Keio University

CALAMVS GLADIO FORTIOR 1858

Statistical analysis by tuning curvature of data spaces

Kei Kobayashi (Keio University, JST PRESTO)

16/Aug/2016 @RIMS

Joint work with Henry P. Wynn (London School of Economics)

arXiv:1401.3020 [math.ST]

Keio University

CALAMVS GLADIO FORTIOR 1858

Outline

Mean (Center) of Population: Japan

Question: Where is the mean (center) of the population of Japan?

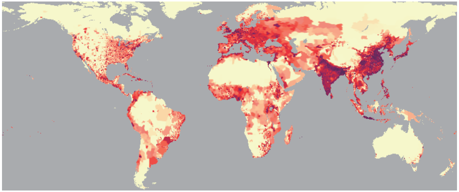
Set every person's coordinate (x_i, y_i) for $i = 1, \ldots, N$, then the mean is

$$(\bar{x}, \bar{y}) = \left(\frac{1}{N}\sum_i x_i, \frac{1}{N}\sum_i y_i\right).$$

$=$ Seki City in Gifu prefecture
(in 2010, after some modification, [Wikipedia])

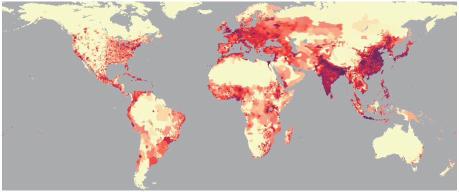
Mean (Center) of Population: World

Question: Where is the mean (center) of the population of the world?

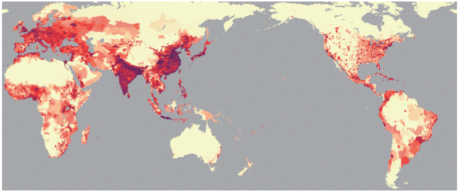


Mean (Center) of Population: World

Question: Where is the mean (center) of the population of the world?

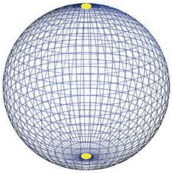


$(\bar{x}, \bar{y}) = \left(\frac{1}{N}\sum_i x_i, \frac{1}{N}\sum_i y_i\right)$ does not make sense:



Mean of the North Pole and the South Pole

Question: Where is the mean of two samples at the north pole and the south pole of a sphere?



The center of the sphere
= the mean on the embedding space (Euclidean space)
But NOT on the sphere

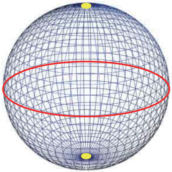
We want the "mean" ON a sphere

The First Candidate: Intrinsic Mean (Fréchet Mean)

Intrinsic mean on a unit sphere:

$$\hat{\mu} = \arg\min_{m \in S^2} \sum_i d(x_i, m)^2$$

where $d(\cdot, \cdot)$ is a geodesic distance(shortest path length) on a sphere.



Every points on the equator attains the minimum.
Intrinsic mean is not necessarily unique.

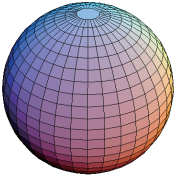
The Second Candidate: Extrinsic Mean

Extrinsic mean on a unit sphere:

$$\hat{\mu} = \arg\min_{m \in S^2} \sum_i \|x_i - m\|^2.$$

Remember the original "outer" mean is

$$\hat{\mu} = \arg\min_{m \in E^3} \sum_i \|x_i - m\|^2.$$



Every points on the sphere attain the minimum.
Extrinsic mean is again not necessarily unique.

Example: Sphere

On a restricted region $A \subset S^2$, the intrinsic mean is unique regardless of the population (empirical) distribution iff the diameter of A is less than $\pi/2$ [Kendall, W.S. 1990]

So in the five continents, only Eurasia can have multiple means.
(e.g. d(Madrid, Singapore) $=$ 11400km $>$ 10000km)

Similar theory holds for metric spaces of positive curvature and CAT(k) spaces.

