation to the others are computed, incurring $O(n^2)$ time. Hence, the total worst case time complexity becomes $O(n) + O(n^2) \equiv O(n^2)$.

Space complexity: The proposed algorithm computes \mathcal{R} , which is not symmetric. Hence, we need not store a matrix representation of computations of neighbourhood within an auxiliary space to avoid re-computations during the runtime. So, for the n lines / line segments (each represented as a doublet) to be clustered, the worst case space complexity is $O(n)$.

6 Results

To test the effectiveness of the proposed approach, we have applied it on various types of datasets. We first detail on the datasets used and then discuss the experimental results obtained.

6.1 Dataset Details

The datasets considered for experimental analysis were of two types – Line Datasets and Point Datasets with Missing Entries. The Line Datasets are further categorised as Synthetic Datasets and Real-world Datasets. The details about the datasets are provided below (see Table 3 for more details).

- **Line Datasets:** These datasets include line segments and are of two types. No domain knowledge is used for this kind of datasets.
 - **Synthetic:** These are prepared by writing programs with carefully controlled parameters. There are two such datasets, namely Convex and Doughnut.
 - **Real-world:** These are prepared by selecting a portion of real-world datasets to create clusters recognisable through human perception to test the efficacy of our algorithm. There are six such datasets namely Sparse Tripod, Dense Tripod, Broken Beads 1, Broken Beads 2, Spectral Band 1, and Spectral Band 2, which include Amtrak routes and subway lines in USA [11].
- **Point Dataset with Missing Entries:** This dataset includes points with missing entries and are of one type. We consider domain knowledge for this kind of datasets.
 - **Real-world:** This is prepared from a real-world dataset by including missing values randomly. There is only one such dataset, namely Sporulation, which contains expres-

sion values of yeast genes measured across seven time points (0, 0.5, 2, 5, 7, 9, and 11.5h) during the sporulation process of budding yeast [12].

Table 3: Statistical and background details of the various datasets used. Some line segments from some real-world datasets were cropped out to allow forming a realistic cluster.

Type	Name	# Data Points	Source
Lines (Synthetic)	Convex	150	Program-generated
	Doughnut	400	Program-generated
Lines (Real-world)	Sparse Tripod	50	amtrak-routes [11]
	Dense Tripod	100	amtrak-routes [11]
	Broken Beads 1	59	Amtrak [11]
	Broken Beads 2	50	Amtrak [11]
	Spectral Band 1	190	Subway Lines [11]
	Spectral Band 2	2000	Subway Lines [11]
Points with Missing Entries (Real-world)	Sporulation	475	Chu et al. [12]

6.2 Empirical Analysis

A set of programs are written in Python for the empirical analysis. All the experiments have been performed on an HP Laptop with 2.2 GHz Intel 13th Generation Core i7 processor, 18 MB Cache memory, and having 16 GB primary memory.

6.2.1 Results on Line Datasets

The results obtained for the different line datasets are reported in Fig. 1. Notably, we apply only the first version of neighborhood computation of Algorithm 1 in all these cases due to the absence of domain knowledge. Moreover, as all the line segments have equal weightage, we take the scaling factor (α_l) to be independent of the line segments. As can be seen from Figs. 1(a) and (b), the proposed approach is capable of exploring groups from both the convex and non-convex (with a little abuse of notation) orientation of lines. However, the cardinality parameter plays a crucial role in capturing the essence of true clusters in non-convex cases. For example, lowering the cardinality threshold of neighbourhood (c), helps to get larger clusters by ignoring small variations in density while increasing it, makes the algorithm sensitive to variation in density (compare Fig. 1(b) with Fig. 1(c)). The identification of outliers is also dependent on the cardinality parameter (more outliers in Fig. 1(c) than Fig. 1(b)). Note that the scaling factors (α_l) were the same for the synthetic datasets because they were generated on a similar scale. However, for the other line datasets prepared from real-word data, we need to set the scaling factor (α_l) separately. On carefully examining the real-word datasets (Figs. 1(d) and (e)), we can see that the algorithm can recognise dense and sparse regions along with outliers.

