

Clustering of Distributions on 1-Dimensional Manifolds

Setup

Suppose we have two manifolds \mathcal{M}_1 and $\mathcal{M}_2 \in \mathbb{R}^d$, each of length 1, defined by $f_1(t)$ and $f_2(t)$ respectively ($f_k : [0, 1] \mapsto \mathbb{R}^d$). Define δ as the minimum distance between any two manifolds, i.e., $\delta = \max_{k,\ell} \min_{s,t} \|f_k(s) - f_\ell(t)\|$, and let $\delta > 0$. Restrict each f_k such that the distance along the manifold between $f_k(t)$ and $f_k(s)$ is equal to the difference between t and s , i.e., f_k is the arclength parameterization of \mathcal{M}_k (this also implies that each manifold is of length 1). Labels are sampled as $Z_1, \dots, Z_n \stackrel{\text{iid}}{\sim} \text{Multinomial}(\alpha_1, \alpha_2)$, and we define n_k as the number of labels with label k . Sample $T_1, \dots, T_n \stackrel{\text{iid}}{\sim} F$ for some distribution F with support $[0, 1]$. Finally, let $X_i = f_{Z_i}(T_i)$ for each $i \in [n]$ be the latent vector of vertex i , and gather the latent vectors as $X = \begin{bmatrix} X_1 & \dots & X_n \end{bmatrix}$. We observe $A \sim \text{RDPG}(X)$ (or $A \sim \text{GRDPG}_{p,q}(X)$) and wish to recover Z_1, \dots, Z_n .

Preliminary Theory

Distributions of differences of order statistics

Let $D_i = X_{(i+1)} - X_{(i)}$. Then if $\max_i D_i < \delta$, we have sufficient separation of points in \mathcal{M}_1 . Then it is sufficient to quantify $P(\max_i D_i > \delta)$ as a function of n and δ and show that this converges to zero as n grows to ∞ .

We denote $f(x)$ as the density of each X_i , $g_i(x)$ as the density of $X_{(i)}$, $g_{ij}(x, y)$ as the joint density of $X_{(i)}, X_{(j)}$, and $h_i(d)$ as the density of D_i (with corresponding capital letters for the cumulative distribution functions).

The following are taken as given¹:

1. $g_i(x) = \frac{n!}{(n-i)!(i-1)!} (F(x))^{i-1} (1 - F(x))^{n-i} f(x)$.
2. $g_{ij}(x, y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} (F(x))^{i-1} (F(y) - F(x))^{j-i-1} (1 - F(y))^{n-j} f(x) f(y)$.
3. By convolution, $h_i(d) = \int_0^1 g_{i,i+1}(x, x+d) dx$.

Lemma 1 (The probability density function of D_i).

$$h_i(d) = \int_0^{1-d} \frac{n!}{(i-1)!(n-i-1)!} (F(x))^{i-1} (1 - F(x+d))^{n-i-1} f(x) f(x+d) dx \quad (1)$$

Proof. This is just a direct consequence of 2 and 3 under the given statements. We also note that because the support of X_i is $[0, 1]$, the integral only needs to be evaluated from 0 to $1 - d$ because of the $f(x+d)$ and $1 - F(x+d)$ terms. \square

¹https://en.wikipedia.org/wiki/Order_statistic

Lemma 2 (The cumulative distribution function of D_i).

$$P(D_i < \delta) = H_i(\delta) = 1 - \int_0^{1-\delta} \frac{n!}{(n-i)!(i-1)!} (F(x))^{i-1} (1 - F(x+\delta))^{n-i} f(x) dx \quad (2)$$

Proof.

$$\begin{aligned} H_i(\delta) &= \int_x^{x+\delta} h_i(d) dd \\ &= \int_x^{x+\delta} \int_0^1 \frac{n!}{(i-1)!(n-i-1)!} ((F(x))^{i-1} (1 - F(x+d))^{n-i-1} f(x) f(x+d)) dx dd \\ &= \int_0^1 \frac{n!}{(i-1)!(n-i-1)!} (F(x))^{i-1} f(x) \int_x^{x+\delta} (1 - F(x+d))^{n-i-1} f(x+d) dd dx \\ &= \int_0^1 \frac{n!}{(i-1)!(n-i-1)!} (F(x))^{i-1} f(x) \int_{F(x)}^{F(x+\delta)} (1-u)^{n-i-1} du dx \\ &= \int_0^1 \frac{n!}{(i-1)!(n-i)!} (F(x))^{i-1} f(x) ((1 - F(x))^{n-i} - (1 - F(x+\delta))^{n-i}) dx \\ &= \int_0^1 g_i(x) dx - \int_0^1 \frac{n!}{(i-1)!(n-i)!} (F(x))^{i-1} (1 - F(x+\delta))^{n-i} f(x) dx \\ &= 1 - \int_0^1 \frac{n!}{(i-1)!(n-i)!} (F(x))^{i-1} (1 - F(x+\delta))^{n-i} f(x) dx \end{aligned}$$