Example: Euclidean Space

For a Euclidean space, the intrinsic mean is unique since $f_i(m) = \|m - x_i\|^2$ is strictly convex, thus $f(m) = \sum_i \|m - x_i\|^2$ is strictly convex and has the unique minimum.

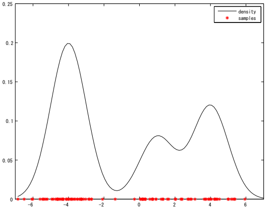
It is easy to see the intrinsic mean is equal to $\bar{x}$.

The L_γ-mean $\arg\min_m \sum_i \|m - x_i\|^\gamma$ for $\gamma \geq 1$ is also unique.

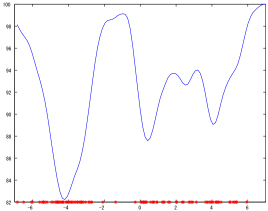
HOWEVER, the uniqueness of the means is sometimes unwelcome.

Clustering

Local minima of the Fréchet function (sometimes called Karcher means) can be used for clustering.



density function and its samples



$f(m)$

However, for clustering, Euclidean space is TOO FLAT.
i.e. curvature of Euclidean metric is so small that f cannot have multiple local minima.

Family of metrics for data analysis

- Ordinary data analysis (e.g. classification, regression):

  Data X_i $(i = 1, \ldots, n)$, Metric d

  $\longrightarrow$ Loss function $\hat{f} \in \mathcal{F}$

  (can be selected by cross validation, resampling)

  $$\longrightarrow \hat{\theta} = \arg\min \sum_i \hat{f}(d(X_i, \theta))$$

- Our approach:

  Data X_i $(i = 1, \ldots, n)$, Loss function f

  $\longrightarrow$ Metric $\hat{d} \in \mathcal{D}$

  (can be selected by cross validation, resampling)

  $$\longrightarrow \hat{\theta} = \arg\min \sum_i f(\hat{d}(X_i, \theta))$$

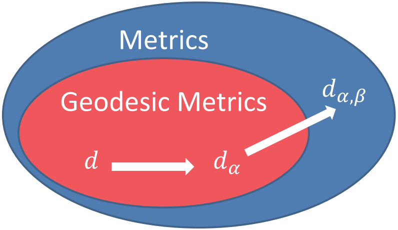
How to set the family $\mathcal{D}$ of metrics?

$\Longrightarrow$ by focusing on their curvature and the intrinsic means.

Our policy: keep the problem in geometry
as much as possible.

Two steps of changing metrics

A geodesic metric space is a metric space such that the distance between two points is equivalent to the shortest path length connecting them.



We assume the original metric is a geodesic metric (usually the Euclidean or the shortest path length of a metric graph).

The α, β-metric and α, β, γ-mean

We propose a family of metrics:

$$d_{\alpha,\beta}(x, y) = g_\beta(d_\alpha(x, y))$$

and intrinsic means:

$$\hat{\mu}_{\alpha,\beta,\gamma} = \arg\min_{m\in\mathcal{M}} \sum_i g_\beta(d_\alpha(x_i, m))^\gamma$$

d_α: a locally transformed geodesic metric ($\alpha \in \mathbb{R}$)
g_β: a concave function corresponding to a specific kind of extrinsic means ($\beta \in (0, \infty]$)
γ: for L_γ-loss ($\gamma \geq 1$)

We will explain α and β one by one.

Data analysis by α, β and γ

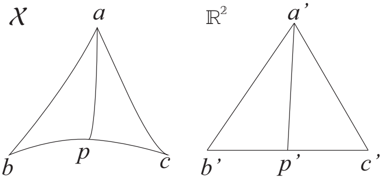
	Euclidean	$d_{\alpha,\beta,\gamma}$
metrics	$d(x,y) = \|x-y\|$	$d_{\alpha\beta\gamma}(x,y) = g_\beta(d_\alpha(x,y))$
intrinsic mean	$\arg\min_{m\in\mathbb{E}^d} \sum \|x_i - m\|^2$	$\arg\min_{m\in\mathcal{M}} \sum g_\beta(d_\alpha(x_i,m))^\gamma$
variance	$\min_{m\in\mathbb{E}^d} \frac{1}{n} \sum \|x_i - m\|^2$	$\min_{m\in\mathcal{M}} \frac{1}{n} \sum g_\beta(d_\alpha(x_i,m))^\gamma$
Fréchet function	$f(m) = \sum \|x_i - m\|^2$	$f_{\alpha\beta\gamma}(m) = \sum g_\beta(d_\alpha(x_i,m))^\gamma$

