Statistical analysis via the curvature of data space

Kei Kobayashi*, Orita Mitsuru† and Henry P. Wynn**

*The Institute of Statistical Mathematics, Japan
†Kumamoto University
**London School of Economics

Abstract. It has been known that the curvature of data spaces plays a role in data analysis. For example, the Frechet mean (intrinsic mean) always exists uniquely for a probability measure on a non-positively curved metric space. In this paper, we use the curvature of data spaces in a novel manner. A methodology is developed for data analysis based on empirically constructed geodesic metric spaces. The population version defines distance by the amount of probability mass accumulated on travelling between two points and geodesic metric arises from the shortest path version. Such metrics are then transformed in a number of ways to produce families of geodesic metric spaces. Empirical versions of the geodesics allow computation of intrinsic means and associated measures of dispersion. A version of the empirical geodesic is introduced based on some metric graphs computed from the sample points. For certain parameter ranges the spaces become CAT(0) spaces and the intrinsic means are unique. In the graph case a minimal spanning tree obtained as a limiting case is CAT(0). In other cases the aggregate squared distance from a test point provides local minima which yield information about clusters. This is particularly relevant for metrics based on so-called metric cones which allow extensions to CAT(κ) spaces. We show how our methods work by using some actual data. This paper is a summary of a longer version [5]. See it for proof of theorems and details.

Keywords: intrinsic mean, extrinsic mean, CAT(0), curvature, metric cone, cluster analysis, non-parametric analysis
PACS: 02.50.-r, 02.40.Ma

INTRODUCTION

The fundamental object in this paper is a geodesic metric space. This is defined in two stages. First define a metric space X with a metric d called the *base data space*, or data space for short which we write (X,d). Sometimes X will be a Euclidean space E_d of dimension d, containing the data points. But it may be some special object such as a graph or manifold including the data. The second stage is to define geodesics between points in the usual way so that the distance between two points is the length of the geodesic connecting the points. The space X endowed with the geodesic metric is called the data space (X,d^*).

The interplay between (X,d) and (X,d^*) will be critical for this paper and we will have a number of ways of constructing d^*. One of the ways to change the metric from d to d^* is to study the different means and other statistical quantities.

It is important to understand difference between extrinsic and intrinsic means. The *Intrinsic mean* is defined by

$$\mu_1 = \arg\inf_{\mu\in X} \sum_i d^*(x_i,\mu)^2.$$

Another possibility is where the data and the space (X,d^*) is embedded in a larger space (X^+,d^+) and the mean is

$$\mu_2 = \arg\inf_{\mu\in X} \sum_i d^+(x_i,\mu)^2.$$

We refer to μ_2 as an *extrinsic mean*. By abuse of notation, we will use notation d even for a geodesic distance of a data space.

There are a number of simple geometric examples where curvature is critical and where discussion on intrinsic and extrinsic means is helpful [1] [3]. In general the intrinsic means are not unique, but for so-called CAT(0) spaces, which (trivially) include Euclidean spaces, they are. Even when they are not unique, the mean function, $f(m) = \sum_i d(x_i,m)^2$ can yield useful information, for example about clustering.

A *CAT(κ) space* for $k \in \mathbb{R}$ is intuitively a metric space whose curvature is bounded above by κ and defined as a geodesic metric space satisfying following CAT(κ) condition. Let (X,d) be a geodesic metric space and take for any geodesic triangle Δabc whose perimeter is less than $2\pi/\sqrt{\max(\kappa,0)}$ on X and for any point x on a geodesic $\widetilde{bc}$. Let $\Delta a'b'c' \subset \mathbb{R}^2$ be a comparison triangle, which has the same edge length as Δabc, on a surface with a constant curvature κ, i.e. a sphere with radius $1/\sqrt{\kappa}$ for $\kappa > 0$, a plane for $\kappa = 0$ and a hyperbolic space for $\kappa < 0$. Set x' on $\widetilde{b'c'}$ satisfying $|\widetilde{bx}| = |\widetilde{b'x'}|$. Then the CAT($\kappa$) condition is that for all a,b,c and all choices of x, $|\widetilde{ax}| \le |\widetilde{a'x'}|$. Thus every CAT($\kappa$) space is a CAT($\kappa'$) space for $\kappa < \kappa'$. Every metric graph is CAT(κ) for $\kappa > 0$ iff there is no loop shorter than $2\pi/\sqrt{\kappa}$ since every metric tree is CAT(0) and, therefore, CAT(κ) for $\kappa > 0$.

