空間回帰モデルのパラメータ推定における問題

力丸　佑紀　　リスク解析戦略研究センター　特任准教授

空間データとは

「空間」といわれると，何をイメージするだろうか．

宇宙は最大の空間であるし，人間の細胞も小さな小さな空間である．多種多様な空間で取得されたデータはすべて一般に空間データと呼ぶことができる．

1854年にイギリスでコレラが大流行した際に，ジョン・スノウ医師がコレラによる死亡者を地図上にプロットし，水道ポンプの周辺に死亡者が多いことをつきとめ，ポンプを止めたところ，感染拡大が収まったという話がある．これが空間情報を取り入れた問題解決のはじまりだと言われており，空間的な広がりの認識や空間データの重要性がよくわかる例である．近年では，技術の進歩により，スマートフォンなどを通して得られる国民の位置情報や人工衛星から送られる大規模な宇宙温度データなど，より複雑な構造の空間データが増え，その分析の需要も増えてきている．

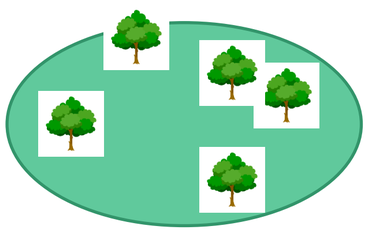
空間データとひとくちに言っても，そのあり方は様々であるが，ここでは空間データを以下のように分類する．

1. point pettern

データ：イベントの発生位置 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$

例：森林で木が生える位置

分析目的：イベントの発生メカニズムを探る等



2. values on a space

データ：固定点で観測された値 $\{Z\_\boldsymbol{v}, \boldsymbol{v} \in \boldsymbol{Z}^2\}$

例：宇宙温度，降水量，地価

分析目的：値の分布や相関構造を探る等

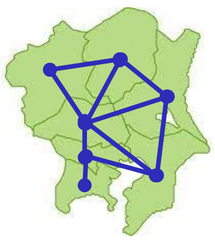
3. graphical data

データ：各区域 $\boldsymbol{s}_i$ で観測された値 $\{Z(\boldsymbol{s}_i)\}$

位置情報なし，各区域の配置情報あり

例：疾病地図

分析目的：各地域のつながりを探る等



空間データから，人や生物の動き，天気の移り変わり，温度分布，地震解析など，互いの関係や空間内での動き，発生メカニズムなどを読み解いていくための方法のひとつとして，空間点過程や空間自己回帰モデル，グラフィカルモデリングやネットワークモデリングなどの空間データモデリングがあるが，本研究は特に values on a space の空間的な相関関係をモデル化するための空間回帰モデルとそのパラメータ推定における問題に注目する．

空間回帰モデルの問題の発見

values on a space の回帰モデルとしてSLM(Spatial Lag Model)

$$\boldsymbol{y} = \lambda W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$$

がよく用いられる．これは，$S(\lambda) = I - \lambda W, X(\lambda) = S(\lambda)^{-1}X$ とすれば，

$$\boldsymbol{y} = X(\lambda)\boldsymbol{\beta} + \boldsymbol{u}, \qquad S(\lambda)\boldsymbol{u} = \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

のように書き換えられる．ただし，$\boldsymbol{y}$ は観測値ベクトル，W は空間重み行列である．パラメータ $(\boldsymbol{\beta}, \lambda, \sigma)$ の推定には，このモデルのもとでの対数尤度

$$L = \log\det S(\lambda) - \frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - X(\lambda)\boldsymbol{\beta})^{\top} S(\lambda)^{\top} S(\lambda)(\boldsymbol{y} - X(\lambda)\boldsymbol{\beta})$$

を最大化する最尤推定量が漸近有効であるので良いと信じられてきた．

しかし，ここには大きな落とし穴が潜んでいる．一貫性のない誤差 $\boldsymbol{u}$ を前提として証明される漸近有効推定量だからである．実際，W を単純な隣接行列としても，$\boldsymbol{u}$ の部分ベクトルの分散共分散行列が $\boldsymbol{u}$ の分散共分散行列 $\sigma^2 S(\lambda)^{-1}(S(\lambda)^{-1})^{\top}$ の部分行列とはならない．つまり，観測領域を変えるとデータをとってきた空間が変わるような設定になっているのである．これは，$\boldsymbol{u}$ の自己回帰を λW で表すとエッジエフェクトが生ずることが原因で，その影響が漸近的に無視可能なら問題ないが，空間的な広がりがある空間データの場合には理論的にどうしても無視できない大きさになる．

誤差が定常性をもつ空間回帰モデルを用いた解決

誤差 $\boldsymbol{u}$ に一貫性をもたせる方法はいくつかあるが，ここでは，$\boldsymbol{u}$ は定常性をもつ Whittle の SAR モデル

$$P(T_1, T_2)u_{\boldsymbol{v}} = \varepsilon_{\boldsymbol{v}}, \qquad P(T_1, T_2) = 1 - \lambda \sum_{\boldsymbol{k} \in K} w_{\boldsymbol{k}} T_1^{k_1} T_2^{k_2}$$

に従うと仮定する．ただし，$P(T_1, T_2)$ は行列演算 $S(\lambda)$ を反映した伝達関数であり，$E\left(u_{\boldsymbol{v}} u_{\boldsymbol{v}'}\right) = \gamma_{\boldsymbol{v}-\boldsymbol{v}'}$ である．

このように一貫性のある誤差 $\boldsymbol{u}$ を仮定した場合，L から得られる最尤推定量 $\widehat{\boldsymbol{\beta}}$, $\hat{\lambda}$ は，実は

$$E\left(\frac{N}{n_1 + n_2}(\hat{\lambda} - \lambda)\right) \longrightarrow \frac{-\sum_{t \neq 0} \gamma_{1,2t} + \sum_{t \neq 0,-1} \gamma_{0,2t+1} - \lambda\left(2\sum_{t \neq 0,1} \gamma_{0,2t} + \sum_{t \neq 0} \gamma_{2,2t} + 8\sum_{t \neq 0,-1} \gamma_{1,2t+1}\right)}{4(\gamma_{2,0} + \gamma_{0,0} + 2\gamma_{1,1}) + C}$$

$$E\left(\hat{\beta}(\hat{\lambda}) - \beta\right) = E\left(\hat{\beta}(\hat{\lambda}) - \hat{\beta}(\lambda)\right) + E\left(\hat{\beta}(\lambda) - \beta\right) = E(\hat{\lambda} - \lambda)\ (X^{\top}X)^{-1}X^{\top}WS(\lambda)^{-1}X\beta$$

のように漸近バイアスをもつため，そのまま使うのは危険である．

では，誤差 $\boldsymbol{u}$ に定常性を仮定して一貫性をもたせた上で良い推定量を得るにはどうすればよいか．そのひとつの方法として，

$$L_A = \frac{1}{2}\log\det A - \frac{N}{2}\log 2\pi - \frac{1}{2}(\boldsymbol{y} - X(\lambda)\boldsymbol{\beta})^{\top}\tilde{A}\ (\boldsymbol{y} - X(\lambda)\boldsymbol{\beta})$$

を近似尤度として使うことを提案する．ただし，A は $S(\lambda)^{\top}S(\lambda)/\sigma^2$ を反映させた巡回行列であり，$\tilde{A}$ は A の和の各項に縮小率を導入したものである．X に関して一般的な回帰モデルと同様の仮定をいくつか置けば，この近似尤度を最大化することによって得られる推定量は一致性，漸近有効性をもつ．