LINSCAN - A Linearity Based Clustering Algorithm

Andrew Dennehy

Abstract

DBSCAN is a very powerful algorithm for clustering in domains where few assumptions can be made about the structure of the data. In this paper, we hope to leverage these strengths for more specialized clustering tasks where the standard algorithm fails. In particular, by embedding points as normal distributions approximating their local neighborhoods and leveraging a distance function derived from the Kullback Leibler Divergence, we hope to create an algorithm which can distinguish highly linear clusters that are spatially close but have orthogonal covariances. We then apply this to the identification of slip faults via measuring seismic activity.

1. Introduction

Many existing clustering algorithms require some prior knowledge of the dataset and are limited in the possible shapes they can identify. For example, both K-Means Clustering and GMM Expectation Maximization require a prior estimate of the number of clusters existing in the dataset and struggle to distinguish clusters with complicated interactions. For example, these algorithms would struggle to distinguish between a cluster and a surrounding ring of points.

In contrast, DBSCAN iteratively generates clusters by leveraging a heuristic for the local behavior of clustered points. In particular, its designers equated clusters to connected regions of high density (Ester et al., 1996). Thus, by identifying points whose local neighborhoods are highly dense, we can generate clusters with little prior knowledge about their possible shapes by iteratively growing clusters from those points. Furthermore, it is done in a way where the number of clusters comes out of the process, rather than being a parameter itself.

In this paper, we seek to leverage this characterization of clusters using distance measures other than Euclidean distance. In particular, we hope to develop an algorithm which can delineate between multiple highly linear clusters which are nearby in space but have nearly orthogonal covariances.

1.1. Notation

Here we summarize notation that will be used throughout the rest of the paper:

1. For $\epsilon > 0$, we let $B_\epsilon(x)$ be the open ball of radius ϵ centered at x (in the standard Euclidean norm).
2. For finite $A \subseteq \mathbb{R}^d$,
   (a) $\mu_A \in \mathbb{R}^d$ is the sample mean of A
   (b) $\Sigma_A^2 \in \mathbb{R}^{d\times d}$ is the sample covariance of A
3. Given $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ with Σ symmetric positive definite, $\mathcal{N}(\mu, \Sigma)$ is the multivariate Gaussian with mean μ and covariance Σ.
4. For two probability measures P and Q on the same space $\mathcal{X}$, we write $P \ll Q$ if P is absolutely continuous with respect to Q, meaning that for any $Q-$measurable $A \subseteq \mathcal{X}$, A is $P-$measurable and if $Q(A) = 0$ then $P(A) = 0$.
5. If $P \ll Q$, $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Q-almost everywhere unique function such that for $Q-$measurable $A \subseteq \mathcal{X}$
$$P(A) = \int_A \frac{\mathrm{d}P}{\mathrm{d}Q}\mathrm{d}Q$$
6. We write $P \sim Q$ if P and Q are equivalent, meaning $P \ll Q$ and $Q \ll P$.
7. We let $\|A\|_F := \mathrm{tr}(A^T A)$ denote the Frobenius norm.
8. For positive definite A, $\|x\|_A := \sqrt{x^T A x}$ is the elliptic norm defined by A.
9. We let $\mathbb{P}(\mathcal{X})$ be the set of probability distributions over $\mathcal{X}$.

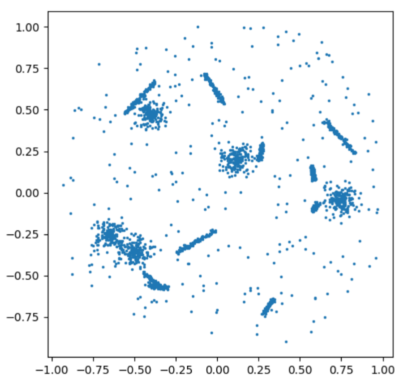
1.2. Motivating Problem

This work was motivated by the desire to isolate highly linear clusters in point clouds and distinguish clusters which were close in physical space but had different local directions. This originated with the goal of designing an algorithm to identify slip faults based on locations of seismic activity. Importantly, because slip faults being perfectly linear, corollary seismic activity be close to linear after accounting for noise.

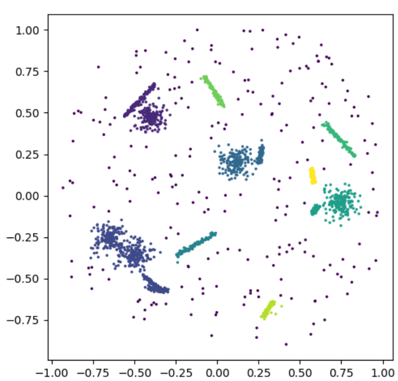
To highlight the insufficiency of existing algorithms for this task, consider Figure 1a. Note how there are clearly highly linear clusters and highly nonlinear clusters, as well as locations (in particular near (-.4,-.6)) where two highly linear clusters intersect obliquely. Figure 1b shows the results obtained by applying DBSCAN (described in the next section) to the data. Note how both the linear and nonlinear clusters are identified, how highly linear clusters are conjoined with highly nonlinear clusters, and how we see no separation between the intersecting linear clusters.

