大規模な地理空間データのための空間混合効果モデリング

統計数理研究所・データ科学研究系　　村上大輔

1. はじめに

観測技術の発達に伴う地理空間データの大規模化は著しく、大規模な地理空間データ（標本数：数千～数百万）を取り扱うための回帰モデリングが重要となってきている。そこで本研究では、空間効果（残差や回帰係数の空間相関等）とそれ以外の効果（グループ効果、非線形効果等）を大規模な地理空間データから推定するための新たな空間回帰手法を開発した。

2. 手法

共変量$\mathbf{z}_p$からの効果を関数$f(\mathbf{z}_p) = \mathbf{A}_p\mathbf{V}(\boldsymbol{\theta}_p)\mathbf{u}_p$で推定しようという以下の加法モデルを考える:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{p=1}^{P} \mathbf{A}_p\mathbf{V}(\boldsymbol{\theta}_p)\mathbf{u}_p + \boldsymbol{\varepsilon} \qquad \mathbf{u}_p \sim N(\mathbf{0}, \tau_p^2\mathbf{I}_p) \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

ここで$\mathbf{A}_P$はP番目の共変量$\mathbf{z}_p$からの効果を表現するための基底関数行列$(N \times M_p)$、$\mathbf{V}(\boldsymbol{\theta}_p)$は各基底関数からの効果を決める行列$(M_p \times M_p)$、$\boldsymbol{\theta}_p$はハイパーパラメータを表す。$\mathbf{z}_p$を緯度経度で与えることで、(a)残差の空間相関を捉えるためのガウス過程モデルや(b)共変量からの場所毎の効果を捉えるための空間可変パラメータモデルを上式に組み込むことができる。即ち上式を用いることで空間効果・非空間効果を考慮した柔軟な回帰モデリングが可能である。一方で、空間効果(a)、(b)を精度良く捉えるには基底数M_pを大きくする必要がある。さらに N が巨大な場合には各$\mathbf{A}_p$が巨大な行列となる。上式はメモリ消費・計算効率の両観点で課題が残されている

そこで、上式を高速推定するために、本研究では、(i)次元が N に依存する行列を予め尤度関数（Type II）から除外して、(ii)同尤度関数を各ハイパーパラメータについての逐次計算で最大化する方法を開発した。手順(i)の結果、パラメータ推定の計算量はNに依存しない。また(ii)の結果、M_pが大きくても計算効率よくモデル推定できる。また、手順(i)における処理をデータのサブセット毎に並列処理しており、メモリ消費はNに依存しない。以上より、計算量、メモリ消費の両効率の良いアルゴリズムが確立できた。

3. 計算速度の比較

提案手法によるモデル推定時間を、大規模データの回帰モデリング用関数である bam 関数（mgcv パッケージ）と比較した（図 1）。この結果よりNの増加に伴う提案手法の計算時間の増大は bam 関数よりも小さく、計算効率の良さを確認した。

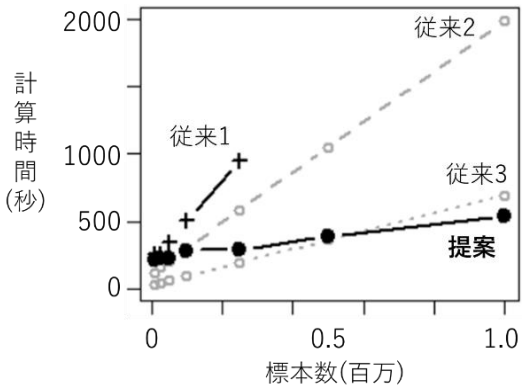

図 1 ：計算時間の比較。従来 1～3 は bam 関数で推定