高次元空間誤差モデル

北里大学　力丸佑紀
慶應義塾大学　柴田里程

空間データモデル

- さまざまなモデル
  - SAR (Simultaneous Autoregressive) モデル
  - CAR (Conditional Autoregressive) モデル
  - SLM (Spatial Lag Model)
  - SEM (Spatial Error Model)
  - ・・・

どのモデルを使うか？
どうモデルを使うか？

モデルの意味

- 平均に関するモデル
  1. 期待値 $\mu = E(y)$ が位置関係とは独立に決まる
  2. 期待値 $\mu = E(y)$ が位置関係の影響を受ける

- 分布に関するモデル
  a. Simultaneous AR
     - サイトの値を周りの値の線形結合で表す
     - 共分散だけでは決まらないが，近傍を定めると決まる

$$\boldsymbol{Z} = B\boldsymbol{Z} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$
$I - B$: 正則

  b. Conditional AR
     - 周りの値を条件にして分布が決まる
     - 偏相関であるc_{ij}が周辺の値の条件付き分布への影響度を表している
     - 枠組みとしてはグラフィカルモデリングと同じ

$$\boldsymbol{Z} = C\boldsymbol{Z} + \boldsymbol{v}, \boldsymbol{v} \sim N(0, \tau^2(I - C)^T)$$
C: 対称，$\lambda_{max}(C) \leq 1$

  c. Independent

モデルの選択

1. μ が周囲のサイトの影響を受けない

1a (**SEM**; Simultaneous Error Model)

$$g(\boldsymbol{\mu}) \equiv \boldsymbol{\mu}$$

$$V(\boldsymbol{\mu}) \equiv \left((I-B)^T(I-B)\right)^{-1}$$

1b (**CEM**; Conditional Error Model)

$$g(\boldsymbol{\mu}) \equiv \boldsymbol{\mu}$$

$$V(\boldsymbol{\mu}) \equiv (I-C)^{-1}$$

ただし，Cは対称で，$\lambda_{max}(C) \leq 1$

1c (**IEM**; Independent Error Model)

$$g(\boldsymbol{\mu}) \equiv \boldsymbol{\mu}$$

$$V(\boldsymbol{\mu}) \equiv I$$

1a,1bの特殊ケース

2. μ が周囲のサイトの影響を受ける

2a (**eSLM**; extended Simultaneous Lag Model)

$$g(\boldsymbol{\mu}) \equiv (I-B_1)\boldsymbol{\mu}$$

$$V(\boldsymbol{\mu}) \equiv \left((I-B_2)^T(I-B_2)\right)^{-1}$$

2b (**eCLM**; extended Conditional Lag Model)

$$g(\boldsymbol{\mu}) \equiv (I-C_1)\boldsymbol{\mu}$$

$$V(\boldsymbol{\mu}) \equiv (I-C_2)^{-1}$$

ただし，C_2は対称で，$\lambda_{max}(C_2) \leq 1$

2c (**ILM**; Independent Lag Model)

$$g(\boldsymbol{\mu}) \equiv (I-B)\boldsymbol{\mu}$$

$$V(\boldsymbol{\mu}) \equiv I$$

2a, 2bの特殊ケース

データ	被説明変量	説明変量	実際の使用例	まず試すべきモデル	ポイント
オハイオ住宅市場データ	住宅販売価格	建築年，階数，居住面積，壁タイプなど	[**eSLM(2a)**] ・$(\mathrm{I}-\rho W)\boldsymbol{\mu}=X\boldsymbol{\beta}$ ・$B_2=\rho W$	[**SEM(1a)**] ・$\boldsymbol{\mu}=X\boldsymbol{\beta}$ ・$B=\rho W$	建築年など物件の客観的条件は位置関係に依存しない.
空気の質データ	空気の質の指数（AQI）など	都市の大きさ，都市の形状など	[**SEM(1a)**] ・$\boldsymbol{\mu}=X\boldsymbol{\beta}$ ・$B=\rho W$ [**eSLM(2a)**]	[**eSLM(2a)**] ・$(I-\rho W)\boldsymbol{\mu}=X\boldsymbol{\beta}$ ・$B_2=\rho W$	都市データは位置関係に依存する.
大統領選挙投票率データ	選挙投票率	19歳以上の住民の所得，大卒者の割合，持ち家の割合	[**SEM(1a)**] [**eSLM(2a)**] [**Spatial Durbin Model(2a)**] ・$(I-\rho W)\boldsymbol{\mu}=X\boldsymbol{\beta}+W\boldsymbol{\gamma}$ ・$B_2=\rho W$	[**IEM(2c)**] ・$(I-\rho W)\boldsymbol{\mu}=X\boldsymbol{\beta}$	まず偏差は独立であると思って平均モデルの精度を高める.

