データ空間の距離グラフ近似手法の改良

慶應義塾大学　JST さきがけ　小林　景

本発表では発表者がここ数年研究しているデータ空間の距離の変換手法について，主に多様体学習との比較を中心に説明する．特に，そこから見えてくる提案手法の新規性，発展性および課題点について説明する．提案手法について，以下箇条書きでまとめる．

【用語の説明】

- データ空間 とは，データの分布のサポートを含むような，幾何学的対象（通常距離空間）である．ただし，データが近似的に幾何学的対象の周りに分布する場合にもデータ空間という用語を用いることがある．

- 測地距離空間 とは，任意の二点間の距離が，それを結ぶ最短の曲線長と一致するような距離空間である．

- 測地距離空間 $(\mathcal{X}, d)$ が CAT(0) 空間 であるとは，任意の $a, b, c \in \mathcal{X}$ と $a', b', c' \in \mathbb{R}^2$ が $\|a'-b'\| = d(a,b), \|b'-c'\| = d(b,c), \|c'-a'\| = d(c,a)$ 等をみたすとき，測地線 $\widetilde{bc}$ 上の点 p と，$d(b,p) = \|b'-p'\|$ をみたす辺 $\overline{b'c'}$ 上の点 p' に対して，$d(a,p) \leq \|a'-p'\|$ が成り立つことである．直感的には，$\mathcal{X}$ 上の測地三角形がユークリッド空間上の対応する三角形より「へこむ」ような空間であり，局所的には非正曲率を持つ．より一般に $k \in \mathbb{R}$ についても，曲率 k の定曲率曲面上の三角形と比較することにより，CAT(k) 空間 が定義でき，k が大きくなるほど局所的に大きな曲率を持ち得ることに対応する．

- 測地距離空間 $(\mathcal{X}, d)$ 上の標本 $x_1, \ldots, x_n$ の 内測平均（intrinsic mean，Fréchet mean） は以下で定義され，その一意性は $\mathcal{X}$ の CAT(k) 特性に依存することが知られている．

$$\hat{\mu} = \arg\min_{m \in \mathcal{X}} \sum_{i=1}^{n} d(x_i, m)^2.$$

  また，関数 $f(m) = \sum_{i=1}^{n} d(x_i, m)^2$ は Fréchet 関数 とよばれ，測地距離空間上のデータの特徴を表す最も基本的な関数の一つである．

- 測地距離空間 $(\mathcal{X}, d)$ が他の距離空間 $(\tilde{\mathcal{X}}, \tilde{d})$ に埋め込まれているとき，埋め込まれた空間の距離を用いて

$$\hat{\mu} = \arg\min_{m \in \mathcal{X}} \sum_{i=1}^{n} \tilde{d}(x_i, m)^2.$$

  としたものを 外測平均（extrinsic mean） とよぶ．埋め込む空間 $(\tilde{\mathcal{X}}, \tilde{d})$ は通常ユークリッド空間で，測地距離を用いた内測平均より計算が簡単であるために応用上有効である．

- 距離グラフ とは，ここでは各辺 $e \in E$ に長さ（非負の重さ）が定義されている連結グラフ $G = (V, E)$ のことを指すとする．距離グラフの 測地部分グラフ (geodesic sub-graph) とは，どの頂点対を結ぶ最短経路にも含まれない辺を除いた部分グラフである．データ点を頂点とする距離グラフにおいて，Fréchet 関数の測地部分グラフ上での値は（よって全頂点上の値も），測地部分グラフのみによって定まる．

- 測地距離空間 $\mathcal{X}$ から構成される 計量錐 (metric cone) $\tilde{\mathcal{X}}$ とは，集合 $\mathcal{X}\times[0,1]/\mathcal{X}\times\{0\}$ 上の各点対 (x,s) と (y,t) に以下の距離を導入したものである：

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

$\tilde{\mathcal{X}}$ は測地距離空間になることが証明できる．直観的には,「原点」O と測地距離空間 $\mathcal{X}$ の各点とを結ぶ長さ 1 の「線分」の集合に，うまく測地距離を定義したものと考えることもできるが，ユークリッド空間に埋め込めるとは限らない．以下のように距離を定義すると,「線分」の長さを r とした計量錐を構成することもできる：

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

【提案手法】

- 本研究では，上記の Fréchet 関数をパラメータ $\alpha\in\mathbb{R},\beta\in(0,\infty],\gamma\geq 1$ を導入して以下のように一般化する．

$$f_{\alpha,\beta,\gamma}(m) = \sum_i g_\beta(d_\alpha(x_i,m))^\gamma$$

ただし，d_α,g_β については以下で説明する．これにより，対内測平均および分散も一般化される．また，一般化された Fréchet 関数は非凸となり得るので，その極小点はクラスタリングにおけるクラスター中心の候補となる．

- （α 距離）データ空間 $\mathcal{M}$ 上の確率分布の密度関数 f に対して，二点 $x_0=z(0)$，$x_1=z(1)$ を結ぶ曲線 $\Gamma=\{z(t),t\in[0,1]\}$ の長さを $d_{\Gamma,\alpha}(x_0,x_1)=\int_0^1 s(t)f^\alpha(z(t))dt$ で定義する．また，この経験分布版を以下のように定義する．
(1) 標本を頂点とするグラフを構成する（完全グラフ，k-NN グラフ等）．
(2) 各辺の長さ d_{ij} を $d_{ij}^{1-\alpha}$ に変換する．
(3) グラフでの最短経路長で標本間の距離を定義する．

α の増大にともない測地グラフ（頂点間の測地線に用いられる辺のみ残したグラフ）が縮小すること，それにともない CAT(k) となる領域が増大すること（曲率の減少に対応），および極限が最小全張木になることなどを証明した．

- （β 距離）パラメータ $\beta\in(0,\infty]$ による距離 d の変換を以下で定義する．

$$d_\beta(x_0,x_1)=g_\beta(d(x_0,x_1))\ \ (\beta>0),\text{ただし }g_\beta(z)=\begin{cases}\sin(\frac{\pi z}{2\beta}),\ \text{for}\ \ 0\leq z\leq\beta,\\ 1,\ \ \text{for}\ \ z>\beta.\end{cases}$$

β の変化により，Fréchet 関数の極小点の数を調整でき，クラスタリングに応用できる．β 距離空間は，一般に測地距離空間ではないので CAT(k) 特性を定義できない．一方，β 距離による内測平均は，動径 β の計量錘に埋め込まれた空間の外測平均として解釈できる．また β が小さくなると，CAT(k) 特性の意味で計量錘の曲率が小さくなることが証明できる．

本研究は Henry P. Wynn(London School of Economics) との共同研究である．

Kei Kobayashi, Henry P. Wynn (2014), Empirical geodesic graphs and CAT(k) metrics for data analysis, arXiv1401.3020.