Figure 1. Test Data

(a) Data



(b) DBSCAN Results



2. Background: DBSCAN

2.1. The Algorithm

We begin by providing a description of DBSCAN. The main principle behind DBSCAN is that clusters are equivalent to connected regions of high density. Thus, the most natural way to identify clusters is to search for points whose local neighborhoods contain a high density of points from the dataset and inductively grow clusters from those points.

Assume we have a dataset $X = \{x_1, ..., x_m\} \subseteq \mathbb{R}^d$. Then, for $\epsilon > 0$ we let

$$\begin{aligned} R_\epsilon : X &\to \mathcal{P}(X) \\ x &\mapsto X \cap B_\epsilon(x) \end{aligned}$$

So, $R_\epsilon(x)$ is the set of points in X within ϵ of x. Letting $\#A$ be the number of elements in A for any finite set A, DBSCAN works as follows:

1. Initialize $n = 0$ and $N = \emptyset$. n is the index of the cluster we want to make, and N represents the set of noise points.

2. Pick any point $x \in X$ that is not labeled as noise (i.e. not contained in N) and is not in any cluster (i.e. not contained in C_n for any n).

3. If $\#R_\epsilon(x) < \text{minPts}$, add x to N then go to step two, as this means that the density of data points around x is insufficient to consider x clustered. Otherwise, start a new cluster C_n and set $C_n = \{x\}$. Then, let $S = R_\epsilon(x) \setminus \{x\}$ be the set of candidate points for addition to the cluster.

4. For each point $y \in S$, if $\#R_\epsilon(y) < \text{minPts}$, add y to N and remove y from S. Otherwise, we add y to C_n, remove y from S, and add $R_\epsilon(y)$ to S. In other words, if the neighborhood around y is not sufficiently dense, label y as a noise point. Otherwise, add the neighborhood around y to the set of candidate points.

5. Repeat step 4 until S is empty.

6. If $\#C_n < \text{minPts}$, then label all the points in C_n as noise and restart from step 2. Otherwise, the cluster C_n is sufficiently large, so we keep it. If $X \setminus (\mathcal{N} \cup \bigcup_{k=0}^{n} C_k)$ is nonempty, set $n \leftarrow n + 1$ and repeat from step 2. Otherwise, terminate the algorithm.

Note that the algorithm will always terminate in finite time because in each loop, at least one point is labeled as either noise or a clustered point and then never considered again. Also note that in step 4 we never have to consider the case where $R_\epsilon(y)$ contains points from C_k for $k < n$, since any point that is reachable from C_k is always labeled as either an element of C_k or noise. In either case, it is never considered again. DBSCAN is described more succinctly in Algorithm 1 (see supplemental document).

DBSCAN satisfies a few important properties. Specifically, because $x \in R_\epsilon(y)$ if and only if $y \in R_\epsilon(x)$, the same clusters will be formed regardless of how the points in X are ordered (up to a reordering of the clusters). Furthermore, we do not need to specify the number of clusters beforehand, and all of the operations are highly efficient so long as one can efficiently calculate $R_\epsilon(x)$. This makes DBSCAN useful for large, noisy datasets in cases where we have little prior knowledge about the structure of the data.

2.2. Remarks

Note that the choice to use Euclidean distance with DBSCAN is somewhat arbitrary choice. The stability of the algorithm only depends on the fact that our distance function is symmetric, so if we define a notion of density of neighborhoods in terms of another symmetric, non-negative function, the algorithm should maintain its stability. Importantly, the function does not need to satisfy the triangle inequality, so we can work with non-metrics.

2.3. Related Work

The idea of extending DBSCAN to domains where we seek linearity is not entirely new. Previously, an algorithm called ADCN was developed to solve this problem by redefining the search neighborhoods from circles to ellipses whose eccentricity reflect the local covariance of the point (Mai et al., 2016). In practice, ADCN performs as well as DBSCAN in many tasks and performs better in cases where clusters are locally linear in otherwise highly noisy datasets.

However, ADCN is not well-suited for our task in particular because it does not provide the desired separation of obliquely intersecting linear clusters. On the contrary, it can make these intersections more likely, say with a T-shaped intersection, where one cluster can accidentally bisect the other. Furthermore, this process is non-symmetric, meaning that in certain cases the clustering behavior may be incredibly unstable to permutations of the points. LINSCAN, by contrast, solves both of these problems.