Outline

CAT(0)

A geodesic metric space $(\mathcal{X}, d)$ is a CAT(0) space iff for any $a, b, c \in \mathcal{X}$ the following condition is satisfied:

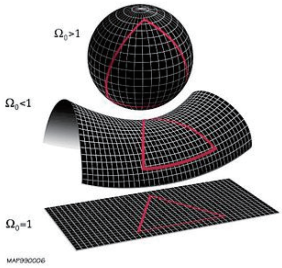
Construct a triangle in $\mathbb{E}^2$ with vertices a', b', c', called the comparison triangle, such that $\|a' - b'\| = d(a, b)$, etc. Select $p \in \widetilde{bc}$ and find the corresponding point $p' \in \overline{b'c'}$ such that $d(b, p) = \|b' - p'\|$. Then for any choice of $p \in \widetilde{bc}$, $d(p, a) \leq \|p' - a'\|$.



Intuitively speaking, each geodesic triangle in $\mathcal{X}$ is "thinner" than the corresponding one in a Euclidean space.

CAT(k)

CAT(k) is defined similarly but by using
(i) geodesic triangles whose perimeter is less than $2\pi/\sqrt{\max(k,0)}$ and
(ii) comparison triangles on a surface with a constant curvature k.



Locally, a simply connected Riemannian manifold with sectional curvatures at most k is CAT(k).

Globally, it requires completeness condition (i.e. two different geodesics can intersect only once) but up to diameter $1/\sqrt{\max(k,0)}$.

Convexity, Geodesic and Unique Mean

Theorem 1 (Known result, e.g. Kendall (1990))

On a CAT(k) space, an empirical/population distribution has a unique local intrinsic mean in any subsets with a diameter smaller than $\pi/(2\sqrt{k})$.

Thus a lower curvature k of the data space makes the intrinsic means "more unique".

Outline

Geodesic metrics on distributions

$\mathcal{M}$: a geodesic metric space.
$\mathcal{X}$: r.v. on $\mathcal{M}$ with density $f(x)$.
$\Gamma = \{z(t) | t \in [0,1]\}$: a parametrised integrable path between $x_0 = z(0), x_1 = z(1)$ in M.
Let

$$s(t) = \sqrt{\sum_{i=1}^{d} \left(\frac{\partial z_i(t)}{\partial t} \right)^2},$$

with appropriate modification in the non-differentiable case, be the local element of length along Γ.
The weighted metric along Γ is

$$d_\Gamma(x_0, x_1) = \int_0^1 s(t) f(z(t)) dt.$$

The geodesic metric is $d(x_0, x_1) = \inf_\Gamma d_\Gamma(x_0, x_1)$.

The d_α Metric: Population Case

$\Gamma = \{z(t), t \in [0,1]\}$ between $x_0 = z(0)$ and $x_1 = z(1)$,

$$d_{\Gamma,\alpha}(x_0, x_1) = \int_0^1 s(t) f^\alpha(z(t)) dt$$

and

$$d_\alpha(x_0, x_1) = \inf_\Gamma d_{\Gamma,\alpha}(x_0, x_1).$$

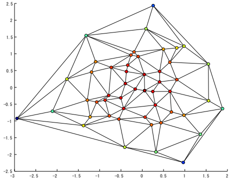
Roughly speaking when α is more negative (positive) so curvature is more negative (positive).

We can prove that for $d = 1$, the intrinsic mean for $\alpha = 1$ is equivalent to the (ordinary) median.