Geodesic metrics on distributions

Let X be a d-dimensional random variable absolutely continuous with respect to the Lebesgue measure, with density $f(x)$. Let $\Gamma = \{z(t), t \in [0,1]\}$ be a parametrised integrable path between two points in $x_0 = z(0), x_1 = z(1)$ in $\mathbb{R}^d$, which is rectifiable with respect to Lebesgue measure. 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 distance along Γ is

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

The geodesic distance is $d(x_0,x_1) = \inf_\Gamma d_\Gamma(x_0,x_1)$. Here we consider a random variable on a Euclidean space, but this can be generalized for Riemannian manifolds and even for singular spaces with a density with respect to a base measure naturally defined by

the metric. It may happen, for example, when the density does not exist that the geodesic may be piecewise smooth.

From the geodesic distances on the distribution we shall develop in three main directions:

1. Transform the geodesic metrics in various ways with parameters α, β, γ to obtain a wide class of metrics
2. Discover CAT(0) and CAT(κ) spaces for certain ranges of the parameters
3. Apply empirical versions of the metrics based on graphs of certain complexes

Then a Fréchet mean proposed in the paper is

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

with a function g_β and a distance d_α defined below.

THE d_α METRIC AND MINIMAL SPANNING TREES

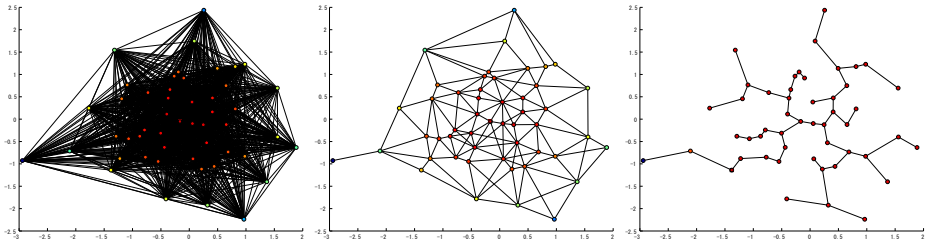
The general d_α metric is what we have referred to as a local metric. It is obtained by transforming the density in (1). Thus for $\Gamma = \{z(t), t \in [0,1]\}$ between $x_0 = z(0)$ and $x_1 = z(1)$,

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

and $d_\alpha(x_0, x_1) = \inf_\Gamma d_{\Gamma,\alpha}(x_0, x_1)$. By changing α essentially change the local curvature. Roughly speaking when α is more negative (positive) so curvature is more negative (positive). Note that the intrinsic mean for $\alpha = 1$ is equivalent to the median for $d = 1$.

There are a number of options to produce an empirical version of the d_α metric based on the data. One such option would be to produce a smooth empirical density $f(t)$ and numerical integration and optimization to compute geodesics. We prefer a much simpler method based on a graph constructed from the data, for example 1) Complete graph with vertices at the data points and all edges, 2) The edge graph (1-skeleton) of the Delaunay simplicial complex with vertices at the data points, 3) Gabriel graph with vertices at the data points. All geodesic computation is then restricted to the graph. Remark that Delaunay and Gabriel graphs are known to hold good properties, e.g. they include the minimum spanning tree of the data points as their subgraph.