3. LINSCAN

3.1. The Embedding

LINSCAN seeks to keep the advantages DBSCAN provides while applying it to the task of distinguishing highly linear clusters. To do this, we embed data points into $\mathbb{P}(\mathbb{R}^d)$, the space of probability measures on $\mathbb{R}^d$, and then cluster the data using a notion of distance between distributions. Letting $\mathrm{eccPts} \in \mathbb{N}$ and defining $R^m(x)$ be the m-nearest neighbors to x in the X, we define a mapping

$$x \in X \mapsto \mathcal{N}\left(\mu_{R^{\mathrm{eccPts}}(x)}, \Sigma_{R^{\mathrm{eccPts}}(x)}\right)$$

Thus, we embed each point in the dataset as the normal distribution best approximating its eccPts-nearest neighbors, which allows us to cluster the points based on the local covariance of the data. To perform clustering in this space, we define a distance function by

$$\begin{aligned} D(P,Q) &= \frac{1}{2}\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F \\ &+ \frac{1}{2}\left\|\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2} - I\right\|_F \\ &+ \frac{1}{\sqrt{2}}\left\|\mu_P - \mu_Q\right\|_{\Sigma_Q^{-1}} \\ &+ \frac{1}{\sqrt{2}}\left\|\mu_P - \mu_Q\right\|_{\Sigma_P^{-1}} \end{aligned}$$

where $P = \mathcal{N}(\mu_P, \Sigma_P)$ and $Q = \mathcal{N}(\mu_Q, \Sigma_Q)$ for positive definite Σ_P and Σ_Q. Note that this function is symmetric and $D(P,Q) = 0$ if and only if $P = Q$. Although D does not satisfy the triangle inequality and is thus not a metric, later we will discuss an approximate form of the triangle inequality which D does satisfy.

While other distance functions exist which have a closed form for Gaussians, D satisfies certain properties which are specifically useful for our purposes. In particular, consider the Wasserstein-2 distance, given by

$$\begin{aligned} W_2(P,Q)^2 &= \|\mu_P - \mu_Q\|_2^2 \\ &+ \mathrm{tr}\left(\Sigma_P + \Sigma_Q - 2\left(\Sigma_Q^{1/2}\Sigma_P\Sigma_Q^{1/2}\right)^{1/2}\right) \end{aligned}$$

While this is a metric, note that the distance between the means and covariances are independent, whereas D punishes differences in mean more heavily in directions orthogonal to the local linearity of the point. Furthermore, the Wasserstein-2 distance punishes differences in covariances at most polynomially in the magnitude of the eigenvalues, whereas D punishes orthogonal covariance inversely to the size of the minimum eigenvalues for high eccentricity clusters, which can grow quite rapidly.

3.2. Motivating the Definition of D

We recall that on a probability space $\mathcal{X}$, the Kullback-Leibler Divergence between two probability measures P and Q with $P \ll Q$ is defined by

$$KL(P|Q) = \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$$

Note the importance of the condition that $P \ll Q$, since by the Radon-Nikodym Theorem $\frac{\mathrm{d}P}{\mathrm{d}Q}$ exists and is finite Q-almost everywhere if and only if $P \ll Q$.

In particular, if $\mathcal{X} = \mathbb{R}^d$ and

$$P = \mathcal{N}(\mu_P, \Sigma_P) \qquad Q = \mathcal{N}(\mu_Q, \Sigma_Q)$$

then

$$KL(P|Q) = \frac{1}{2}\log\frac{|\Sigma_Q|}{|\Sigma_P|} + \frac{1}{2}\operatorname{tr}(\Sigma_Q^{-1}\Sigma_P - I) + \frac{1}{2}(\mu_P - \mu_Q)^T\Sigma_Q^{-1}(\mu_P - \mu_Q)$$

where $|\Sigma|$ is the determinant of Σ.

Next, one can show (see supplemental document) that if

$$\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F < 1$$

then

$$KL(P|Q) = \frac{1}{4}\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F^2 + o\left(\operatorname{tr}\left(\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right)^3\right)\right) + \frac{1}{2}(\mu_P - \mu_Q)^T\Sigma_Q^{-1}(\mu_P - \mu_Q)$$

So, we can define an approximation of $KL(P|Q)$ by

$$M(P|Q) = \frac{1}{4}\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F^2 + \frac{1}{2}(\mu_P - \mu_Q)^T\Sigma_Q^{-1}(\mu_P - \mu_Q)$$

This motivates the symmetric distance function

$$D(P,Q) = \frac{1}{2}\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F + \frac{1}{2}\left\|\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2} - I\right\|_F + \frac{1}{\sqrt{2}}\|\mu_P - \mu_Q\|_{\Sigma_Q^{-1}} + \frac{1}{\sqrt{2}}\|\mu_P - \mu_Q\|_{\Sigma_P^{-1}}$$

obtained by taking the square root term-wise of $M(P|Q)$ and $M(Q|P)$ and then adding them together.

3.3. Local Behavior of D

While D does not satisfy the triangle inequality, one can show that it satisfies a relaxed version, as follows:

Theorem 3.1 *Let $P = \mathcal{N}(\mu_P, \Sigma_P)$, $Q = \mathcal{N}(\mu_Q, \Sigma_Q)$, and $K = \mathcal{N}(\mu_K, \Sigma_K)$, and let $\epsilon > 0$. If $D(P,Q) \le \epsilon$ and $D(Q,K) \le \epsilon$, then*

$$D(P,K) \le D(P,Q) + D(Q,K) + \sqrt{2}\epsilon + \sqrt{2}\epsilon\sqrt{1+\epsilon} + \epsilon^2 + E(P,Q,K)$$

where $E(P,Q,K) = 0$ if Σ_P, Σ_Q, and Σ_K commute.

