2022年度共同利用研究報告書

2022年09月30日

所属・職名 滋賀大学データサイエンス学部・准教授

松井 秀俊

			整理番号	2022a008	
1.研究計画題目	大規模高次元データに基づく統計的モデリングとスマート農業への応用				
2.新規・継続	新規				
3.種別	一般研究				
4.種目	短期研究員				
5.研究代表者	氏名	松井 秀俊			
	所属 部局名	滋賀大学データサイエンス学部	職名	准教授	
6.研究実施期間	2022年09月12日(月曜日)～2022年09月16日(金曜日)				
7.キーワード	統計的モデリング，経時測定データ解析，スパース推定				
8.参加者人数	2人				

9.本研究で得られた成果の概要

観測個体それぞれが時間の経過等に伴い複数の観測値を得たデータを、経時測定データという。 経時測定データは多変量データとして扱うことも可能であるが、計測時点や時点数が個体ごとに異なる場合などは、古典的な多変量解析手法を適用することが困難になる場合がある。 これに対して、経時測定データを個体ごとに関数化処理し、得られた関数集合を対象とした分析方法およびその理論は総称して、関数データ解析とよばれている。 申請者はこれまでに関数データ解析に基づく統計的モデリング手法の開発と、さまざまな分野のデータ分析への応用を行ってきた。 特に近年は，農業のデータに関数データ解析を適用することで、農作物の収穫量と、栽培期間中の気温といった環境要因との関係を表現するためのモデリングについて検討している (Matsui, 2020; 2021; 2022)。

本共同研究では、関数データ解析手法の発展や応用先の展開について、IMIの廣瀬慧准教授と議論を行った。 廣瀬准教授は電力需要の予測やマテリアルインフォマティクスの分野のデータ分析の経験があり、相互の知見や成果から新しい成果に繋げられる可能性があると考え、本共同研究に至った。廣瀬准教授との議論により，農業データの分析のさらなる展開に役立つ情報を得ることができた。今後も実際のデータ分析の進捗に応じて、議論を進める予定である。

IMI 共同利用 報告書
「大規模高次元データに基づく統計的モデリングとスマート農業への応用」

滋賀大学データサイエンス学部
松井 秀俊

1　概要

観測個体それぞれが時間の経過等に伴い複数の観測値を得たデータを，経時測定データという．経時測定データは多変量データとして扱うことも可能であるが，計測時点や時点数が個体ごとに異なる場合などは，古典的な多変量解析手法を適用することが困難になる場合がある．これに対して，経時測定データを個体ごとに関数化処理し，得られた関数集合を対象とした分析方法およびその理論は総称して関数データ解析とよばれている．申請者はこれまでに関数データ解析に基づく統計的モデリング手法の開発と，さまざまな分野のデータ分析への応用を行ってきた．特に近年は，農業のデータに関数データ解析を適用することで，農作物の収穫量と，栽培期間中の気温といった環境要因との関係を表現するためのモデリングについて検討している．

本共同研究では，関数データ解析手法の発展や応用先の展開について，IMI の廣瀬慧准教授と議論を行った．廣瀬准教授は電力需要の予測やマテリアルインフォマティクスの分野のデータ分析の経験があり，その知見や成果について議論を行った結果，農業データの分析のさらなる展開に役立つ情報を得ることができた．

本研究で行った議論を，以下にまとめる．

- スパース推定法の 1 つとして，パラメータの積に対する制約を課すことで，いずれかのパラメータのうち一方の値を縮小せずもう一方の値を 0 に縮小する方法について知見を得た．この方法は，農作物の収穫量に関連すると考えられる環境要因の選択に役立つと考えられる．

- 目的変数が複数与えられた場合に用いられる多変量回帰モデルに対して，回帰係数と共分散行列に対して同時にスパース推定を導入することで，目的変数に関連する説明変数と目的変数間の独立性を同時に選択する方法について知見を得た．

- 関数データに基づく回帰モデルにおいて，説明変数に対応する関数データの定義域が観測個体ごとに異なるデータに対して用いられる変動ドメイン関数線形モデルとその推定法について議論を行った．この方法は物性予測にも用いられるのではないかという知見を得た．

- 電力需要の予測に対する関数データ解析アプローチとして，曲線アライメントと関数データに基づく自己回帰モデリングの方法や，関数因子モデルに基づく方法について議論を行った．

上記についての共同研究で議論した内容の詳細について，次節で紹介する．

2　関数データに基づく統計モデリング

関数データ解析では，各個体が経時的に観測値を得たデータに対して，基底関数展開などの方法を使って関数として表現し，得られた関数データ集合を分析対象とする Ramsay and Silverman (2005)．本研究では，関数データに対するさまざまな統計的モデリング手法と，農業データの分析への適用可能性について議論を行った．

2.1　変動ドメイン関数線形モデル

関数データの説明変数と，スカラーの目的変数に関する n 組のデータ $\{(x_i(t), y_i);\ i = 1, \ldots, n,\ t \in \mathcal{T} \subset \mathbb{R}\}$ が得られたとする．このとき，説明変数と目的変数の関係を表す関数線形回帰モデルは，次で与えられる (Ramsay and Silverman, 2005)．

$$y_i = \beta_0 + \int_{\mathcal{T}} x_i(t)\beta_1(t)dt + \varepsilon_i. \tag{1}$$