The discussion below applies to the complete graph, or any connected sub-graph. For any such graph, define a version of the d_α distance, just for edges $\tilde{d}_{\alpha,ij} = d_{ij}^{1-\alpha}$, where d_{ij} is the Euclidean distance from x_i to x_j. This can be explained by making a transformation $ds \to \frac{ds}{d_{ij}}$. We refer to this as *edge regularization*. We then apply α in the usual way to obtain: ds/d_{ij}^α. The new “length" of each edge e_{ij} is obtained by integrating this “density" along the edge. In this sense, d_{ij} also plays a role of density estimation. Though we need a regularization $d_{ij}^{-1/p}$ with respect to the dimension p for density estimation [4], we manage the regularization by rescaling the parameter α. Note that $\alpha = 1$ gives the unit length and $\alpha = 0$ restores the original length.



(a) $\alpha = 0$ (b) $\alpha = -0.3$ (c) $\alpha = -30$

FIGURE 1. The geodesic graphs with different values of α for 50 samples from the standard 2-d Normal distribution. Initial graph ($\alpha = 0$) is the complete graph.

Now we consider only the set of edges E of the graph $G(V,E)$ as a metric space with metric defined by the geodesic:

$$\tilde{d}_\alpha(x_0,x_1) = \inf_\Gamma \sum_{(i,j)\in\Gamma} \tilde{d}_{\alpha,ij},$$

where the summation is taken over all (connected) paths Γ between x_0 and x_1. Note that even if $d_{ij}^{-1/p}$ can estimate the local density well, it does not imply that the metric d_α can be approximated by the metric $\tilde{d}_\alpha$ since edge lengths $d_i j$ on each path Γ are not independent. It is suggested that further theoretical work is necessary. Here we will admit $\tilde{d}_\alpha$ as an approximation of d_α.

If the graph is not a complete Euclidean graph with weights equal to the Euclidean length of edges, some edges may not be in any edge-geodesics between any pair of vertices. For an edge-weighted graph G with weights $\{d_{ij}\}$ on the graph G^* which is the union of all edge-geodesics between all pairs of vertices is called the *geodesic sub-graph* of G. We will see how the geodesic graphs transform as the value of α changes. Especially for a very small α, we can prove the following theorem.

Theorem 1 *There is an α^* such that for any $\alpha \le \alpha^*$ the geodesic sub-graph becomes the minimal spanning tree $T^*(G)$ endowed with the d_α metric and, therefore, becomes a CAT(0) space.*

We may study the geometry as α increases away from $-\infty$. Following Theorem 2 we are interested in two cases: when the graph of interest is the Euclidean Delaunay graph and the complete Euclidean graph. In both cases we consider the d_α metric.

Theorem 2 *Let G_α be an edge-weighted graph with distinct weights $\{d_{ij}^{1-\alpha}\}$ and let G^*_α be its geodesic subgraph then: $|1-\alpha'| > |1-\alpha| \Rightarrow G^*_{\alpha'} \subseteq G^*_\alpha$*

Figures 1 (a)-(c) are geodesic graphs with different values of α for 50 samples of the standard 2-d Normal distribution.

α and CAT(κ)

Let $C(X,p,r) := \{x \in X \mid d(p,x) \le r\}$ be a geodesic disc of the radius $r \ge 0$ centred at $p \in X$. Define the maximum radius $D_\kappa(X,x)$ of the disk centred at x being CAT(κ),

that is

$$D_\kappa(X,x) := \sup\{r \geq 0 \mid X \cap C(X,x,r) \text{ is CAT}(\kappa)\}.$$

If X is a metric graph, $D_\kappa(X,x) = \sup\{r \geq 0 \mid X \cap C(X,x,r)$ does not include a cycle whose length is less than $2\pi/\sqrt{\max(k,0)}\}$.

Consider a rescaling of X such that the shortest (longest) edge length is 1 and denote it $\bar{X}$ for $\alpha \leq 1$ ($\alpha > 1$, respectively).

Theorem 3 *If* $|\alpha' - 1| > |\alpha - 1|$

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

By the theorem $\bar{G}^*_\alpha$ becomes "more CAT(κ)" for a smaller $\alpha < 1$. 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 $\alpha < 1$.