A proof of this theorem can be found in the supplemental document. Importantly, this shows that for small values of ϵ, D behaves approximately like a metric, which allows us to bound the diameter of any cluster in terms of ϵ and the number of steps between points in the cluster. This ensures that points whose local neighborhoods are nearly orthogonal are not clustered together.

Compare this to the best results proven previously for the approximate triangle inequality of the KL-Divergence between Gaussians in Zhang et al., which was of exponential order (Zhang et al., 2021).

Numerically this bound does not seem to be tight. To see this, Figure 2 plots $D(X,Y) + D(Y,Z)$ on the x axis and $D(X,Z)$ on the y axis for an ensemble of randomly generated Gaussians on $\mathbb{R}^2$. If D satisfied the triangle inequality, one would expect that all the points to be below the line $y = x$. Instead, they seems to lie below a curve of degree between $\frac{3}{2}$ and 2, leading to the conjecture that an upper bound on the order of ϵ^δ for some $\delta \in \left[\frac{3}{2}, 2\right]$ is possible. Since we care about the behavior at small scales, this would be a much stronger statement than the current bound, which is on the order of ϵ.

3.4. The Algorithm

With the previous sections in mind, defining LINSCAN is quite simple. We begin by embedding each point x_i as a multivariate Gaussian P_i by

$$x_i \mapsto P_i := \mathcal{N}\left(\mu_{R^{\mathrm{eccPts}}(x_i)}, \Sigma_{R^{\mathrm{eccPts}}(x_i)}\right)$$

Then, we run DBSCAN on $\mathcal{P} = \{P_i\}$ with Euclidean distance replaced by $D(\cdot,\cdot)$, so that

$$R_\epsilon(P) := \{Q \in \mathcal{P} | D(P,Q) < \epsilon\}$$

and cluster X based on the results. The full process is described in Algorithm 2 (see supplemental document).

4. Numerical Results

In practice, we have discovered that some clusters with sufficiently high spacial density can develop even without being sufficiently linear. To counter this phenomenon, we can apply a linearity condition to the clusters at the end. In the case of $\mathbb{R}^2$ we set a minimal threshold on the correlation of the clusters and discard any clusters which do not meet the threshold.

Figure 3a shows the results of applying LINSCAN to the same data as in Figure 1a, Figure 3b shows the clusters with the noise points removed, and Figure 3c shows the results of removing clusters with correlation less than $\frac{1}{2}$. Note the separation of clusters in the bottom left and top left in comparison to the results from DBSCAN.

Figure 4 shows the results of applying LINSCAN to real data representing seismic activity recorded in Southern California. Not only does LINSCAN identify linear clusters, it is also able to identify them at multiple distinct scales simultaneously.

Figure 5 shows the example of two noisy lines intersecting orthogonally. Note that LINSCAN is the only algorithm which is able to distinguish between the two.

References

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 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*, KDD '96, pp. 226–231. AAAI Press, 1996.

Mai, G., Janowicz, K., Hu, Y., and Gao, S. Adcn: An anisotropic density-based clustering algorithm. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPACIAL '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450345897. doi: 10.1145/2996913.2996940.

Zhang, Y., Liu, W., Chen, Z., Li, K., and Wang, J. On the properties of kullback-leibler divergence between gaussians, 2021. URL https://arxiv.org/abs/2102.05485.

Acknowledgements

I would like to thank my advisor Dr. Alexander Cloninger for his support throughout this project. Further, I would like to thank Dr. Yuri Fialko for bringing to our attention the initial problem from which this project arose.

Figure 2. Triangle Inequality

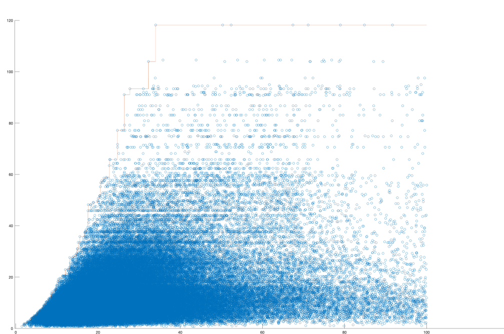
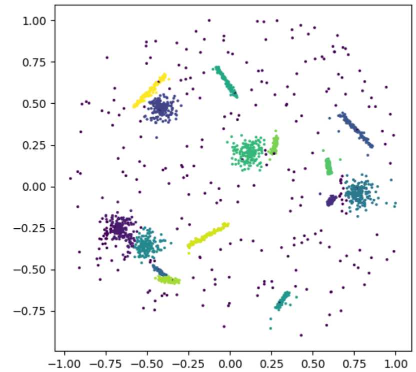
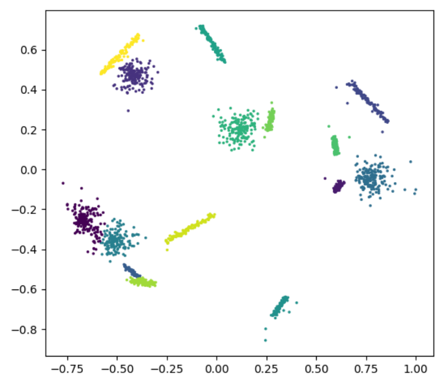