データ	被説明変量	説明変量	実際の使用例	まず試すべきモデル	ポイント
SIDSデータ	乳幼児突然死症候群での死亡数	出生数，非白人出生数	[**CEM(1b)**] ・Freeman-Tukey変換 ・$g(\boldsymbol{\mu}) = X\boldsymbol{\beta}$ ・$C = \rho W$	・独立ポアソン分布 ・$(I - B)\log \boldsymbol{\mu} = X\boldsymbol{\beta}$	感染症だとしたら？
疾病発生数データ	呼吸器疾患の入院数	COなどの大気汚染データ，求職者手当の受給率，都市レベルなど	・条件付ポアソン分布	・独立ポアソン分布 ・$(I - B)\log \boldsymbol{\mu} = X\boldsymbol{\beta}$	平均のモデル化だけでよいのでは？
アカギデータ	アカギの在セル数	標高，傾斜角度，集水域面積，スカイラインなど	・条件付二項分布	・独立二項分布 ・$(\mathrm{I} - B)\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = X\boldsymbol{\beta}$	

データ	被説明変量	説明変量	実際の使用例	まず試すべきモデル	ポイント
川のリン酸塩濃度データ	リン酸塩濃度	座標	[**CEM(1b)**] ・$\boldsymbol{\mu = X\beta}$ ・$C = \rho W$	[**SEM(1a)**] ・$\boldsymbol{\mu} = \beta \mathbf{1}$ ・$B = \rho W$	川の場所によって基本的な値は変わる．リン酸塩濃度は周りの値に影響されて決まる． 時間も本格的に入れる可能性．
差し押さえ率データ	差し押さえ率	緯度，経度，売却日，失業率など	$\boldsymbol{y = X\beta + z + \varepsilon}$ ・$\boldsymbol{z}$ が**CEM(1b)** ・$\boldsymbol{\varepsilon}$ は独立	[**IEM(1c)**] ・$\boldsymbol{\mu = X\beta}$	「空間」がこのモデルに必要かよく考える．

CEM, CLM

- CEM, CLMは使う必要がない
  - Cは対称
    - 影響の方向性がない．現実的ではない．
  - モデルの適合度のチェックが難しい．
    - 誤差に相当するものがない．
    - 繰り返し観測ができない．
    - 地理的な近傍から作ったCEM, CLMを前提としたモデルでデータの共分散をうまく近似できるかどうかわからない．
  - 解釈も難しい．
    - 周りの値によって条件付き分布が決まるというモデル．

オハイオ住宅市場データ

- オハイオ州ルーカス郡で1993年～1998年の間に販売された一戸建て住宅の記録2537件
  - Spatial Econometrics tool boxのdata/house.txtにある.
    - 住宅販売価格
    - 建築年，階数，居住面積，壁タイプ，ベッドルーム数，バスルーム数，ハーフバスルーム数，間口，奥行き，ガレージタイプ，ガレージ面積，部屋数，敷地面積，販売年月日，評価額，築年数

（使用例）“Variable selection via penalized quasi-maximum likelihood method for spatial autoregressive model with missing response”, 2024, Yuanfeng Wang, Song, *Spatial Statistics.*

- $E(y)$を築年数などでモデル化するとしたら，位置関係に依存しない.
- 偏差 $y - E(y)$ は，各位置は同等で，注目するサイトの値が周りの値によってどう定まるかに注目.
- 住宅市場において，物件の客観的条件以外の部分が偏差のモデルで表せる.

空気の質データ

- 2015年，中国の288の都市データと1333か所の空気観測所での1時間ごとの空気の質の観測値
  - 都市計画と管理方針の最適化のため，都市の空気の質と都市形態の関係を探りたい.
    - AQI（空気の質を表す指標），6つの基準汚染物質（PM2.5, PM10, CO, SO2, NO2, O3）
    - 都市被覆データから計算した都市の大きさ，都市の形状，都市の断片化，交通アクセス，スプロール現象の値を計算

（使用例）“Effects of urban form on air quality in China: An analysis based on the spatial autoregressive model”, 2019 , Fan Li, Tao Zhou, *Cites*

- 都市データは位置関係に依存する.
- 偏差も当然，位置関係に依存する.

