F-028

単体測度に基づく位相多様体の次元推定と多様体学習への応用
Dimension Estimation of Topological Manifolds based on Measure of Simplexes and Application to Manifold Learning

田崎 元† 趙 晋輝†,‡
Hajime Tasaki Jinhui Chao

1. 序論

近年, 機械学習やパターン認識などの研究が盛んに行われており, その多くの分野で音声や画像などの多量の高次元データが扱われる. これらのデータは高次元であるため, 多くの計算時間や記憶領域が必要となる一方で, データ点の集合は, 実質的に低い次元の部分多様体上に存在することが多い. そのため, 高次元データの特徴を学習し, 低次元空間においてそれを再現する多様体学習と呼ばれる手法の研究が進められている. しかし, これまでのデータ多様体の次元推定は, 主成分分析に代表されるように主に部分線形空間で近似するような手法が多く, 多様体の位相構造に対応できない.

本研究では, 単体測度に基づく位相多様体の新しい次元推定手法を提案し, データ集合のもつ位相構造に注目した次元推定を行う. また, 手書き文字のデータに適用し, その有効性を検証する.

2. 次元推定

与えられたデータ点の集合が$\mathbb{R}^D$空間に存在し, そのデータ集合全体がd次元の部分空間に含まれるとき, その部分空間の次元dがデータ集合の実効次元となる. 多様体学習では, 局所線形埋め込み(LLE)[4]や ISOMAP[5]をはじめとする数々の次元削減手法が提案されているが, それらの手法において低次元空間への写像を求めるには, あらかじめ実効次元を求めておく必要があるものが多い. もし, 推定した次元が実際の実効次元より大きければ, 冗長な情報を含むことになり, 一方で小さければ, 次元削減を行った際にデータ集合の情報を損失することになるため, 多様体学習においてデータ集合の実効次元の推定は, 重要な問題の 1 つである. この問題を解決するために, 主成分分析のような射影を用いた推定手法やフラクタルに基づく推定手法[7]などが提案されている. しかし, 主成分分析のような射影を用いた手法は, 多様体が局所線形近似できる場合は有効であるが, 非線形な構造が含まれる場合に適切な次元の推定ができない場合がある.

3. 単体測度

本研究で, 次元推定に利用するr次元単体とは, $\mathbb{R}^n$内の 1 次独立な$r+1$個の点を$\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_r$として,

$$\{\, \boldsymbol{x} \mid \boldsymbol{x} = \sum_{i=0}^{r} \lambda_i \boldsymbol{x}_i, \sum_{i=0}^{r} \lambda_i = 1, \lambda_i \geq 0 \,\}$$

をみたす$\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_r$の集合のことをいい, r-単体と表す. 1-単体は線分, 2-単体は三角形, 3-単体は三角錐を形成し, 一般のn次元まで定義することが可能である[1].

† 中央大学大学院理工学研究科情報工学専攻 Department of Information and Systems Engineering, Graduate School of Science and Engineering, Chuo University
‡ 中央大学理工学部情報工学科 Department of Information and Systems Engineering, Faculty of Science and Engineering, Chuo University

また, r-単体の測度とは, 大きさを表す概念を一般化したものであり, 1-単体の測度は線分の長さ, 2-単体の測度は三角形の面積, 3-単体の測度は三角錐の体積のことをいう. 測度V_rは, $k-1$次元の超平面に含まれない点から$(k-1)$-単体に直交射影したときの長さをh_k として,

$$V_r = \frac{1}{r!} \prod_{k=1}^{r} h_k$$

により求めることができる[8].

4. 提案手法

データ点ごとに近傍を設定して, その近傍内の点で単体を構成し, 構成された単体の測度を求めることで, 多様体の次元推定を行う. r-単体は, 多様体の次元をmとしたとき, $r \leq m$であれば単体の構成が可能であるが, $m < r$のときには, 多様体上の点では, $(r-1)$-単体に直交する方向に点をとることができないため, r-単体を構成することができない. これより, 単体測度は単体を構成できた場合に比べ, 単体を構成できなかった場合は, 著しく小さくなることが期待できる.

提案手法のアルゴリズムを以下に示す.