Figure 3. Test Data

(a) LINSCAN Results



(b) LINSCAN Results with Noise Removed



(c) LINSCAN Results with Low-Correlation Clusters Removed

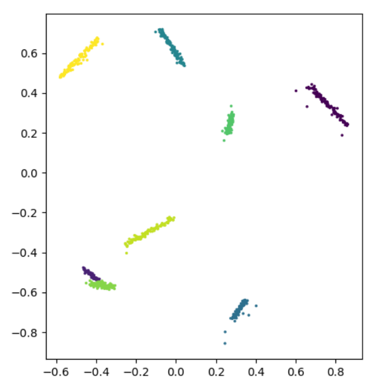
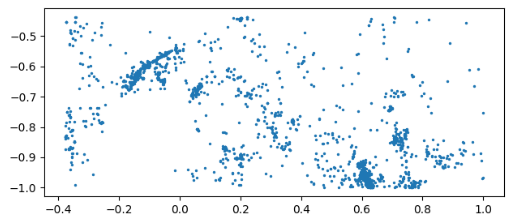
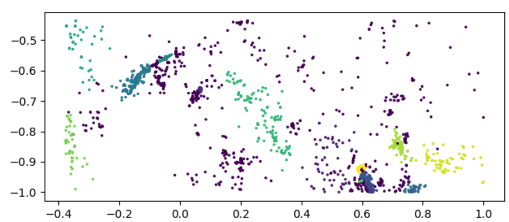


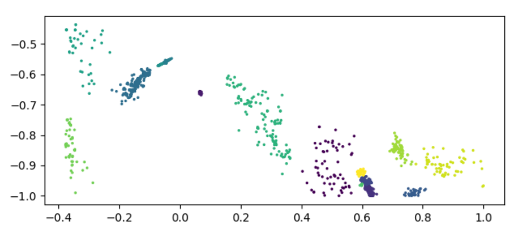
Figure 4. Real Data
(a) Data



(b) LINSCAN results



(c) LINSCAN Results with Noise Removed



(d) LINSCAN Results with Low-Correlation Clusters Removed

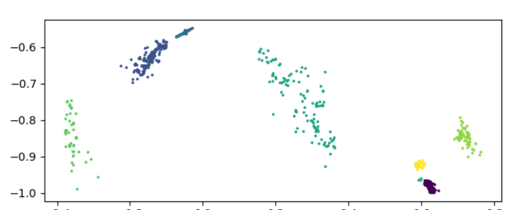
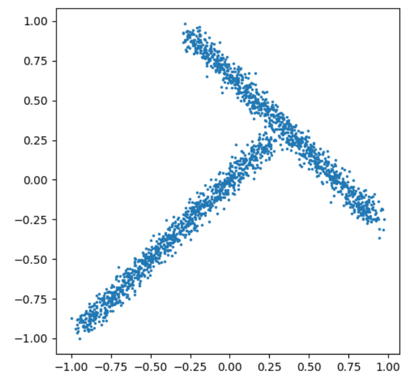
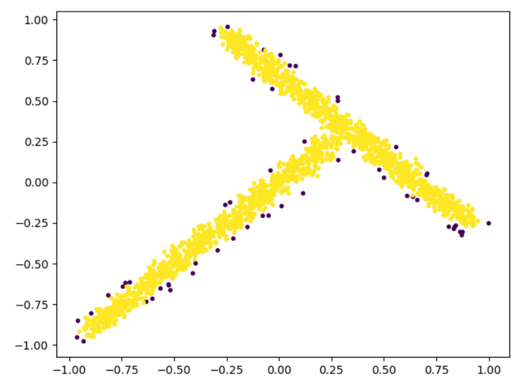


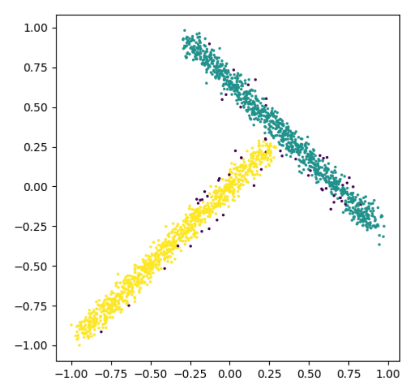
Figure 5. Crossing Lines
(a) Data



(b) DBSCAN results



(c) LINSCAN Results



LINSCAN SUPPLEMENTAL DOCUMENT

1. Approximation of $KL(P|Q)$

First, $\log|A|$ is the logarithm of the product of the eigenvalues of A, which is the same as the sum of the logarithms of the eigenvalues. Therefore,

$$\log|A| = \mathrm{tr}(\log(A))$$

where $\log(A)$ is the matrix logarithm, i.e. the solution B to

$$A = e^B := \sum_{n=0}^{\infty} \frac{B^n}{n!}$$

Such a matrix exists and is unique for any positive definite matrix. In particular, if $A = Q\Lambda Q^T$ for orthogonal Q and diagonal Λ,

$$\log A = Q\log(\Lambda)Q^T$$

where $\log\Lambda$ is the diagonal matrix given by applying the logarithm entrywise to each diagonal entry. Given this,

$$\begin{aligned}\log\frac{|\Sigma_Q|}{\Sigma_P} &= \log|\Sigma_Q| - \log|\Sigma_P| \\ &= \mathrm{tr}\left(\log(\Sigma_Q) - \log(\Sigma_P)\right)\end{aligned}$$