大統領選挙投票率

- 1980年のアメリカの大統領選挙の各郡の投票率3107件
  - Spatial Econometrics Toolbox の data/elect.txtにある
    - 19歳以上の住民の投票率
    - 19歳以上の住民一人当たりの所得，19歳以上の人口に対する大卒者の割合，19歳以上の人口に対する持ち家の割合

（使用例）“Comparing Estimation Methods for Spatial Econometrics Techniques using R.”, 2010, Bivand, R., *NHH Dept. of Economics Discussion Paper (26)*.

- E(y)を所得などでモデル化するとしたら，位置関係に依存する．所得が高いエリアの近傍のエリアでも所得は高い．このモデルの精度を高めるべき．
- 偏差 y-E(y) は独立ではないか．投票率そのものが位置関係に依存して決まるか？
- 分布に関しては独立としてやって，平均のモデルの精度を高めて，それでも表現でいない部分は分布のモデルで検討する．

SIDS

- 1974年7月1日～1978年6月30日，1979年7月1日～1984年6月30日の２期間において，ノースカロライナ州の各郡におけるSIDS（乳幼児突然死症候群）のデータ100件
  - SIDSの原因は不明なので解明したい．
    - SIDSの数
    - 出生数，非白人出生数

（使用例）”Spatial Modeling of Regional Variables”, 1989, Noel Cressie and Ngai H. Chan, *Journal of the American Statistical Association*, Vol. 84, No. 406, pp.393-401

- 平均でのモデル化だけでよいのでは？
- E(y)をモデル化して，ポアソン分布で考える．出生数などでモデル化するとしたら，位置関係に依存する．
- 偏差 y-E(y) は独立ではないか？まずは平均のモデル化を丁寧に取り組む．
- 感染症の場合はまた別に考える必要がある．

疾病発生数

- スコットランドのグラスゴー市とクライド河口からなる地域の271の行政単位における，呼吸器疾患を主診断とする非精神科・非産科病院への入院数
  - 大気汚染と呼吸器疾患の関係を探りたい．貧困と不健康の関係も考慮している．
    - 入院数
    - 大気汚染データとしてCO, NO2, SO2, PM10, PM2.5，求職者手当の受給率，学童の非白人の割合，診療所までの車での平均移動時間，都市レベル

（使用例）”A Bayesian Localized Conditional Autoregressive Model for Estimating the Health Effects of Air Pollution”, 2014, Duncan Lee, Alastair Rushworth and Sujit K. Sahu, *Biometrics,* 70(2), 419-429

- やはりポアソン分布で考える．
- 大気汚染データなどは位置関係に依存する．
- 入院数は位置関係に依存しないのでは？

アカギ

- 2003年，2321の100mメッシュの中のアカギの林冠個体の分布
  - 野外生物の分布パターンを知りたい.
  - 近い地点同士ほど残差が類似するため，残差の非独立性を表現したい.
    - 100mメッシュの中のアカギの在セル数
    - 標高，傾斜角度，集水域面積，スカイライン，1977年分布域からの距離

（使用例）「条件付自己回帰モデルによる空間自己相関を考慮した生物の分布データ解析(<特集2>始めよう!ベイズ推定によるデータ解析)」，2009, 深澤　圭太・石濱　史子・小熊　宏之・武田　知己・田中　信行・竹中　明夫，日本生態学会誌 59 (2), 171-186

- やはりポアソン分布で考える.
- 偏差ではなく，E(y)が位置関係に依存すると考えてモデル化する.

リン酸塩濃度

- ギリシャのラコニア地方のエプロタス川で16×16の正方形の格子上で数年間測定したリン酸塩濃度
  - リン酸塩濃度
  - 座標

（使用例）”Bayesian analysis of conditional autoregressive models”, 2012, Victor De Oliveira, *Ann Inst Stat Math* , 64, 107–133

- 平均というのは川の場所や環境によって変わる基本的な値だと考える.
- 数年間測定だから時間や季節の影響もランダムとして入っているかもしれない.

→ 結果によっては時間も本格的に入れる可能性.

- どこまでが平均でどこからがランダムか.

差し押さえ率

- 2005年～2014年のミシガン州ウェイン郡の全住宅取引に占める差し押さえ率データ370517件
  - 差し押さえ率
  - 緯度，経度，売却日，失業率など

（使用例）“On the formal specification of sum-zero constrained intrinsic conditional autoregressive models”, 2018, Matthew J. Keefe, Marco A.R. Ferreira, Christopher T. Franck , *Spatial Statistics*, 24,54–65

- 「Spatial」に騙されてはいけない！確かにSpatialなデータだけど，Spatialであることがモデルに必要なのか？をまずよく考える.
- Spatialであることに拘るから本質が見えなくなる・・・