1) 近傍半径εを入力として与え, $\mathbb{R}^D$におけるデータ点$\boldsymbol{x}_i$ $(i = 1, \dots, N)$の各点に対するε-近傍を設定する. また, この近傍に含まれる点集合を$N(\boldsymbol{x}_i)$とする.
2) 点$\boldsymbol{x}_i$の各近傍に対して, 次の i), ii)を行う.
   i) 1-単体は, 近傍に含まれる 2 点$\boldsymbol{x}_j, \boldsymbol{x}_k$の距離$\|\boldsymbol{x}_j - \boldsymbol{x}_k\|$が最大となる点で構成される. 1 次元の場合のみ 2 点間の距離の最大値は2εであるのに対し, k次元おける高さの最大値は近傍半径εであることから, その整合性のために高さh_1は, $h_1 = \frac{1}{2}\|\boldsymbol{x}_j - \boldsymbol{x}_k\|$とし, $V_1 = h_1$とする.
   ii) r-単体$(r \geq 2)$は, $r-1$次元の超平面に対して, $N(\boldsymbol{x}_i)$に含まれる点$\boldsymbol{x}_j$から直交射影された点を$\boldsymbol{y}_j$とするとき, 最大の$\|\boldsymbol{x}_j - \boldsymbol{y}_j\|$を$h_r$として, r-単体の測度V_rを求める.
3) これまでに求めた測度$V_r (r = 1, \dots, D)$に対して,

$$M_r = \frac{V_r}{\frac{1}{r!}\varepsilon^r}$$

を計算し, 正規化を行う.

この正規化した測度M_rをもとに, 各次元での測度の推移から, 多様体の次元推定を行う. 実際に, M_{d+1}の値がM_dよりも急激に小さくなっているとき, 推定される多様体の次元は, d次元となる.

5. 実験

5.1 概要

今回計測に用いるデータは, 図 1 のような MNIST の手書き文字データセットを利用する. その中から 1 枚の画像

に対し，ランダムに平行移動，もしくは輝度の変化を加えた画像を生成し，データ集合とする．また，画像には，あらかじめガウシアンフィルターを適用し，平滑化処理をしてから計測を行う．この画像データに対して，データ点の個数Nと近傍半径εを設定して提案手法を適用し，得られた測度の平均をもとにグラフから次元を確認する．既存手法との比較として，主成分分析も同様に適用し，得られる固有値から次元を確認する．どちらの手法も縮尺をそろえるために，1 次元での測度あるいは，固有値で各次元の値を割って，グラフにしている．

図 1 手書き文字

5.2 結果

実験による結果を以下に示す．

まず，手書き文字の画像に対して，1 方向(横方向)の平行移動を施した画像に対する次元推定である．データ点の個数を$N = 2000$, 近傍半径を$\varepsilon = 100$とした条件のもとで計測を行った結果を図 2, 図 3 に示す．

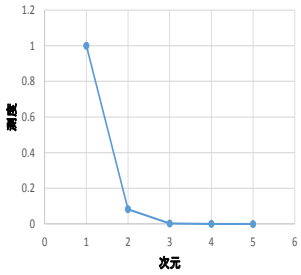

図 2 単体測度の推移

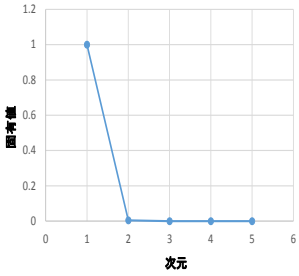

図 3 固有値の推移

図 2 より，1 次元から 2 次元にかけて，測度の大きさが小さくなっていることから，この多様体の局所次元は 1 次元であることがわかる．また，図 3 においても，2 次元での固有値が十分に小さな値になっていることから，データの局所次元が 1 次元であることがわかる．実際に，1 方向の平行移動は，自由度 1 の変形であるため，データ多様体の次元は正しく推定された．

次に，手書き文字の画像に対して，2 方向(横・縦方向)の平行移動を施した画像に対する次元推定である．データ点の個数を$N = 10000$, 近傍半径を$\varepsilon = 150$とした条件のもとで計測を行った結果を図 4, 図 5 に示す．