Next, for any positive definite matrices A and B,

$$\mathrm{tr}\left(\log(AB)\right) = \mathrm{tr}(\log(A)) + \mathrm{tr}(\log(B))$$

$$\log(A^{-1}) = -\log(A)$$

Furthermore, if $\|A - I\|_F < 1$, then the sum

$$\sum_{n=1}^{\infty}(-1)^{k+1}\frac{(A-I)^k}{k}$$

converges in $\|\cdot\|_F$ to $\log(A)$. Combining all of this, if $\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\| < 1$ then

$$
\begin{aligned}
&\operatorname{tr}\left(\log(\Sigma_Q) - \log(\Sigma_P)\right) \\
&= -\operatorname{tr}\left(\log\left(\Sigma_Q^{-1/2}\right) + \log\left(\Sigma_P\right) + \log\left(\Sigma_Q^{-1/2}\right)\right) \\
&= -\operatorname{tr}\left(\log\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right)\right) \\
&= -\operatorname{tr}\left(\sum_{k=1}^{\infty}(-1)^{k+1}\frac{\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right)^k}{k}\right) \\
&= -\sum_{k=1}^{\infty}(-1)^{k+1}\frac{\operatorname{tr}\left(\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right)^k\right)}{k} \\
&= -\operatorname{tr}\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right) + \frac{1}{2}\operatorname{tr}\left(\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right)^2\right) + o\left(\operatorname{tr}\left(\left(\Sigma_q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right)^3\right)\right) \\
&= -\operatorname{tr}\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right) + \frac{1}{2}\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F^2 + o\left(\operatorname{tr}\left(\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right)^3\right)\right)
\end{aligned}
$$

where in the last line we used the fact that $\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I$ is symmetric and for any symmetric matrix A

$$\operatorname{tr}(A^2) = \operatorname{tr}(A^T A) = \|A\|_F^2$$

Next, note that

$$\operatorname{tr}(\Sigma_Q^{-1}\Sigma_P - I) = \operatorname{tr}(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I)$$

So, combined with the prior derivations,

$$
\begin{aligned}
&\frac{1}{2}\log\frac{|\Sigma_Q|}{|\Sigma_P|} + \frac{1}{2}\operatorname{tr}(\Sigma_Q^{-1}\Sigma_P - I) \\
&= \frac{1}{4}\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F^2 + o\left(\operatorname{tr}\left(\left(\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right)^3\right)\right)
\end{aligned}
$$

from which the rest of the approximation follows.

2. Proof of Relaxed Triangle Inequality

We recall that

$$
\begin{aligned}
D(P,Q) = &\frac{1}{2}\left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F + \frac{1}{2}\left\|\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2} - I\right\|_F \\
&+ \frac{1}{\sqrt{2}}\|\mu_P - \mu_Q\|_{\Sigma_Q^{-1}} + \frac{1}{\sqrt{2}}\|\mu_P - \mu_Q\|_{\Sigma_P^{-1}}
\end{aligned}
$$

These terms are all nonnegative, so if $D(P,Q) \le \epsilon$ then each term is at most ϵ. To show the relaxed triangle inequality, we define

$$D_1(P,Q) := \left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F + \left\|\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2} - I\right\|_F$$

and

$$D_2(P,Q) := \|\mu_P - \mu_Q\|_{\Sigma_Q^{-1}} + \|\mu_P - \mu_Q\|_{\Sigma_P^{-1}}$$

so that

$$D(P,Q) = \frac{1}{2}D_1(P,Q) + \frac{1}{\sqrt{2}}D_2(P,Q)$$

Then,

$$\begin{aligned}
&D_2(P,K)\\
&= \|\mu_P - \mu_K\|_{\Sigma_K^{-1}} + \|\mu_P - \mu_K\|_{\Sigma_P^{-1}}\\
&\leq \|\mu_P - \mu_Q\|_{\Sigma_K^{-1}} + \|\mu_Q - \mu_K\|_{\Sigma_K^{-1}} + \|\mu_P - \mu_Q\|_{\Sigma_P^{-1}} + \|\mu_Q - \mu_K\|_{\Sigma_P^{-1}}\\
&= D_2(P,Q) + D_2(Q,K) + \|\mu_P - \mu_Q\|_{\Sigma_K^{-1}} - \|\mu_P - \mu_Q\|_{\Sigma_Q^{-1}} + \|\mu_Q - \mu_K\|_{\Sigma_P^{-1}} - \|\mu_Q - \mu_K\|_{\Sigma_Q^{-1}}
\end{aligned}$$