Empirical Metric Graphs

There are various empirical graphs whose vertices are the data points:

1. Complete graph
2. Delaunay graph
3. Gabriel graph
4. k-NN graphs
5. ϵ-NN graphs



Delaunay empirical graph

We introduce a metric on the graph by the shortest path length:

$$d(x_0, x_1) := \inf_{\Gamma} \sum_{e_{ij} \in \Gamma} d_{ij},$$

where d_{ij} is the length of an edge e_{ij}.

The d_α Metric: Empirical Graph Case

d_α metric for an empirical graph is defined by the shortest path length with powered edge lengths:

$$d_\alpha(x_0, x_1) := \inf_\Gamma \sum_{e_{ij} \in \Gamma} d_{ij}^{1-\alpha}.$$

This is an empirical version of

$$d_\alpha(x_0, x_1) = \inf_\Gamma \int_0^1 s(t) f^\alpha(z(t)) dt.$$

Here we use a fact, under some regularity conditions, $d_{ij}^{-1/p}$ is an unbiased estimator of the local density where p is the dimension of $\mathcal{M}$. Thus a natural rescaling of d_{ij} is

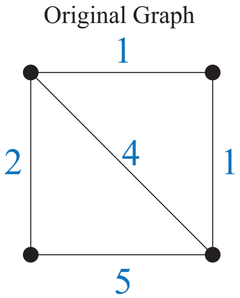
$$d_{ij} d_{ij}^{-\alpha/p} = d_{ij}^{1-\alpha/p}.$$

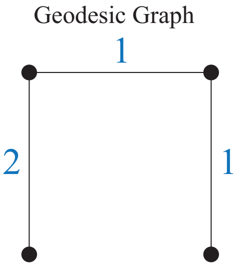
By resetting $\alpha := \alpha/p$, $d_{ij}^{1-\alpha}$ is obtained.

Geodesic Graphs

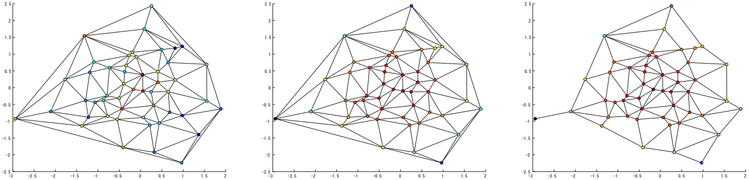
Definition 2

For an edge-weighted graph G, the union of all edge-geodesics between all pairs of vertices is called the geodesic sub-graph of G and denoted as G^*.

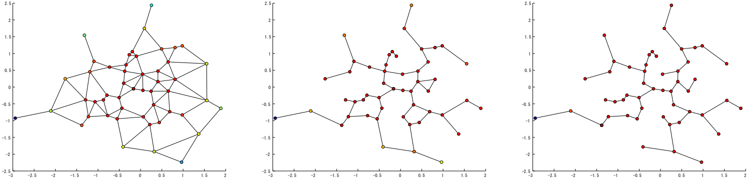




Ex: Geodesic Graph (Delaunay Graph with α)

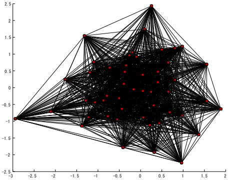


(a) $\alpha = 1$ (b) $\alpha = 0$ (c) $\alpha = -0.3$

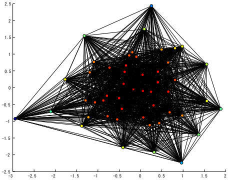


(d) $\alpha = -1$ (e) $\alpha = -5$ (f) $\alpha = -30$

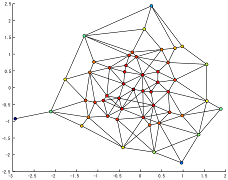
Ex: Geodesic Graph (Complete Graph with α)