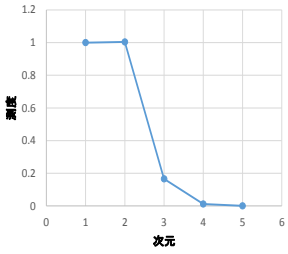

図 4 単体測度の推移

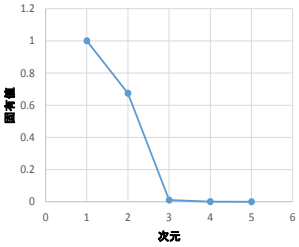

図 5 固有値の推移

図 4 より，2 次元から 3 次元にかけて，測度の大きさが小さくなっていることから，この多様体の局所次元は 2 次元であることがわかる．また，図 5 においても，3 次元での固有値が十分に小さな値になっていることから，データの局所次元は 2 次元であることがわかる．実際に，2 方向の平行移動は，自由度 2 の変形であるため，データ多様体の次元は正しく推定された．

最後に，手書き文字の画像に対し，2 方向の平行移動をさせ，さらにその輝度を変化させた画像に対する次元推定である．データ点の個数を$N = 30000$, 近傍半径を$\varepsilon = 200$とした条件のもとで計測を行った結果を図 6, 図 7 に示す．

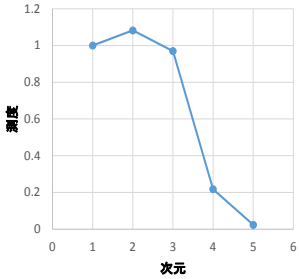

図 6 単体測度の推移

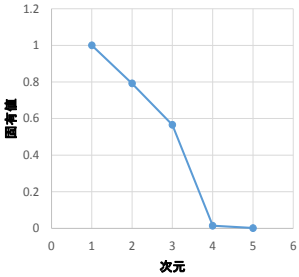

図 7 固有値の推移

図 6 より，3 次元から 4 次元にかけて，測度の大きさが小さくなっていることから，この多様体の局所次元は 3 次元であることがわかる．また，図 7 においても，4 次元での固有値が十分に小さな値であることから，データの局所次元は 3 次元であることがわかる．実際に，2 方向の平行移動と輝度の変化は，自由度 3 の変形であるため，正しく推定された．

6. 結論

本研究では，単体測度を用いた新しい次元推定手法を提案し，平行移動や輝度の変化を加えた手書き文字に適用した．主成分分析が有効に働くと思われる条件のもとで，それによる推定との比較を行い，次元の局所推定が適切に行われていることを確認した．さらに実データへの適用や雑音対策，また異なる次元の複体から構成される多様体の次元推定を可能にするため，複数の近傍による次元推定の手法や位相不変量の計算による位相多様体の構造を検証する手法の検討が今後の課題である．

参考文献

[1] 瀬山 士郎, “トポロジー:柔らかい幾何学”, 日本評論社 (1988)
[2] Yunqian Ma, Yun Fu, “Manifold Learning Theory and Applications”, CRC Press (2012).
[3] C.M.ビショップ, 元田 浩, 栗田 多喜夫, 樋口 知之, 村田 昇, “パターン認識と機械学習 下: ベイズ理論による統計予測”, 丸善出版(2008)
[4] Sam T. Roweis, Lawrence K. Saul, “Nonlinear Dimensionality Reduction by Locally Linear Embedding”, Science, Vol.290, pp.2323-pp.2326 (2000)
[5] Joshua B. Tenenbaum, Vin de Silva, John C. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction”, Science, Vol290, pp.2319-pp.2322 (2000)
[6] Tong Lin, Hongbin Zha, “Riemannian Manifold Learning”, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE VOL.24, NO.5, pp.796-809, (2008)
[7] Francesco Camastra, Allessandro Vinciarelli, “Estimating the Intrinsic Dimension of Data with a Fractal-Based Method”, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL.24, NO.10, pp.1404-1407 (2002)
[8] “Simplex Volumes and the Cayley-Menger Determinant”, http://mathpages.com/home/kmath664/kmath664.htm, (最終閲覧日 2015 年 6 月 23 日)