高速かつノンパラメトリックな一般化加法モデル学習を用いた関数データ回帰

IBIS2024
2024.11.6
ソニックシティ

武田優真　(東京大学総合文化研究科)
松島慎　(東京大学総合文化研究科)

概要

Function-on-scalar (FOS) 回帰とはスカラーな独立変数から関数形の従属変数を予測するために用いられる回帰手法であり、医療データ分析などで有用性が認められている。FOS回帰は従来、計算コストの高さやパラメトリックな学習に伴う性能の限界があった。本研究では、これらの課題を解決するため、一般化加法モデルを基にした新たな学習手法を提案し、より効率的かつ柔軟な FOS 回帰を実現した。

関数データ回帰

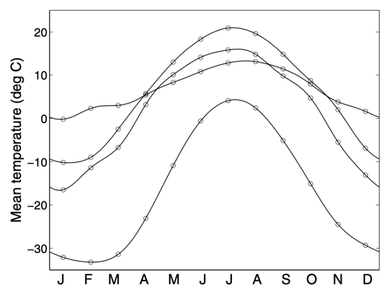
関数データ: 時間や波長など、連続する領域において計測された複数の点データを一つの関数とみなしたもの。
関数データ解析: 関数データを用いたさまざまな分析手法の総称。時系列分析、医療データ分析などで応用が進められている。
メリット:

1. 高次元データで生じる推定値の**不安定化を防ぐ**
2. 欠損がある場合でも**分析が容易**
3. 微分など**関数としての性質**を活かした分析が可能

本研究では関数データを用いた回帰問題の一種である **Function-on-scalar 回帰** (FOS 回帰) に注目する。$\{x_i, y_i(t)\}$ を教師データとして以下のモデルを学習する。

$$y(t) = f(x,t) + \epsilon(t)$$

$$x \in \mathbb{R}^d,\ y(t) \in L^2,\ f(x,\cdot)\colon \mathbb{R}^d \to L^2$$



カナダの4都市における気温変化の関数データ。一つの関数が一つの都市に対応[1]

既存手法

	モデル	計算量
線形モデル[2] **(FOSR)**	$y(t) = \sum_{j=1}^{d} x_j \gamma_j(t) + \epsilon(t)$	$\boldsymbol{O(nd^2M^2)}$
FAMM[3]	$y(t) = \sum_{r=1}^{R} f_r(\mathscr{X}_r, t) + \epsilon(t)$	$\boldsymbol{O(nd^2M)}$
GFOSR[4]	$y(t) = \mu(t) + \sum_{j=1}^{d} \beta_j(x_j,t) + \sum_{k=1}^{q} z_k \gamma_k(t) + \epsilon(t)$	$\boldsymbol{O(n(d+M))}$

n: データ数, d: 説明変数の次元数,
M: 関数データをコンピュータ上で表現するのに用いる点の数

線形モデルや FAMM は計算量が多く、大規模データには適用が難しい。また、いずれの手法も予測関数を基底関数の和として表現し、係数を学習するパラメトリックな手法であり、基底関数の選択により予測性能が制限される可能性がある。

提案手法

以下の加法的なモデルの学習方法を考える。

$$y(t) = \mu(t) + \sum_{j=1}^{d} \beta_j(x_j,t) + \epsilon(t)$$

関数データ $(y_1(t), \ldots, y_n(t))$ に対し、**関数主成分分析 (FPCA)** を用いて計算される正規直交基底関数 ϕ により以下のように近似的に表現できる (Karhunen-Loève 展開) 。

$$y_i(t) \approx \mu(t) + \sum_{l=1}^{L} \xi_{i,l}\phi_l(t) \quad (\text{ただし } \xi_{i,l} = \langle y_i - \mu, \phi_l \rangle)$$

このように、応答変数はデータから計算される基底関数によって表現できる。逆に、　未知の x に対し適切な基底関数の係数 ξ を得ることで $y(t)$ を予測できると考えられる。

提案するモデルは、KL 展開における基底関数を用いて

$$\xi_l(x) = \sum_{j=1}^{d} \beta_{j,l}(x_j) \triangleq \left\langle \sum_{j=1}^{d} \beta_j(x_j,\cdot), \phi_l \right\rangle$$

と書くことができるから、

$$\sum_{j=1}^{d} \beta_j(x_j,t) = \sum_{l=1}^{L} \left\langle \sum_{j=1}^{d} \beta_j(x_j,\cdot), \phi_l \right\rangle \phi_l(t)$$