THE d_β METRIC

As we mentioned on a Euclidean space in the introduction, the uniqueness of the intrinsic mean caused by the CAT(0) property avoids to capture the multiple local means for multi-modal distributions. A concave transformation of the metric can play the role to modify the base data space "less CAT(0)" and we introduce d_β metric by g_β transformation as a candidate for that.

For any geodesic metric space (X,d) with metric $d(x_0,x_1)$ and a parameter $\beta > 0$, we can define the following metric with

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

where

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

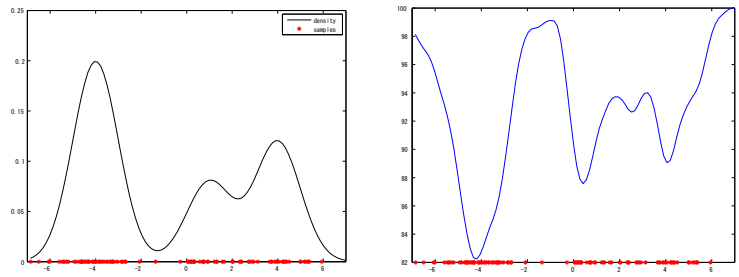
Since g_β is a concave function on $[0,\infty)$, d_β becomes a metric but not necessarily a geodesic-metric. Since, as we will see soon, d_β can be recognized as a geodesic metric of a cone embedding X, we refer to the mean

$$\hat{\mu}_\beta = \arg\inf_{\mu \in X} \sum_{i=1}^n g_\beta(d(x_i,\mu))^2$$

as the β-*extrinsic mean*.

The extrinsic mean keeps the mean within the original data space but uses the d_β metric. We will see below that the d_β metric involves, even in the general case, an extension of the original space by a single dimension. This distinguishes it from the embeddings for extrinsic means used in the literature which involves high-dimensional Euclidean space, for example the case of the tree space.

Controlling β, as we will see below, controls the value of κ when the embedding space is considered as a CAT(κ) space. We have an indirect link between clustering



(a) Density function (b)$\beta = 1$

FIGURE 2. (a) The density function is a mixture of three normal distributions. and 100 i.i.d. samples from it. (b) the graph of $\sum_i d_\beta(x_i, m)^2$ against m for $\beta = 1$.

and CAT(κ) spaces. As β decreases while the embedding space becomes more CAT(0) (κ decreasing) the original space becomes less CAT(0). This demonstrates, we believe, importance of CAT(κ) property in geodesic based clustering.

In a Euclidean space, the standard Euclidean distance does not exhibit multi local means since the space is trivially CAT(0). But using d_β-metric with a sufficiently small β, the space can have multi-local means as in Figure 2.

Metric cone

The above construction is a special case of a general construction which applies to any geodesic metric space and hence to those of this paper. Let $\mathscr{X}$ be a geodesic metric space with a metric $d_{\mathscr{X}}$. A metric cone $\tilde{\mathscr{X}}_\beta$ with $\beta \in (0,\infty)$ is a cone $\mathscr{X} \times [0,1]/\mathscr{X} \times \{0\}$ with a metric

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

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

The intuitive explanation is as follows. Let $\mathscr{X}_\beta$ be a subset $\{(x,1) \mid (x,t) \in \tilde{\mathscr{X}}_\beta\}$ with the geodesic metric on it. Then $\mathscr{X}_\beta$ is a rescaling of the metric on $\mathscr{X}$ by β. For any $(x,s),(y,t) \in \tilde{\mathscr{X}}_\beta$, their projections $(x,1),(y,1)$ give two points $x,y \in \mathscr{X}_\beta$, respectively. For a geodesic $\gamma \subset \mathscr{X}_\beta$ between x and y, consider a cone $\{(x,s) \mid x \in \gamma, s \in [0,1]\}$ spanned by γ. This cone can be isometrically embedded to an "extended unit circular sector", i.e. a covering $\{(r,\theta) \mid r \in [0,1], \theta \in (-\infty,\infty)\}/\{(0,\theta) \mid \theta \in (-\infty,\infty)\}$ of the unit disk corresponding to $\theta \in [0, \pi d_{\mathscr{X}}(x,y)/\beta]$. Then (x,s) and (y,t) are also mapped to the extended unit circular sector, the distance $\tilde{d}_\beta((x,s),(y,t))$ for $\beta = 1$ corresponds to the case (D2) of a disk if we set $(r,r') = (s,t)$ and $(\theta,\theta') = (\pi x, \pi y)$. For more on metric cones, [2] is a good summary.