6.2.2 Results on Point Datasets with Missing Entries

Clustering results are obtained for the Sporulation dataset (that we prepared by adding missing entries) considering $\alpha_l = 0.6$ and $c = 7$. The domain knowledge was incorporated to generate the neighbourhood by setting up an appropriate probability distribution. To be precise, we set the function f_l as Unif[-4, 4] based the range of the values in respective columns in the dataset. The output yielded 5 clusters of yeast genes from the dataset along with a set of 120 outlier genes. This was similar to the results obtained with the original dataset in which no missing values were present. A detailed comparison of results is reported in the Table 4. The same algorithm (with the same parameters) applied on the original dataset yielded 4 clusters of yeast genes from the dataset along with a set of 128 outlier genes. There was a significant overlap between the outliers identified from the original dataset and the dataset prepared by introducing missing entries. This demonstrates the effectiveness of the proposed approach when domain knowledge is brought in the model.

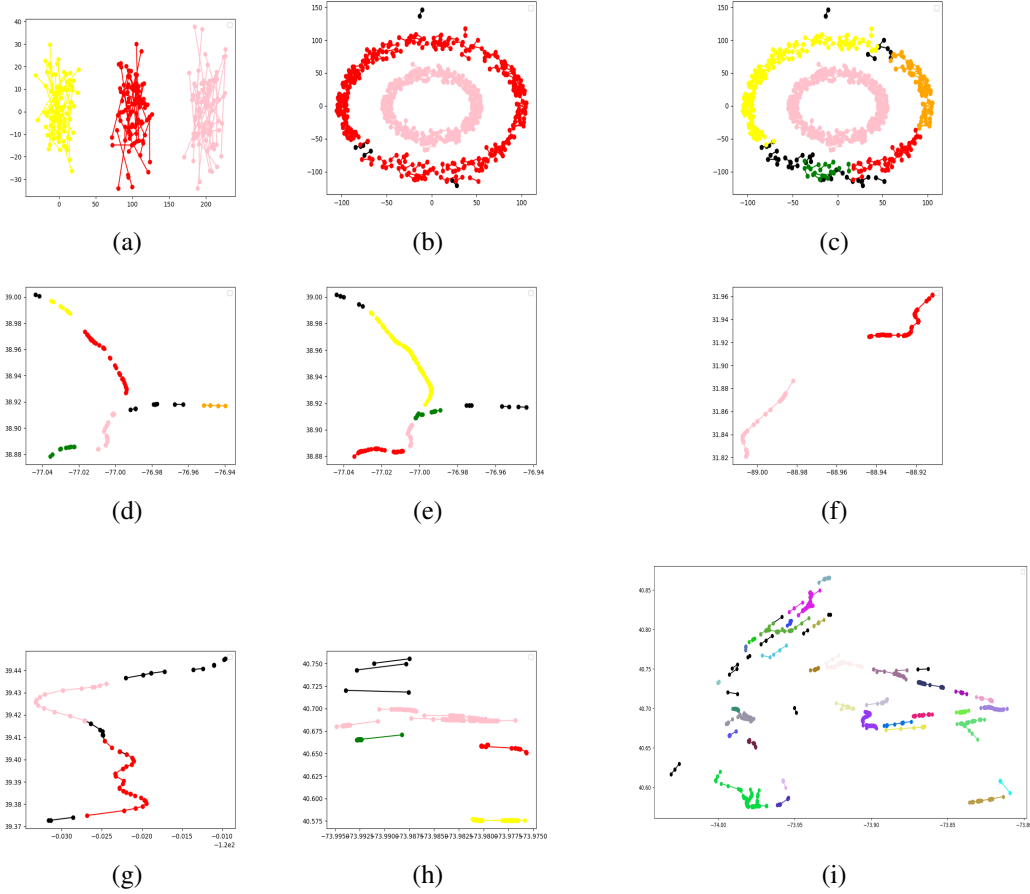


Figure 1: Clustering results for (a) Convex dataset considering $\alpha_l = 12$ and $c = 10$, (b) Doughnut dataset considering $\alpha_l = 12$ and $c = 5$, (c) Doughnut dataset considering $\alpha_l = 12$ and $c = 8$, (d) Sparse Tripod dataset considering $\alpha_l = 0.008$ and $c = 3$, (e) Dense Tripod dataset considering $\alpha_l = 0.005$ and $c = 3$, (f) Broken Beads 1 dataset considering $\alpha_l = 0.005$ and $c = 3$, (g) Broken Beads 2 dataset considering $\alpha_l = 0.005$ and $c = 8$, (h) Spectral Band 1 dataset considering $\alpha_l = 0.008$ and $c = 5$, (i) Spectral Band 2 dataset considering $\alpha_l = 0.005$ and $c = 3$.

