データ空間の距離変形を用いた解析手法

小林　景　　数理・推論研究系　助教

※本研究はHenry P. Wynn教授 (LSE) との共同研究である.

【研究の背景・動機】

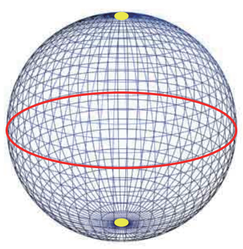
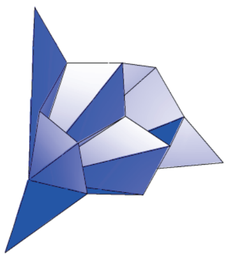
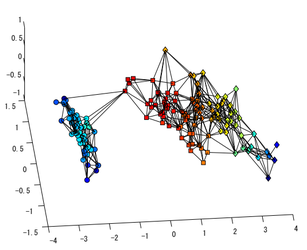
- 多様体などの距離空間（データ空間とよぶ）上に標本が分布するとき，平均などの統計的特徴量の性質はデータ空間の曲率に依存する.
- 特に距離空間$(\mathcal{M}, d)$上の標本(X_i)の内測平均（intrinsic mean，Fréchet mean）は以下で定義され，その一意性は$\mathcal{M}$の曲率に依存する.

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

【例】ユークリッド空間E^d上では内測平均は常に一意に存在し，通常の標本平均と一致する.

【例】球面S^d上では内測平均は一般に複数存在する．ただし，仰角$\pi/2$以下の領域内に標本が分布しているときは，常に一意に定まる．よって，北米大陸上の人口分布の内測平均は一意に存在するが，ユーラシア大陸上の内測平均は一意とは限らない.

【例】木構造グラフの空間は単体的扇とよばれる複雑な形となるが，CAT(0)とよばれる非正の局所的曲率をもつ空間なので，内測平均は一意に存在する.

- これらの研究では，データ空間は曲面などの単純な図形に限って議論する場合が多かったが，本研究ではより複雑な図形上のデータについて，曲率を用いた手法と理論評価を行う.（参考：多様体学習）
- データ空間の距離を変化させる（変形させる）ことにより，曲率も変化させた上で統計的解析を行う.
- 内測平均の一意性を保証するために曲率を小さくする場合に加え，クラスタリングのためにあえて曲率を大きくする場合も提案する.

【提案手法】

以下のような内測平均のクラスを提案する.

α, β, γ-内測平均

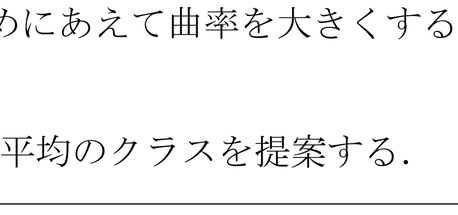
$$\hat{\mu}_{\alpha,\beta,\gamma} = \arg\min_{m\in\mathcal{M}} \sum_i g_\beta(d_\alpha(x_i, m))^\gamma$$

d_α: 局所的に変換した測地距離（$\alpha \in \mathbb{R}$）
g_β: $\beta > 0$でパラメトライズされる凹関数
$(\cdot)^\gamma$: L_γ損失関数（$\gamma \geq 1$）

α, β, γそれぞれについて順を追って説明する.

【CAT(0), CAT(k)と内測平均】

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$\|a' - b'\| = d(a, b)$等をみたすとき，測地線上の$p \in \widetilde{bc}$と，$d(b, p) = \|b' - p'\|$をみたす$p' \in \overline{b'c'}$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.

同様に，$k \in \mathbb{R}$についても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もつことに対応する.

CAT(k)空間内の直径$\pi/(2\sqrt{k})$以下の領域上の確率分布はL_γ-内測平均を持つことが証明できる．つまりデータ空間の曲率が小さいほど，内測平均は「より一意」になる.