Note that

$$\begin{aligned}
&\|\mu_P - \mu_Q\|_{\Sigma_K^{-1}} - \|\mu_P - \mu_Q\|_{\Sigma_Q^{-1}}\\
&= \left\|\Sigma_K^{-1/2}(\mu_P - \mu_Q)\right\|_2 - \left\|\Sigma_Q^{-1/2}(\mu_P - \mu_Q)\right\|_2\\
&\leq \left\|\Sigma_K^{-1/2}(\mu_P - \mu_Q) - \Sigma_Q^{-1/2}(\mu_P - \mu_Q)\right\|_2\\
&= \left\|\left(\Sigma_K^{-1/2} - \Sigma_Q^{-1/2}\right)(\mu_P - \mu_Q)\right\|_2\\
&= \left\|\left(\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right)\Sigma_Q^{-1/2}(\mu_P - \mu_Q)\right\|_2\\
&\leq \left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2 \left\|\Sigma_Q^{-1/2}(\mu_P - \mu_Q)\right\|_2\\
&= \left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2 \|\mu_P - \mu_Q\|_{\Sigma_Q^{-1}}\\
&\leq \left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2 \epsilon
\end{aligned}$$

Now, note that $\left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2$ is the square root of the maximal eigenvalue of

$$(\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I)^T(\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I)$$

Therefore,

$$\begin{aligned}
&\left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2\\
&= \sqrt{\left\|(\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I)^T(\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I)\right\|_2}\\
&= \sqrt{\left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - \Sigma_K^{-1/2}\Sigma_Q^{1/2} - \Sigma_Q^{1/2}\Sigma_K^{-1/2} + I\right\|_2}\\
&\leq \sqrt{\left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - I\right\|_2 + \left\|2I - \Sigma_K^{-1/2}\Sigma_Q^{1/2} - \Sigma_Q^{1/2}\Sigma_K^{-1/2}\right\|_2}\\
&\leq \sqrt{\left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - I\right\|_2 + \left\|I - \Sigma_K^{-1/2}\Sigma_Q^{1/2}\right\|_2 + \left\|I - \Sigma_Q^{1/2}\Sigma_K^{-1/2}\right\|_2}\\
&= \sqrt{\left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - I\right\|_2 + 2\left\|I - \Sigma_K^{-1/2}\Sigma_Q^{1/2}\right\|_2}\\
&= \sqrt{\left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - I\right\|_2 + 2\left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2}
\end{aligned}$$

Solving this for $\left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2$, we get

$$\left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2 \leq 1+\sqrt{1 + \left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - I\right\|_2} \leq 1+\sqrt{1 + \left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - I\right\|_F} \leq 1+\sqrt{1+\epsilon}$$

So,

$$\|\mu_P - \mu_Q\|_{\Sigma_K^{-1}} - \|\mu_P - \mu_Q\|_{\Sigma_Q^{-1}} \leq \left\|\Sigma_K^{-1/2}\Sigma_Q^{1/2} - I\right\|_2 \epsilon \leq \epsilon + \epsilon\sqrt{1+\epsilon}$$

A similar statement holds for $\|\mu_Q - \mu_K\|_{\Sigma_P^{-1}} - \|\mu_Q - \mu_K\|_{\Sigma_Q^{-1}}$, so

$$D_2(P,K) \le D_2(P,Q) + D_2(Q,K) + 2\epsilon + 2\epsilon\sqrt{1+\epsilon}$$

Next,

$$\begin{aligned}
&\left\|\Sigma_P^{-1/2}\Sigma_K\Sigma_P^{-1/2} - I\right\|_F - \left\|\Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2} - I\right\|_F - \left\|\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2} - I\right\|_F \\
&\le \left\|\Sigma_P^{-1/2}\Sigma_K\Sigma_P^{-1/2} - \Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2}\right\|_F - \left\|\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2} - I\right\|_F \\
&\le \left\|\Sigma_P^{-1/2}\Sigma_K\Sigma_P^{-1/2} - \Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2} - \Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2} + I\right\|_F \\
&= \left\|\left(I - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right)\left(I - \Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right) + \Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&\le \left\|I - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right\|_F\left\|I - \Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F + \left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&\le \epsilon^2 + \left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F
\end{aligned}$$

A similar argument shows

$$\begin{aligned}
&\left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - I\right\|_F - \left\|\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2} - I\right\|_F - \left\|\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - I\right\|_F \\
&\le \epsilon^2 + \left\|\Sigma_P^{-1/2}\Sigma_K\Sigma_P^{-1/2} - \Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2}\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2}\right\|_F
\end{aligned}$$

Combining these,

$$\begin{aligned}
D_1(P,K) \le D_1(P,Q) + D_1(Q,K) + 2\epsilon^2 \\
+ \left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
+ \left\|\Sigma_P^{-1/2}\Sigma_K\Sigma_P^{-1/2} - \Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2}\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2}\right\|_F
\end{aligned}$$