Because of the $x + \delta$ term, we can't actually evaluate this integral all the way up to 1, and so we are left with

$$= 1 - \int_0^{1-\delta} \frac{n!}{(i-1)!(n-i)!} (F(x))^{i-1} (1 - F(x+\delta))^{n-i} f(x) dx.$$

□

Uniform case

Lemma 3 (Differences between order statistics of a uniform distribution). *If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1)$, then each $D_i \sim \text{Beta}(1, n)$.*

Proof. We begin with Eq. (1), plugging in $f(x) = 1$ and $F(x) = x$:

$$h_i(d) = \int_0^{1-d} \frac{n!}{(i-1)!(n-i-1)!} x^{i-1} (1-x-d)^{n-i-1} dx$$

Then we proceed with integration by parts, setting $u = x^{i-1} \implies du = (i-1)x^{i-2}$ and $dv = (1-x-d)^{n-i-1} dx \implies v = -\frac{1}{n-i}(1-x-d)^{n-i}$. Note that $uv|_0^{1-d} = 0$ in this case. This yields

$$= \frac{n!}{(i-1)!(n-i-1)!} \int \frac{i-1}{n-i} x^{i-2} (1-x-d)^{n-i} dx$$

Then applying integration by parts again until the x^p term disappears, we get:

$$\begin{aligned}
&= \frac{n!}{(i-1)!(n-i-1)!(n-i)\cdots(n-2)} \int_0^{1-d} (1-x-d)^{n-2} dx \\
&= -\frac{n(n-1)}{n-1} (1-x-d)^{n-1} \Big|_0^{1-d} \\
&= n(1-d)^{n-1}
\end{aligned}$$

This the density function for $\text{Beta}(1, n)$, completing the proof. \square

Theorem 1. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1)$. Then for any ϵ and $\delta > 0$, there exists an $N = O(\frac{-\log \epsilon}{\delta})$ such that $P(\max_i X_{(i+1)} - X_{(i)} < \delta) \geq 1 - \epsilon$ when $n > N$.

Proof (sketch). Since $X_{(i+1)} - X_{(i)} = D_i \sim \text{Beta}(1, n)$, $P(X_{(i+1)} - X_{(i)} < \delta) = 1 - (1 - \delta)^n$. This yields

$$\begin{aligned}
P(\max_i D_i < \delta) &\geq (P(D_i < \delta))^{n-1} \\
&= (1 - (1 - \delta)^n)^{n-1} \\
&\approx e^{-n \exp(-n\delta)}.
\end{aligned}$$

In the limit $n \rightarrow \infty$, this goes to 1. \square

Test for intersecting manifolds

Theorem 2. Let \mathcal{M}_1 and \mathcal{M}_2 be two one-dimensional manifolds of unit length with $n/2$ points drawn uniformly along each manifold.

Suppose we want to perform a hypothesis test for whether the ends of the manifolds meet. The H_0 is the case where $\delta = 0$ and H_1 is the case where $\delta > 0$. We define the test statistic T_n as follows: Construct an ϵ -neighborhood graph using the smallest ϵ required to make the graph connected. Let $\{d_i\}$ be the set of edge lengths of this graph. Define $T_n = \max_i d_i$. Then $T_n/2 \xrightarrow{d} \text{Beta}(1, n)$ under H_0 .

This theorem isn't 100% correct—in reality, we want to draw edges along the “super-manifold” that is created by joining the two manifolds (under H_0). But if H_0 is true, then these lengths should converge to the true distances along the super-manifold.