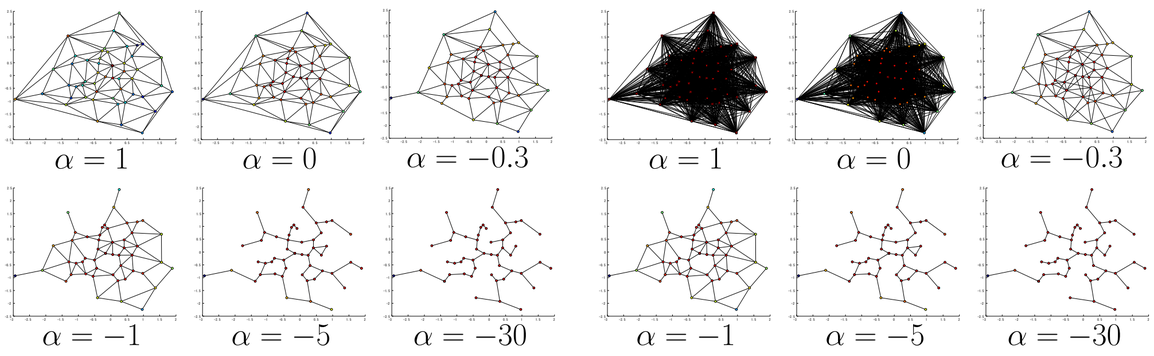
【α距離】

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$\Gamma = \{z(t), t \in [0, 1]\}$の長さを以下で定義する.

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

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，Delaunayグラフ，Gabriel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の値を変化させたときの，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）の移り変わりは以下のようになる.
（左：Delaunayグラフから，右：完全グラフから）

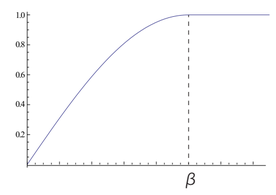


αの増大にともない測地グラフが縮小すること，それにともないCAT(k)となる領域が増大すること（曲率の減少に対応），および極限が最小全張木になることなどを証明した.

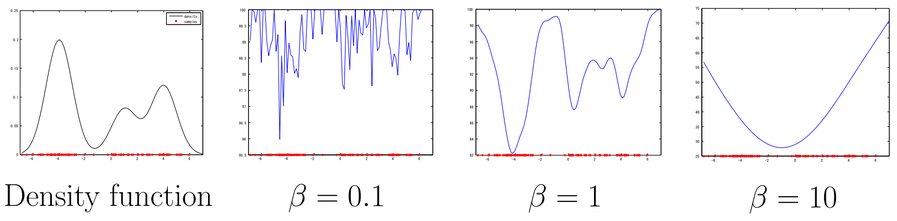
【β距離】

$$d_\beta(x_0, x_1) = g_\beta(d(x_0, x_1)) \ (\beta > 0),$$

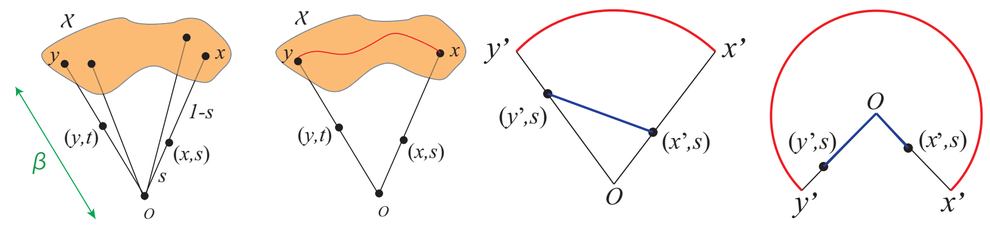
$$g_\beta(z) = \begin{cases} \sin(\frac{\pi z}{2\beta}), & \text{for } 0 \leq z \leq \beta, \\ 1, & \text{for } z > \beta. \end{cases}$$



βの変化により，内測平均の目的関数$f(m) = \sum_i g_\beta(d(x_i, m))^2$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距離による内測平均は，動径βの計量錘（metric cone）における外測平均(extrinsic mean)として解釈できる.



βが大きくなると，計量錘の曲率が小さくなることが証明できる．これにより，データ空間上での目的関数fの極小点が不要に多くなりすぎることを防いでいると期待できる.

【その他】

- γについても，より大きなγの値に対して，CAT(k)になりやすい（曲率が小さくなる）ことが証明できる.
- $\mathbb{R}$上の$\alpha = 1$の内測平均はメディアンとなる．また，β距離はHuber損失関数と同様のロバスト性をもつ．一方，$\gamma < 2$に対するL_γメディアンもロバスト性を持つことが知られている．このように，α, β, γ-内測平均は，平均の一意性とロバスト性のトレードオフをデータ空間の曲率を介して調整していると解釈できる.

- Kei Kobayashi, Henry P. Wynn (2014), Empirical geodesic graphs and CAT(k) metrics for data analysis, arXiv1401.3020.
- Kei Kobayashi, Mitsuru Orita (2014), Permutation test for dendrograms and its application to the analysis of mental lexicons, arXiv1403.2845.