Next, if $[A,B] = AB - BA$ is the commutator of A and B,

$$\begin{aligned}
&\left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&\le \left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - \Sigma_K^{-1/2}\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&\quad + \left\|\Sigma_K^{-1/2}\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_K^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&= \left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_K^{-1/2} - \Sigma_K^{-1/2}\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&\quad + \left\|\left[\Sigma_K^{-1/2}, \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right]\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&= \left\|\Sigma_K^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_Q^{-1/2}\Sigma_Q\Sigma_K^{-1/2} - \Sigma_K^{-1/2}\Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&\quad + \left\|\left[\Sigma_K^{-1/2}, \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right]\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&= \left\|\Sigma_K^{-1/2}\left[\Sigma_P, \Sigma_Q^{-1/2}\right]\Sigma_Q^{-1/2}\Sigma_Q\Sigma_K^{-1/2}\right\|_F + \left\|\left[\Sigma_K^{-1/2}, \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right]\Sigma_Q\Sigma_K^{-1/2}\right\|_F \\
&= \left\|\Sigma_K^{-1/2}\left[\Sigma_P, \Sigma_Q^{-1/2}\right]\Sigma_Q^{1/2}\Sigma_K^{-1/2}\right\|_F + \left\|\left[\Sigma_K^{-1/2}, \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right]\Sigma_Q\Sigma_K^{-1/2}\right\|_F
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left\|\Sigma_P^{-1/2}\Sigma_K\Sigma_P^{-1/2} - \Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2}\Sigma_P^{-1/2}\Sigma_Q\Sigma_P^{-1/2}\right\|_F \\
&\le \left\|\Sigma_P^{-1/2}\left[\Sigma_K, \Sigma_Q^{-1/2}\right]\Sigma_Q^{1/2}\Sigma_P^{-1/2}\right\|_F + \left\|\left[\Sigma_P^{-1/2}, \Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2}\right]\Sigma_Q\Sigma_P^{-1/2}\right\|_F
\end{aligned}$$

So finally, if we let

$$E(P,Q,K) := \frac{1}{2}\left\|\Sigma_K^{-1/2}\left[\Sigma_P, \Sigma_Q^{-1/2}\right]\Sigma_Q^{1/2}\Sigma_K^{-1/2}\right\|_F + \frac{1}{2}\left\|\left[\Sigma_K^{-1/2}, \Sigma_Q^{-1/2}\Sigma_P\Sigma_Q^{-1/2}\right]\Sigma_Q\Sigma_K^{-1/2}\right\|_F$$
$$+ \frac{1}{2}\left\|\Sigma_P^{-1/2}\left[\Sigma_K, \Sigma_Q^{-1/2}\right]\Sigma_Q^{1/2}\Sigma_P^{-1/2}\right\|_F + \frac{1}{2}\left\|\left[\Sigma_P^{-1/2}, \Sigma_Q^{-1/2}\Sigma_K\Sigma_Q^{-1/2}\right]\Sigma_Q\Sigma_P^{-1/2}\right\|_F$$

then the theorem follows.

Note we could have defined E in terms of the original Frobenius norm terms and the theorem would have still held, since if the three matrices commute, those terms would also be 0. However, rewriting the bound in terms of commutators clarifies the relationship between the size of E and the degree to which the matrices commute. Furthermore, in the future this form may be more useful in bounding E for matrices which have similar eigenvectors.

3. Algorithms

Algorithm 1 DBSCAN

Input: Data $X = \{x_1, ..., x_m\}$, $\epsilon > 0$, minPts $\in \mathbb{N}$
Output: Clusters $\{C_k\}$
$n \leftarrow 0$
$N \leftarrow \emptyset$
while $X \setminus (N \cup \bigcup_{k=0}^{n-1} C_k) \neq \emptyset$ **do**
  Pick $x \in X \setminus (N \cup \bigcup_{k=0}^{n-1} C_k)$
  if $\#R_\epsilon(x) <$ minPts **then**
    $N \leftarrow N \cup \{x\}$
  else
    $C_n \leftarrow \{x\}$
    $S \leftarrow R_\epsilon(x) \setminus (N \cup \{x\})$
    while $S \neq \emptyset$ **do**
      Pick $y \in S$
      if $\#R_\epsilon(y) <$ minPts **then**
        $N \leftarrow N \cup \{y\}$
        $S \leftarrow S \setminus \{y\}$
      else
        $C_n \leftarrow C_n \cup \{y\}$
        $S \leftarrow (S \cup R_\epsilon(y)) \setminus (N \cup C_n)$
      end if
    end while
    if $\#C_n <$ minPts **then**
      $N \leftarrow N \cup C_n$
      $C_n \leftarrow \emptyset$
    else
      $n \leftarrow n + 1$
    end if
  end if
end while

Algorithm 2 LINSCAN

Input: Data $X = \{x_1, ..., x_m\}$, $\epsilon > 0$, minPts $\in \mathbb{N}$, $n = 0$, $N = \emptyset$, eccPts $\in \mathbb{N}$
Output: Clusters $\{C_k\}$
$n \leftarrow 0$
$N \leftarrow \emptyset$
$\mathcal{P} \leftarrow \emptyset$
for $x \in X$ **do**
  $\mu \leftarrow \mu_{R^{\text{eccPts}}(x)}$
  $\Sigma \leftarrow \Sigma_{R^{\text{eccPts}}(x)}$
  $P \leftarrow \mathcal{N}(\mu, \Sigma)$
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{P\}$
end for
$\{D_k\} \leftarrow \text{DBSCAN}(\mathcal{P}, \epsilon, \text{minPts})$
for $k \in \{0, 1, ..., n\}$ **do**
  $C_k \leftarrow \{x_i \in X : P_i \in D_k\}$
end for