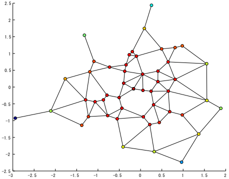
(a) $\alpha = 1$



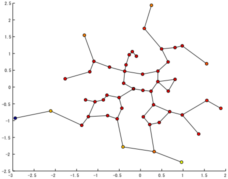
(b) $\alpha = 0$



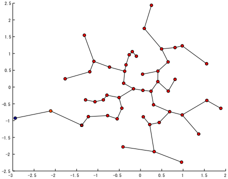
(c) $\alpha = -0.3$



(d) $\alpha = -1$



(e) $\alpha = -5$



(f) $\alpha = -30$

α-Chain and minimal spanning trees

The geodesic subgraph G^*_α gives a filter:

Theorem 3

Let G_α be an edge-weighted graph with distinct weights $\{d_{ij}^{1-\alpha}\}$ and let G^*_α be its geodesic subgraph then:

$$\alpha' < \alpha \leq 1 \Rightarrow G^*_{\alpha'} \subseteq G^*_\alpha.$$

For sufficiently small α, G^*_α becomes the minimal spanning tree and, therefore, CAT(0):

Theorem 4

There is an α^* such that for any $\alpha \leq \alpha^*$ the geodesic sub-graph becomes the minimal spanning tree $T^*(G)$ endowed with the d_α metric and, therefore, becomes a CAT(0) space.

Smaller α implies CAT(k) for smaller k

Assume $\alpha \leq 1$.
$D_k(X, x)$: the maximum radius of a disk centred at x being CAT(k).
$\bar{X}$: a rescaling of X such that the shortest edge length is 1.
For metric graphs, $D_k(X, x)$ can be computed only from the shortest cycle length and we can prove

Theorem 5

If $\alpha' < \alpha \leq 1$

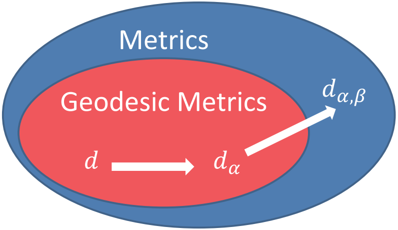
$$D_k(\bar{G}^*_{\alpha'}, x) \geq D_k(\bar{G}^*_{\alpha}, x) \text{ for each } k \in \mathbb{R}.$$

i.e. $\bar{G}^*_\alpha$ becomes "more CAT(k)" for smaller α. Since rescaling of the graph does not affect the uniqueness of the intrinsic mean, G^*_α tends to have a unique mean for a smaller α.

Outline

The second step is by β

A geodesic metric space is a metric space such that the distance between two points is equivalent to the shortest path length connecting them.



We assume the original metric is a geodesic metric (usually the Euclidean or the shortest path length of a metric graph).

Keio University

CALAMVS GLADIO FORTIOR 1858

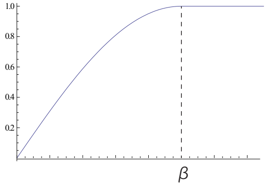
d_β Metric

Let (X, d) be a geodesic metric space. For $\beta > 0$, transform the metric d:

$$d_\beta(x_0, x_1) = g_\beta(d(x_0, x_1))$$

where

$$g_\beta(z) = \begin{cases} \sin(\frac{\pi z}{2\beta}), & \text{for } 0 \le z \le \beta, \\ 1, & \text{for } z > \beta. \end{cases}$$



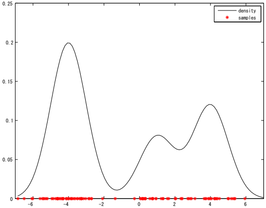
For $\beta = \infty$, $d_\beta = d$.

d_β satisfies the triangle inequality and becomes a metric but not a geodesic metric.

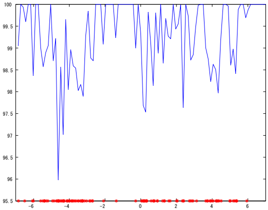
d_β-mean: $\hat{\mu}_\beta = \arg\min_{m \in X} \sum_i g_\beta(d(x_i, m))^2$.