The following result indicates that the metric cone space preserves the CAT(0) property of the original space and the smaller values of β continue this process.

Theorem 4 *1. If $\mathscr{X}$ is a CAT(0) space, the metric cone $\tilde{\mathscr{X}}_\beta$ is also CAT(0) for every $\beta \in (0,\infty)$.*

2. If $\tilde{\mathscr{X}}_{\beta_2}$ is CAT(0), $\tilde{\mathscr{X}}_{\beta_1}$ is also CAT(0) for $\beta_1 < \beta_2$.

3. *If $\mathscr{X}$ is CAT(κ) for $k \geq 0$, $\tilde{\mathscr{X}}_\beta$ becomes CAT(0) for $\beta \leq \pi/\sqrt{k}$.*

As noted above, the metric cone embedding the original data space becomes CAT(0) for β sufficiently small. The following is a possible strategy: start with the Euclidean graph apply the cone method decreasing β until the intrinsic mean is unique (this may happen at a larger value of β than that required to be CAT(0)). Then project that unique mean back onto the original graph; the projection is unique by construction provided the intrinsic mean is not at the apex of the cone. Rather than judge this method now we will wait for some empirical studies.

In the β case rather than the projection of the intrinsic mean in the cone back to the original geodesic space, we prefer the extrinsic mean in which (i) the mean lies in the geodesic space but (ii) we use the β cone metric d_β. As we decrease β we tend to get more local minima. This is the antithesis of obtaining a CAT(0) space.

Because the β metric becomes more concave as β decreases for very small β, each data point will yield a single "local mean". For reasonable values of β, each data cluster will have a local mean because the function g_β is locally linear and intra-cluster distances are small. But when the intra-cluster distances are larger, the concavity dominates, and we obtain multiple local minima.

The transformation of a metric by the parameter $\alpha \in \mathbb{R}$ maps a geodesic metric d to another geodesic metric d_α while the transformation by $\beta > 0$ maps a (not necessarily geodesic) metric d to another (not necessarily geodesic) metric d_β. The transformation by α can be said to be 'local' since it depends on a density function and computed by the integration while the transformation by β is 'global' in this sense. The difference between α and β is discussed in a longer version [5] of the paper with some examples.

CURVATURE, DIAMETER AND EXISTENCE OF MEANS

Let $\mathscr{X}$ be a geodesic metric space and fix it throughout this section. The diameter of a subset $A \subset \mathscr{X}$ is defined as the length of the longest geodesic in A. We define classes $\mathscr{C}_{\text{convex}}$, $\mathscr{C}_{L_\gamma}$ and $\mathscr{C}_{\text{geodesic}}$ in the following manner.

1. $\mathscr{C}_{\text{convex}}$: the class of the subsets $A \subset \mathscr{X}$ such that the geodesic distance function $f_p(x) := d(p,x)$ is strictly convex on A for each $p \in A$. Here, "convex" means geodesic-convex, i.e. a function f on $\mathscr{X}$ is convex iff for every geodesic $\{z(t) \mid t \in (t_0, t_1)\}$ on $\mathscr{X}$, $f(z(t))$ is convex with respect to t.
2. $\mathscr{C}_{L_\gamma}$ for $\gamma \in [1,\infty]$: the class of the subsets $A \subset \mathscr{X}$ such that for any probability measure whose support is in A and non-empty, intrinsic L_γ-mean exists uniquely. We call $\mathscr{C}_{L_2}$ as $\mathscr{C}_{\text{mean}}$, especially.
3. $\mathscr{C}_{\text{geodesic}}$: the class of the subsets $A \subset \mathscr{X}$ such that for every pair $p, q \in A$, the geodesic between p and q is unique.

