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 ($f_i : [0,1] \mapsto \mathbb{R}^d$). Define δ as the minimum distance between the two manifolds, i.e., $\delta = \min_t \|f_1(t) - f_2(t)\|$, and let $\delta > 0$. For now, restrict each f_i such that the distance along the manifold between $f_i(t)$ and $f_i(s)$ is equal to the difference between t and s (this also implies that each manifold is of length 1). We sample $T_1, ..., T_n \overset{iid}{\sim} F$ for continuous F with support $[0,1]$ and use f_1 to map the first n_1 points to $\mathcal{M}_1$ and f_2 to map the remaining $n_2 = n - n_1$ points to $\mathcal{M}_2$. Let $X_i = f_1(T_i)$ and $Y_j = f_2(T_j)$. Without loss of generality, assume $n_1 \leq n_2$.

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 given[1]:

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

Lemma 1 (The probability density function of D_i)**.**

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

Proof. This is just given by setting the integrand as $g_{i,i+1}(x, x+d)$. We briefly note that the integral can only be evaluated from 0 to $1 - d$ in this case because of the $x + d$ terms. $\square$

Lemma 2 (The cumulative distribution function of D_i)**.**

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

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

Proof.

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

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

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

□

Thus if we are able to show that Eq. (2) converges to 1, we are done.

Uniform case

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

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

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

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

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

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

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

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

Theorem 1. *Let $X_1,...,X_n \overset{\text{iid}}{\sim} \mathrm{Uniform}(0,1)$. Then for any $\delta > 0$, $P(\max_i X_{(i+1)} - X_{(i)} < \delta) \to 1$ as $n \to \infty$.*

Proof (sketch). Since $X_{(i+1)} - X_{(i)} = D_i \sim \mathrm{Beta}(1,n)$, $P(X_{(i+1)} - X_{(i)} < \delta) = 1-(1-\delta)^n$. 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 $n \to \infty$, this goes to 1. □

General case

Theorem 2. *Let $X_1,...,X_n \overset{\text{iid}}{\sim} F$ 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 $P(\max_i D_i < \delta) > 1-\epsilon$ when $n > N$.*

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

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

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

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

Then making the substitution $u = F(x) \implies du = f(x)\mathrm{d}x$, we obtain

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

Evaluating the integral yields

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

Then as before,

$$
\begin{aligned}
P(\max_i D_i < \delta) &\geq (P(D_i < \delta))^{n-1} \\
&\approx (1 - (1 - a\delta)^n + (1 - F(1-\delta) - a\delta)^n)^{n-1} \\
&\to 1
\end{aligned}
$$

$\square$

If F is absolutely continuous, then it can be shown that

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

Making the approximation $F(x+\delta) \approx F(x) + \delta f(x)$ and bounding $f(x) \geq a > 0$, we get:

$$P(D_i \leq \delta) \geq 1 - \int_0^{1-\delta} \frac{n!}{(n-i+1)!(i-2)!}(F(x))^{i-2}(1 - F(x) - a\delta)^{n-i} f(x)dx$$

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

$$1 - \int_0^{F(1-\delta)} \frac{n!}{(n-i+1)!(i-2)!} u^{i-2}(1 - u - a\delta)^{n-i} du$$

Which becomes $1 - CP(n)\delta^n$ for some $C > 0$ and $P(n)$ is a polynomial of n, giving us the desired result $D_i \xrightarrow{p} 0$.

Algorithms

TBD

Computational Results

TBD