ここで，β_0 は切片，$\beta_1(t)$ は係数を表す関数で，いずれもパラメータである．また，ε_i は誤差を表す確率変数とする．$\beta_1(t)$ に対しては，基底関数展開で表される仮定を置くことで，古典的な最小二乗法などの枠組みで推定できる．(1) 式のモデルを推定することで得られる $\beta_1(t)$ の推定値から，説明変数 $x_i(t)$ が，任意の時点 t において目的変数 y_i にどのように関連しているかを定量化できる．

関数回帰モデル (1) 式では，関数データは $x_i(t)$ は各 i で共通のドメイン $\mathcal{T}$ 上で与えられていることを仮定した．一方で，例えば複数の苗を生育したとき，作物が収穫できる時期が苗によって異なるように，データによっては i に応じてドメインが異なる場合がある．いま，関数データ $x_i(t)$ のドメインが $\mathcal{T}_i = [0, T_i]$ で与えられたとする．このようなデータに対して，Gellar et al. (2014) は，次のモデルを用いることを提案した．

$$y_i = \beta_0 + \int_0^{T_i} x_i(t)\beta_1(t, T_i)dt + \varepsilon_i. \tag{2}$$

ここで，$\beta_1(t, T_i)$ は 2 変数関数で与えられる係数関数である．また，$t \leq T_i$ であることに注意されたい．これより，ドメインの終点 T_i が各 i で異なっていても，ドメインの長さに応じた回帰係数の推定が可能になる．変動ドメイン関数線形モデル (2) 式は，2 変数の係数関数のドメインが三角形の形状をしていることから，(1) 式のモデルと同様の推定法を適用することが困難である．本共同研究では，(2) 式のモデルの推定方法と，農業データの分析への応用可能性について議論を行った．加えて，廣瀬准教授が携わっている材料科学への応用への適用可能性についても検討した．

2.2　関数因子モデル

ある時点の電力価格は，その時間帯の電力需要に依存して変動する．これらの時系列データから 1 日ごとの関係性を表現する方法として，Liebl (2013) は次の関数因子モデルを提案した．いま，電力需要量に対応する変数を u とする．このとき，第 i の電力価格に対応する関数を次で表現する．

$$x_i(u) = \sum_{k=1}^{K} \beta_{tk} f_k(u), \quad u \in [a_i, b_i]. \tag{3}$$

ここで，β_{tk} は未知パラメータ，$f_k(u)$ は正規直交基底とする．また，a_i, b_i はそれぞれ第 i 日の電力需要量の下限と上限とする．データから β_{tk} および $f_k(u)$ を適切に推定することで，第 i の電力価格の関係性を定量化できる．本共同研究では，関数因子モデル (3) 式の農業データの分析への適用可能性について議論を行った．

2.3 関数線形モデルに対する Prenet の適用

p 個の特徴量からベクトル $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ に対して，未観測の潜在要因を表現する因子分析モデルは次で与えられる．

$$\boldsymbol{X} = \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\varepsilon}.$$

ただし，$\boldsymbol{\Lambda} = (\lambda_{ij})_{ij}$ は $p \times m$ 因子負荷行列，$\boldsymbol{F} = (F_1, \ldots, F_m)^T$ は m 次元共通因子ベクトルとする．また，$\boldsymbol{\varepsilon}$ は p 次元独自因子ベクトルとする．因子分析モデルに対しては，最尤法などが古典的な推定法として用いられている．これに対して，Hirose and Terada (2022) は，Prenet とよばれる，次の制約を課した正則化最尤法に基づき因子負荷量 λ_{ij} を推定する方法を提案した．

$$P(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \sum_{j=1}^{m-1} \sum_{k>j} \left\{ \gamma |\lambda_{ij}\lambda_{ik}| + \frac{1}{2}(1-\gamma)\lambda_{ij}^2\lambda_{ik}^2 \right\}.$$

これにより，最尤法よりも単純かつ解釈が容易な因子分析モデルを構築できる．関数因子モデル (3) 式の推定に prenet を適用することで，効率的に因子を推定する方法を確立できるのではないかと考えられる．

以上のように，関数データに基づく統計モデリングは農業だけでなく，さまざまな分野のデータ分析への適用の可能性を有している．本共同研究は，その具体的なアイデアを複数議論できた点で，非常に有意義なものとなった．

References

Gellar, J. E., Colantuoni, E., Needham, D. M., and Crainiceanu, C. M. (2014). Variable-Domain Functional Regression for Modeling ICU Data Jonathan. *J. Amer. Statist. Assoc.*, 109(508):1425–1439.

Hirose, K. and Terada, Y. (2022). Sparse and simple structure estimation via prenet penalization. *Psychometrika*, To appear.

Liebl, D. (2013). Modeling and forecasting electricity spot prices : a functional data perspective. *Ann. Appl. Statist.*, 7(3):1562–1592.

Ramsay, J. and Silverman, B. (2005). *Functional data analysis (2nd ed.)*. Springer, New York.