β and clustering

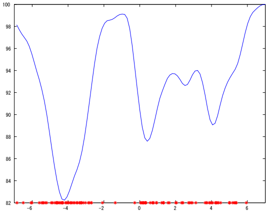
$f(m) = \sum_i g_\beta(|x_i - m|)^2$ with various β:



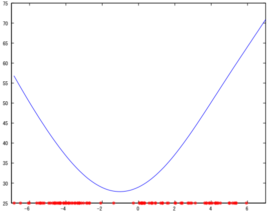
(a) Density function



(b) $\beta = 0.1$



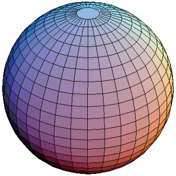
(c) $\beta = 1$



(d) $\beta = 10$

Extrinsic Mean on a Sphere: Revisited

Extrinsic mean on a unit sphere:



$$\hat{\mu} = \arg\min_{m \in S^2} \sum_i \|x_i - m\|^2.$$

Merit of extrinsic means: Euclidean distance is easier to compute than geodesic length on the data space.

Extrinsic mean on a metric space (X, d) embedded in $(\tilde{X}, \tilde{d})$:

$$\hat{\mu} = \arg\min_{m \in X} \sum_i \tilde{d}(x_i, m)^2.$$

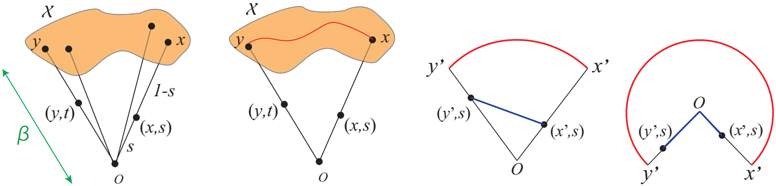
d_β-mean can be redefined as an extrinsic mean when the data space is embedded in a "metric cone".

Metric Cone

$\mathcal{X}$: a geodesic metric space
A metric cone $\tilde{\mathcal{X}}_\beta$ with $\beta > 0$ is a (truncated) cone $\mathcal{X} \times [0,1]/\mathcal{X} \times \{0\}$ with a metric

$$\tilde{d}_\beta((x,s),(y,t)) = \sqrt{t^2 + s^2 - 2ts\cos(\pi \min(d_{\mathcal{X}}(x,y)/\beta, 1))}$$

for any $(x,s),(y,t) \in \tilde{\mathcal{X}}_\beta$.



d_β-mean can be redefined as an extrinsic mean when the data space is embedded in a "metric cone".

β and CAT(k) of the Metric Cone

Theorem 6

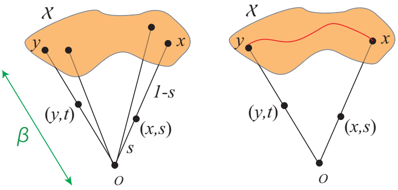
1. *If $\mathcal{X}$ is a CAT(0) space, the metric cone $\tilde{\mathcal{X}}_\beta$ is also CAT(0) for every $\beta \in (0, \infty)$.*
2. *If $\tilde{\mathcal{X}}_{\beta_2}$ is CAT(0), $\tilde{\mathcal{X}}_{\beta_1}$ is also CAT(0) for $\beta_1 < \beta_2$.*
3. *If $\mathcal{X}$ is CAT(k) for $k \geq 0$, $\tilde{\mathcal{X}}_\beta$ becomes CAT(0) for $\beta \leq \pi/\sqrt{k}$.*

Roughly speaking, the theorem insists that smaller β makes the metric cone less curved.

Extrinsic Mean in Metric Cone

Compared with ordinary extrinsic means for embedding in Euclidean space,

- "Curvature" of embedding space can be tuned by β.
- the embedding space is only 1-dimensional higher than the original data space.



