外部距離によるノンパラメトリック多様体回帰

今泉　允聡　統計的機械学習研究センター　特別研究員

共同研究者：矢野恵佑（東京大学）

【導入：ノンパラメトリック回帰の精度】

設定

入力 $X = (x_1, \dots, x_D) \in \mathbb{R}^D$, 出力 $Y \in \mathbb{R}$ が

$$Y = f^*(X) + \epsilon$$ に従う

($f^*: \mathbb{R}^D \to \mathbb{R}$ 未知の関数, ϵ:ノイズ)

最適な推定の精度

観測 $\{(X_i, Y_i)\}_{i=1}^n$ から f^* を推定するとき、適切な推定量 $\hat{f}_n$ の最適な収束レート[1]は

$$E[\|\hat{f}_n - f^*\|^2] = O\left(n^{-2\alpha/(2\alpha+D)}\right)$$

(α: f^*の微分可能回数)

ノンパラ回帰の高次元問題

入力の次元 D が大きいと収束レートが悪化

⇨**ノンパラメトリック回帰の大きな問題点**

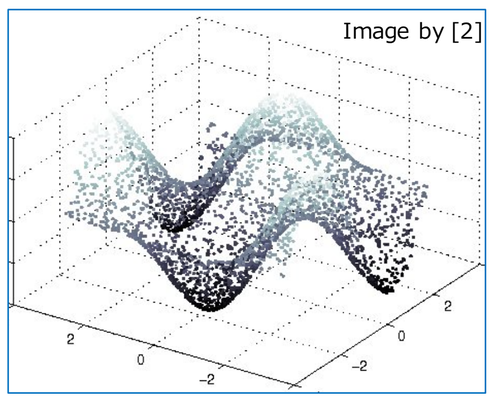
【トピック：多様体データ回帰】

出発点

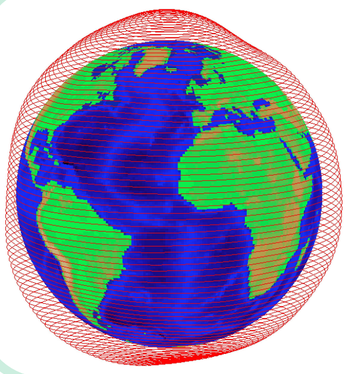
多くのデータは

低次元多様体上にある

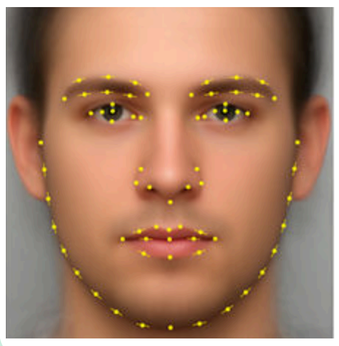
※多様体：局所座標系で被覆できる（局所的に$\mathbb{R}^d$で表現可能）空間



多様体とその次元の例



球面[3]
$D = 3$
$d = 2$



画像（矩形）[4]
D =#pixel
$d = 2\#ad - 4$

e.g.,
#pixel=10000
#ad=20

多様体データ回帰

入力 $Z = (z_1, \dots, z_d) \in \mathcal{M}$と未知関数 $f^o: \mathcal{M} \to \mathbb{R}$ で回帰モデル $Y = f^o(Z) + \epsilon$ を考える

多様体データ回帰の困難さ

データから多様体$\mathcal{M}$を特定するのは困難[4]

e.g., 観測ノイズは多様体をはみ出す / 多様体間の入れ子構造

$\mathcal{M}$特有の距離(測地線)は計算コスト大[5]