Lemma 5 $\mathscr{C}_{\text{convex}} \subset \mathscr{C}_{L_\gamma} \subset \mathscr{C}_{\text{geodesic}}$ *for any* $\gamma \in [1,\infty]$.

Let $D_{\text{convex}}, D_{L_\gamma}$ and D_{geodesic} be the largest values (including ∞) such that every subset whose diameter is less than the value belongs to $\mathscr{C}_{\text{convex}}, \mathscr{C}_{L_\gamma}$ and $\mathscr{C}_{\text{geodesic}}$, respectively.

Then evidently from Lemma 5, $D_{\text{convex}} \leq D_{L_\gamma} \leq D_{\text{geodesic}}$ for $1 \leq \gamma \leq \infty$. Note that if $\mathscr{X}$ is CAT(0), $D_{\text{convex}} = D_{L_\gamma} = D_{\text{geodesic}} = \infty$. In general, the following theorem holds.

Theorem 6 *(1) If $\mathscr{X}$ is CAT(k), $D_{\text{convex}} \geq \pi/(2\sqrt{\max(k,0)})$.*
(2) If $\mathscr{X}$ is CAT(k), $D_{\text{geodesic}} \geq \pi/\sqrt{\max(k,0)}$.
(3) If $\mathscr{X}$ is the surface with a constant curvature $k > 0$, $D_{L_1} \geq \pi/(2\sqrt{k})$

By Theorem 6(1), $D_{L_\gamma} \geq D_{\text{convex}} \geq \pi/(2\sqrt{k})$. Thus a lower curvature k gives a wider area where the intrinsic L_γ-mean is unique. Theorem 6(3) says this lower bound for D_{L_1} is the best as a universal upper bound for any $\mathscr{X}$ with CAT(k) property.

SUMMARY

In classification, regression and many other statistical methods, an optimal function (e.g. best classifier or regression function) is selected from a pre-defined set of functions. In this paper, we define instead a set of metrics from where the optimal one is selected. In order to define the set of metrics, we use transformations of a metric by real parameters α and β. The transformation by α maps a geodesic metric d to another geodesic metric d_α. The intrinsic mean with respect to d_α is a generalization of one-dimensional Euclidean median. Since it is hard to compute the metric d_α analytically or numerically, we introduced an approximation by the shortest path length on the empirical graph. The transformation by β maps a (not-necessarily geodesic) metric d to another metric d_β. Though the metric d_β is not a geodesic metric in general, it can be recognized as a geodesic metric. The curvature of a space where the data lie plays an important role in statistical analysis by controlling the number of the intrinsic means and is theoretically analysed by the CAT(κ) property. The curvature of the space and the embedding metric cone is controlled by α and β, respectively. In a longer version [5] of the paper, we show application of our methods to some real data.

ACKNOWLEDGMENTS

This work was partly supported by JSPS KAKENHI Grant Numbers 24700288 and 25370634, ISM Research Collaboration Grant 26-Kyoken-2064 and JST, PRESTO.

REFERENCES

1. Bhattacharya, A., and Bhattacharya, R., *Nonparametric inference on manifolds: with applications to shape spaces*, vol.2. Cambridge University Press, 2012.
2. Deza, M. M. and Deza, E., *Encyclopedia of distances*, Springer Berlin Heidelberg, 2009.
3. Kendall, D. G., Barden, D., Carne, T. K. and Le, H., *S*hape and shape theory, vol.500, John Wiley & Sons, 2009.
4. Kendall, M. G., and Moran, P. A. P. *Geometrical Probability*. Griffin, London, 1963.
5. Kobayashi, K. and Wynn, H. P., Empirical geodesic graphs and CAT(κ) metrics for data analysis, arXiv1401.3020, 2014.