The α, β, γ-mean: Summary

We proposed a class of intrinsic means:

$$\hat{\mu}_{\alpha,\beta,\gamma} = \arg\min_{m \in \mathcal{M}} \sum_i g_\beta(d_\alpha(x_i, m))^\gamma$$

and corresponding variances:

$$V_{\alpha,\beta,\gamma} = \min_{m \in \mathcal{M}} \frac{1}{N} \sum_i g_\beta(d_\alpha(x_i, m))^\gamma$$

d_α: a locally transformed geodesic metric
g_β: a concave function corresponding to extrinsic means in metric cones
γ: L_γ-loss

Outline

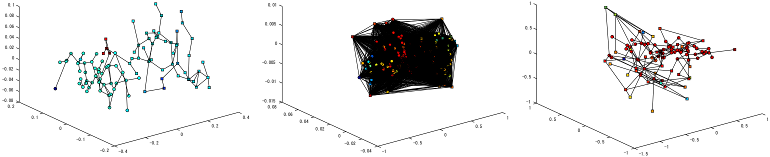
Application: Clustering

Data: five kinds of data from UCI Machine Learning Repository (iris, wine, ionosphere, breast cancer, yeast)
- The clustering error by k-means method decreases significantly by selecting an adequate value.
- $\alpha \in \{-5.0, -4.8, \ldots, 0.8, 1\}$ and
  $\beta \in \{2^{-3}, 2^{-2}, \ldots, 2^{6}, \infty\}$.

data set	k-means with $d_{\alpha,\beta}$			Euclid
	$\hat{\alpha}$	$\hat{\beta}$	r^*	r
(i) iris	-4.4	0.125	0.0333	0.1067
(ii) wine	0.8	8	0.2753	0.2978
(iii) ionosphere	-5.0	16	0.0798	0.2877
(iv) cancer	0.8	16	0.0914	0.1459
(v) yeast	-0.6	2	0.4447	0.4515

Application: Clustering

- The structure of the "optimal" geodesic graphs differs depending on the data:



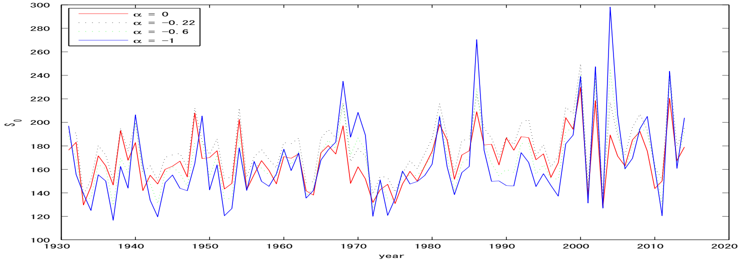
(a) iris ($\hat{\alpha} = -5.0, \hat{\beta} = 0.125$) (b) wine ($\hat{\alpha} = 0.8, \hat{\beta} = 4$) (c) ionosphere ($\hat{\alpha} = -3.2, \hat{\beta} = 8$)

Figure 1 : The geodesic graph of each data set with an optimum value of α and β for a randomly selected 100 sub-samples.

Application: Rainfall Data

Time series of "variance" $s_0^2 := \{\min_i \sum_j d_\alpha(x_i, x_j)^2\}^{1/(1-\alpha)}$

are plotted for $\alpha = 0$ (red solid line), -0.22 (black dashed line) and -1 (blue solid line).



This generalized "variance" is expected to detect change of another type of volatility incorporating spacio-temporal geometrical structure of the precipitation data.

Summary

- Curvature of the data space should be focused again in the recent development of studies on empirical geodesic graphs (e.g. manifold learning).
- The α-metric is a deformation of a geodesic metric. For empirical graphs, α can control the power law on an estimated density.
- Smaller $\alpha < 1$ makes the data space CAT(k) with a smaller k.
- β-metric is non-geodesic but embeddable in a geodesic metric cone.
- Smaller β makes the embedding metric cone CAT(k') with a smaller k'.

Summary

- This maybe the first study of an extrinsic mean by embedding in non-Euclidean spaces and the first application of metric cones to statistics and data analysis.
- Uniqueness of the L_γ-mean depends on γ for non-Euclidean spaces.
- Trade-off between uniqueness of the mean and robustness of the estimation can be managed by the curvature of the data space and the embedding metric cone via α, β and γ.
- See arXiv:1401.3020 [math.ST] for the proofs and details.

Thank you very much for listening!