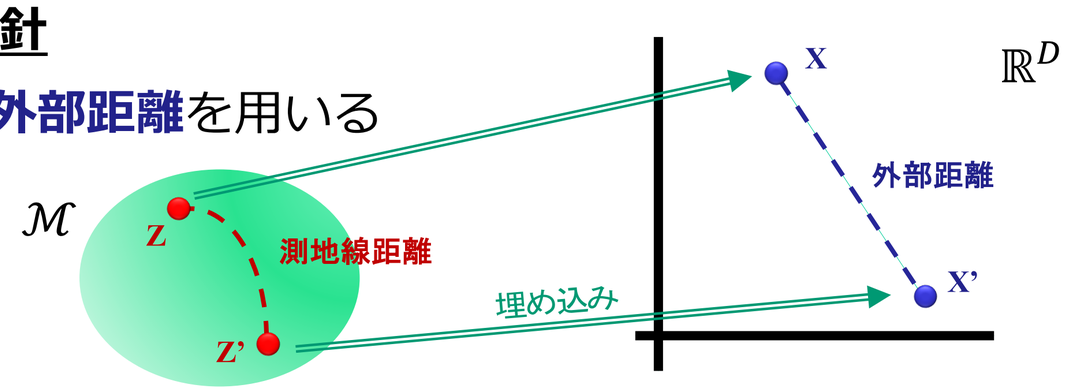
e.g., 球面上の測地線平均の計算はn=30で10秒かかる

⇨**汎用的な理論・手法を作るのが難しい**

【手法：外部距離によるカーネル推定量】

方針

外部距離を用いる



$\mathcal{M}$の測地線距離の代わりに、埋め込み先（$\mathbb{R}^D$）のユークリッド距離で分布を評価

手順

1. 多様体上のデータを$\mathbb{R}^D$に埋め込む

$$X = e(Z) \in \mathbb{R}^D$$

($e: \mathcal{M} \to \mathbb{R}^D$ 埋込写像, $\mathcal{E}$：埋込写像の候補集合)

2. $\mathbb{R}^D$上のユークリッド距離を用いてカーネル推定

$$\hat{f}_{e,n}(Z) := \frac{\sum_{i=1}^n K(h^{-1}\|e(Z_i) - e(Z)\|_2) Y_i}{\sum_{i=1}^n K(h^{-1}\|e(Z_i) - e(Z)\|_2)}$$

（カーネルK：正値かつ指数の速度で単調減少）

3. 埋め込み$e \in \mathcal{E}$の選択

$$\hat{e} := \arg\min_{e \in \mathcal{E}} \sum_{(Z_j, Y_j) \in D'} (\hat{f}_{e,n}(Z_j) - Y_j)^2$$

【提案推定量の利点】

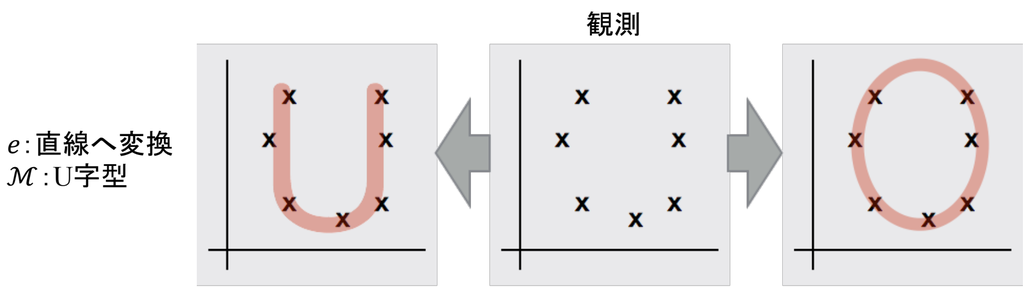
効率性：外部距離の次元に依存しない収束レート

定理1 f^oがγ-Holder連続とすると、 $h = n^{-1/(2\gamma+d)}$ のとき

$$E[\|\hat{f}_{e,n} - f^o\|^2] = O\left(n^{-2\gamma/(2\gamma+d)}\right)$$

- ユークリッド距離なので計算コストは小さい
- 収束レートは埋め込みによって悪化しない

適応性： eの選択を通して$\mathcal{M}$の形式が選択可能



拡張：多様体値関数の推定問題へ拡張可能

$\tilde{f}_n$を写像 $f: \mathcal{M} \to \mathcal{M}'$に拡張したカーネル推定量

定理2 $f^o: \mathcal{M} \to \mathcal{M}'$が$\mathcal{M}'$上の距離$'d_g$に関して$\gamma$-Holder連続のとき、$h = n^{-1/(2\gamma+d)}$ とすると

$$E\left[\int_{\mathcal{M}} d_g(f^o(x), \tilde{f}_{e,n}(x))^2\right] = O\left(n^{-2\gamma/(2\gamma+d)}\right)$$

[1] Tsybakov (2005), Introduction to Nomparametric Estimation, Springer.
[2] http://people.kyb.tuebingen.mpg.de/mmaier/ManifoldDenoising.html
[3] Lai Schmaker (2007), Spline functions on triangulations, Cambridge.
[4] Windhager, et al. (2017). Patterns of correlation of facial shape .... Scientific Reports.
[5] Bhattacharya Bhattacharya (2012), Nonparametric Inference on Manifolds, Cambridge.
[6] Lin Zhu Dunson (2015), Extrinsic local regression on manifold-valued Data, JASA.