Table 4: Comparison of clustering results obtained by applying the proposed algorithm on the original dataset and the dataset prepared by introducing missing entries.

Dataset	Algorithm Version	Parameters			Clustering Results			
		α_l	c	f_l	Count	Min	Max	#Outliers
Sporulation (Original)	-	0.6	7	-	4	17	219	128
Sporulation (With Missing Entries)	3	0.6	7	Unif[-4, 4]	5	5	222	120

7 Conclusion and Future Work

This paper presents a density based clustering algorithm that is particularly useful for grouping n -dimensional data points with missing entries. This approach does not require a pre-defined number of clusters. Moreover, it identifies outliers as noises. Interestingly, it can find arbitrarily-sized and arbitrarily-shaped clusters quite well. But this algorithm is not deterministic in nature. Through empirical analysis, we observe that it is effective on synthetic and real-world datasets. The proposed approach considers a number of parameters like volume, cardinality, and the density function to probabilistically generate the neighbourhood of a line. To fix the said parameters (or tune them easily), we require appropriate domain knowledge for the efficacy of the algorithm. Note that the current paper presents a simple implementation of the algorithm. If this is implemented using a range query approach with an appropriately maintained database index, we can obtain a better performance. The proposed clustering algorithm is capable of dealing with data that has at most one missing value per line. This approach can be further extended to the scenarios having higher number of missing entries per line.

References

- [1] Duda, R. O., & Hart, P. E. (2006) *Pattern classification*. John Wiley & Sons.
- [2] Marom, Y. (2019) k-Clustering of Lines and Its Applications. *Masters' Dissertation*, Department of Computer Science, University of Haifa.
- [3] Marom, Y., & Feldman, D. (2019) k-Means clustering of lines for big data. In *Advances in Neural Information Processing Systems*, 32.
- [4] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023) K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178-210.
- [5] Peretz, T. (2011) Clustering of lines. *Masters' Dissertation*, Department of Mathematics and Computer Science, Open University of Israel.
- [6] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Vol. 96, No. 34, pp. 226-231.
- [7] Allison, P. D. (2009) Missing data. *The SAGE Handbook of Quantitative Methods in Psychology*, 72-89.
- [8] Wilson, S. E. (2015) Methods for clustering data with missing values.
- [9] Zass, R., & Shashua, A. (2005) A unifying approach to hard and probabilistic clustering. In *Tenth IEEE International Conference on Computer Vision*, vol. 1, pp. 294-301.
- [10] Stylianopoulos, K., & Koutroumbas, K. (2021) A probabilistic clustering approach for detecting linear structures in two-dimensional spaces. *Pattern Recognition and Image Analysis*, 31:671-687.
- [11] Line Datasets – data.world. Accessed in May, 2024. Link: <https://data.world/datasets/line>
- [12] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., & Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, 282(5389), 699-705.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. See section 6 for a detailed outcome of this work.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper clearly reflect on the factors that influence the performance of the approach in section 5.1. It also discusses the computational efficiency of the proposed algorithms and how they scale with dataset size in section 5.3. The limitations of our work are also discussed in section 7.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All the theorems, formulas, and proofs in the paper are numbered and cross-referenced. All assumptions are clearly stated and the proofs appear in section 5.1 of the paper.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper is completely reproducible and the code and data are provided as supplementary materials. The datasets and how they are processed are detailed in section 6. Others can replicate the model with the same dataset based on the detailed instructions provided in the paper.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: The entire deck of codes and the associated datasets are provided as supplementary materials. The paper includes exact instructions and environment details such that the codes provided can be run on the datasets provided and new data to reproduce the results.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: The experimental settings are clearly presented in section 6 of the paper. The data splits, hyperparameters used, and how they were chosen, etc. are mentioned therein.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Error bar calculation or any kind of significance testing is not applicable for the paper as it shows results mostly with visualizations and physical interpretations.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Both the computational and memory resources are mentioned in the beginning of section 2 of the paper.

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: This paper preserves anonymity and the other required NeurIPS Code of Ethics.

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed on grouping lines and points.

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks as it addresses an unsupervised learning problem in a conventional approach.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The authors are the original owners and the codes are freely distributed as a part of supplementary materials. The codes are licensed under CC-BY 4.0.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve any research methodology, data or tools that requires IRB approval (or equivalent).