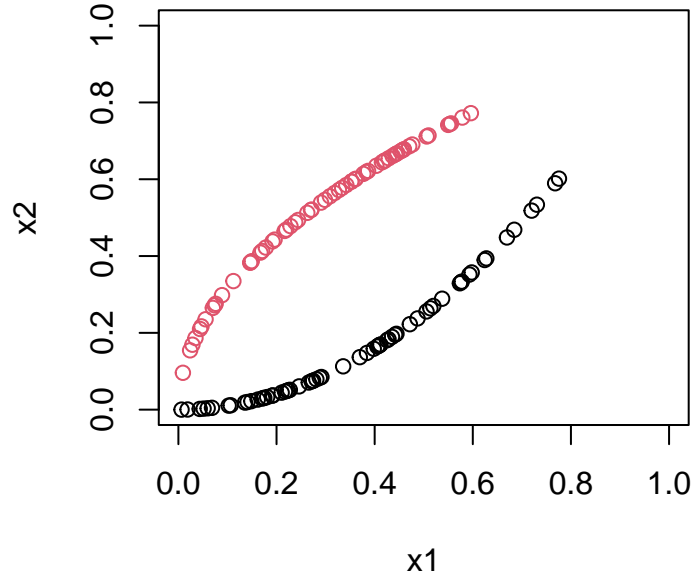
Community detection for “non-affine” one-dimensional manifolds

For non-intersecting manifolds, distance-based graph techniques provide a straightforward method for community detection. However, this is not the case in many models with community structure. For example, the latent structure that induces the degree-corrected block model consists of line segments that emanate from the origin, so each community is a one-dimensional subspace, or linear manifold, and these manifolds meet at the origin. For the DCBM, a straightforward approach to community detection involves an ASE followed by clustering based on angles (e.g., cosine similarity). As an extension to this case, we consider community detection for two one-dimensional manifolds that meet at the origin.

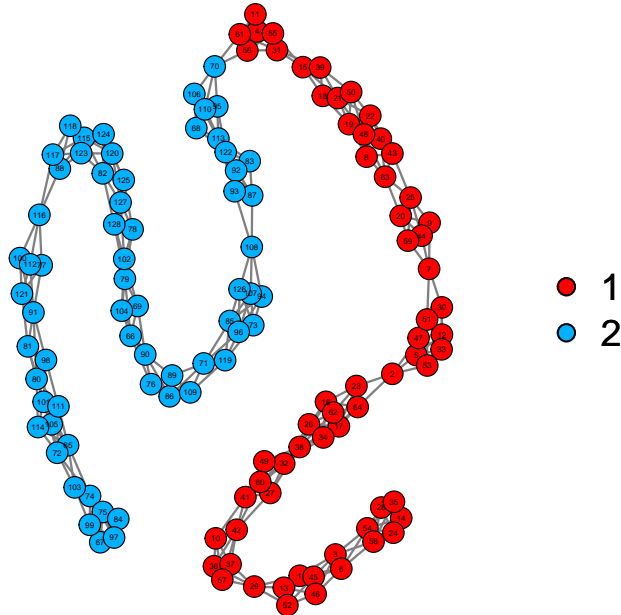
The method that is proposed is as follows:

1. For each point x_i , compute $\|x_i\|$.
2. Construct an ϵ -neighborhood graph for X , choosing ϵ large enough that the graph is connected. Alternatively, construct a K -NN graph, again, choosing K so that the graph is connected.
3. Remove the vertex v_j that corresponds to the point $x_j = \min_i \|x_i\|$. If this results in two disconnected subgraphs, we are done. Otherwise, continue removing vertices in ascending order of magnitude until the graph is disconnected. Then map each component to a community.
4. Estimate curves for each community (e.g., using polynomial regression). Map each removed point to a community based on its distance to the curve.

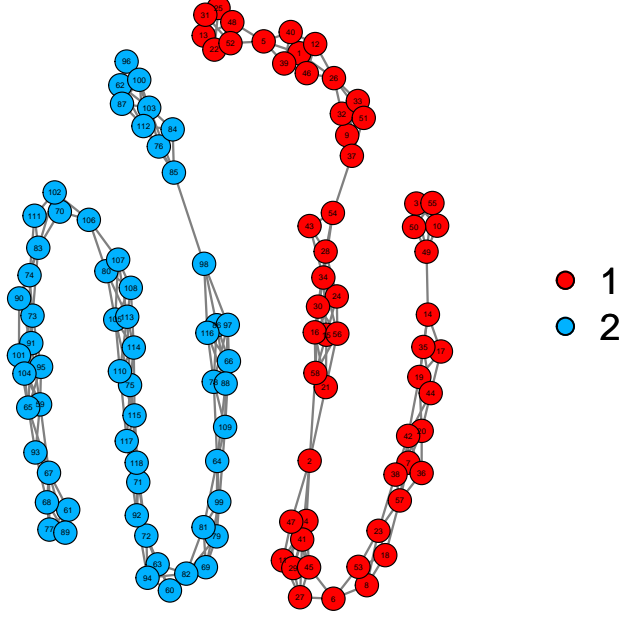
In the following example, we draw 64 points each uniformly along two manifolds in \mathbb{R}^2 that intersect at the origin:



The 5-NN graph is as follows:



Removing the 10 points closest to the origin results in the following:



General case

Theorem 3. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ with support $[0, 1]$, and suppose $f(x)$ is continuous and $f(x) \geq a > 0$ everywhere on the support. Let $D_i = X_{(i+1)} - X_{(i)}$. Then for any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $P(\max_i D_i < \delta) \geq 1 - \epsilon$ when $n > N$.

Proof (sketch). We start with Eq. (2):

$$P(D_i \leq \delta) = 1 - \int_0^{1-\delta} \frac{n!}{(n-i)!(i-1)!} (F(x))^{i-1} (1 - F(x + \delta))^{n-i} f(x) dx.$$

Making the approximation $F(x + \delta) \approx F(x) + \delta f(x)$ and bounding $f(x) \geq a$, we get:

$$P(D_i \leq \delta) \geq 1 - \int_0^{1-\delta} \frac{n!}{(n-i)!(i-1)!} (F(x))^{i-1} (1 - F(x) - a\delta)^{n-i} f(x) dx.$$

Then making the substitution $u = F(x) \implies du = f(x) dx$, we obtain

$$1 - \int_0^{F(1-\delta)} \frac{n!}{(n-i)!(i-1)!} u^{i-1} (1 - u - a\delta)^{n-i} du$$

Evaluating the integral yields

$$P(D_i < \delta) = 1 - (1 - a\delta)^n + (1 - F(1 - \delta) - a\delta)^n.$$

Then as before,

$$\begin{aligned}
P(\max_i D_i < \delta) &= P(\text{all } D_i < \delta) \\
&= 1 - P(\text{some } D_i > \delta) \\
&\geq 1 - \sum_i^{n-1} P(D_i > \delta) \\
&= 1 - (n-1)(1 - a\delta)^n + (n-1)(1 - F(1 - \delta) - a\delta)^n
\end{aligned}$$

This converges to 1 in the limit $n \rightarrow \infty$.

We can also approximate $F(1 - \delta) \approx 1 - a\delta$, which yields $1 - (n-1)(1 - a\delta)^n$. Setting this $\geq 1 - \epsilon$... \square

Extension to multidimensional manifolds

Here, we extend the results on the line to the unit hypercube. The following theorem is a direct consequence of Lemma 2 of Trosset and Buyukbas [1].

Theorem 4. *Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ with support $[0, 1]^r$, and $f(x) \geq a > 0$ everywhere on the support. Define E_n as the event that an η -neighborhood graph constructed from the sample is connected. Then for any $\epsilon > 0$, there exists $N = O\left(\frac{\log \epsilon \eta^r / r^{r/2}}{\log(1 - \frac{a\eta^r}{r^{r/2}})}\right)$ such that $P(E_n) > 1 - \epsilon$ when $n \geq N$.*

Proof (sketch). Divide the hypercube $[0, 1]^r$ into a grid of sub-hypercubes of side length at most η/\sqrt{r} . E_n is satisfied if each sub-hypercube contains at least one X_i from the sample.

$$\begin{aligned}
P(E_n) &= 1 - P(\text{some cells don't contain } X_i) \\
&\geq 1 - \sum_k^{\lceil \sqrt{r}/\eta \rceil^r} \prod_i^n P(X_i \text{ is not in the } k^{\text{th}} \text{ hypercube}) \\
&\geq 1 - \lceil \sqrt{r}/\eta \rceil^r (1 - a\eta^r / r^{r/2})^n
\end{aligned}$$

Setting this $\geq 1 - \epsilon$ and solving for n yields the desired rate. \square

Algorithms

TBD

Computational Results

TBD

References

- [1] Michael W. Trosset and Gokcen Buyukbas. Rehabilitating isomap: Euclidean representation of geodesic structure, 2020.