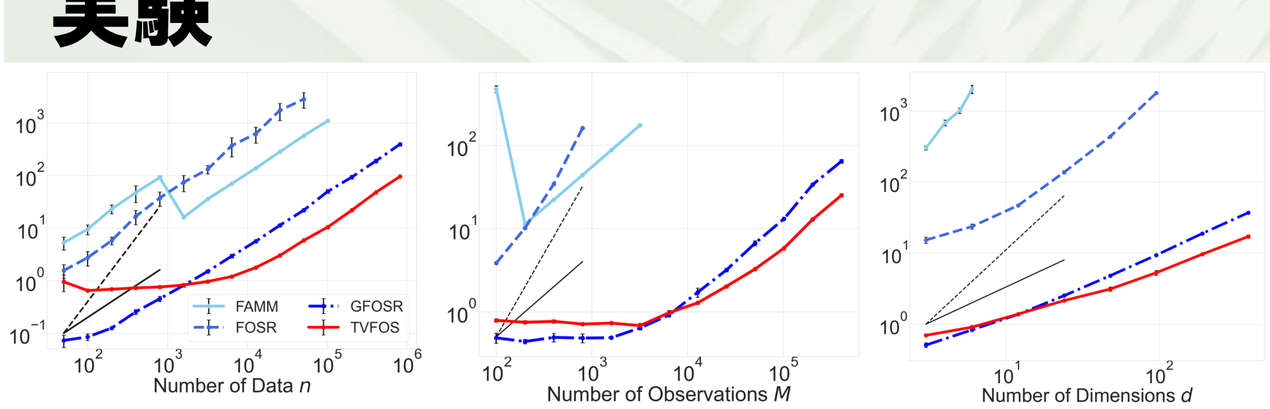
を学習できれば良い。
この関数の学習は $\{x_i, \xi_{i,l}\}$ を教師データとする**一般化加法モデル** (GAM)の学習と見ることができる。高速かつノンパラメトリックな GAM の学習アルゴリズムである**全変動正則化付き一般化加法モデル**[5] (TVGAM) を利用することで

	GFOSR	提案手法
最適化計算量	$\boldsymbol{O(nd)}$	$\boldsymbol{O(n\lceil \frac{d}{\tau} \rceil)}$
$\xi(x)$ の推定	パラメトリック	ノンパラメトリック

(注) GFOSR と提案手法は共通して前処理に KL 展開を利用しており、その計算量は $O(nM)$ である。

既存手法と同等以上に早く、また予測性能が高い学習が可能になると考えられる。
関数の学習における計算量は $\boldsymbol{O(n\lceil \frac{d}{\tau} \rceil)}$ となる (τ は並列化数)。 十分に計算が並列化されれば、この学習の計算量は **FPCA に比べて無視できる**ほどの計算量になる。

実験



左から n、M、d のうち二つを固定し、残り一つを変化させた時の実行時間 (秒) の変化を示している。図中の実線はパラメータに対する線形オーダー、破線は二乗オーダーでの増加を表す。

実験設定

真の関数を設定し、データ数 n、関数データ上の代表点数 M、説明変数の次元数 d を様々に変化させてデータセットを作成した。作成したデータセットに対する既存手法・提案手法の計算時間を測定した。

実験結果

いずれの実験でも、理論的に得られる計算量オーダーと一致する結果が得られた。また、提案手法は全てのパラメータに対し、その数が増加した際に実行時間ベースで GFOSR を上回る結果となり、特に n が十分大きい場合、提案手法は GFOSR のおよそ 4 倍ほど速くなった。

考察

提案手法が GFOSR よりも実行時間ベースで優れているのは実装方法の違いによる部分が大きいと思われるが、提案手法は大規模データに対し有効な手法となっていると考えられる。

今後の展望

FOS 回帰における汎化性能の理論的解析

- 近年、FOS 回帰の手法は様々に提案されているが、提案手法を含め、汎化性能の理論的解析はほとんど進んでいない。
- 一方、別種の関数データを用いた回帰（Function-on-function 回帰）では解析を行なった研究がある[6]。

今後の研究の方向性

- 上記のような研究を参考に、提案手法について汎化性能の理論的解析を行い、予測性能を評価する。
- 得られた理論に関する実験を行う。

補足資料（既存手法詳細、汎化性能についてなど）はこちらから→

参考文献

[1]Ramsay, J. O. and Silverman, W., Functional Data Analysis, Springer, 2005
[2]Reiss., P. T., Huang, L. and Mennes, M., Fast function-on-scalar regression with penalized basis expansions, International Journal of Biostatistics, 2010
[3]Scheipl, F., Staicu, A.-M. and Greven, S., Functional additive mixed models, Journal of Computational and Graphical Statistics, 2010
[4]Ghosal, S. and Maity, M., Variable selection in nonlinear function-on-scalar regression, Biometrics, 2021
[5]Takeda, Y. and Matsushima, S., An accelerated and parallel algorithm for TV-regularized generalized additive model, submitted, 2024
[6]Oliva, J. B., Neiswanger, W., Póczos, B., Xing, E., Trac, H., Ho, S. and Schneider, J., FastFunction to Function Regression, AISTATS, 2015