2019年6月5日　統計数理研究所　オープンハウス

事前圧縮による地理的加重回帰の高速化

村上　大輔　　データ科学研究系　助教

【背景】

場所毎の回帰係数(SVC: Spatially varying coefficients)を推定する方法である Geographically weighted regression、(GWR)は応用分野で幅広く用いられてきた。
しかしながらGWRの推定に必要なLeave-one-out cross-validation (LOOCV)の計算量は標本数Nの2乗のオーダーであり、大規模な空間データには不向きである。

高速なSVC推定手法の比較

SVCを高速推定する数多くの手法が存在するが、
過度な平滑化を招かず(非縮退)かつ高速にSVCを推定する方法が存在しない

手法	推定法	計算量	非縮退
Fast GWR (Li et al., 2019)	局所	$O(N^2)$	○
Expansion method (Casetti, 1972)	大域	$O(N)$	×
Bivariate spline method (Mu et al., 2018)			
ME approach (Murakami and Griffith, 2019)			
Predictive process (Finley et al., 2011)		$O(N)$だがMCMCが必要。遅い。。	
Nearest-neighbor GP (Finley et al., , 2019)	局所		○

→ そこで本研究ではGWRを高速化(＋**Rパッケージscgwrで公開**)

提案型GWR	局所	$O(N)$	○

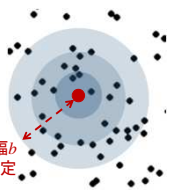
【GWR】

- 地点iの回帰係数を下式で推定する
- LOOCVでバンド幅を最適化した後に回帰係数を推定

$$y_i = \sum_{k=1}^{K} x_{i,k}\beta_{i,k} + \varepsilon_i \qquad \begin{matrix} E[\varepsilon_i] = 0 \\ Var[\varepsilon_i] = \sigma^2 \end{matrix}$$

カーネル$g(d_{i,j};b)$のイメージ

$d_{i,j}$:地点i-j間の直線距離
b　:バンド幅。影響の空間範囲を決める

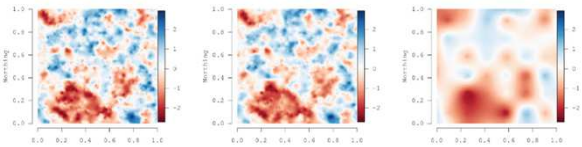


地点iの回帰係数
- 近隣の標本に大きな重みを与えて推定

$$\hat{\beta}_{i,k} = \underset{\beta_{i,k}}{\operatorname{argmin}} \sum_{j=1}^{N} g(d_{i,j};b)\left(y_j - \sum_{k=1}^{K} x_{j,k}\beta_{i,k}\right)^2$$

推定法と縮退
- **大域**:空間基底関数L個($L<<N$)の線形和でSVCをモデル化(i.e., $\beta=E\gamma$)
  → 計算負荷は小さいが、**縮退する(SVCが過度に平滑化される)**
- **局所**:場所毎の局所モデルを推定
  →**縮退しない**。並列化しやすい

Datta et al. (2016)

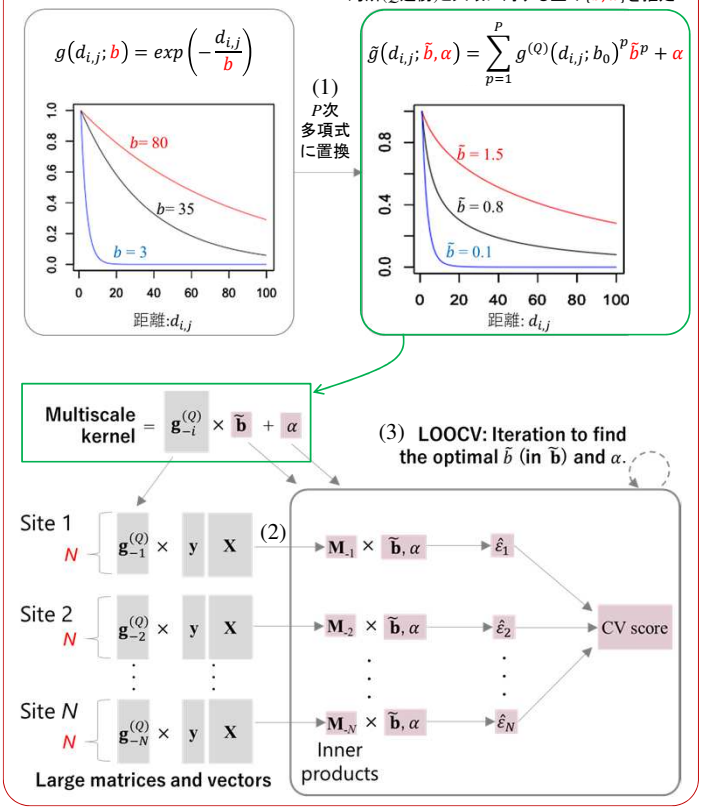


【提案手法】

以下の手順でLOOCVを高速化した(詳しくは**右図**):
(1) カーネル関数を線形式で近似
(2) パラメータを含まない行列を事前処理することで、サイズがNに依存する行列・ベクトルを消去
(3) LOOCVを実行

提案するLOOCVの手順

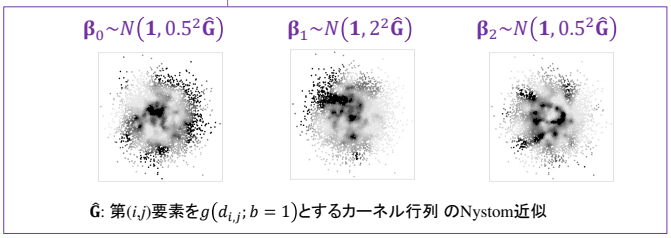
バンド幅bは予めb_0で固定。代わりに局所(Q近傍)と大域に対する重み$\{\tilde{b},\alpha\}$を推定



【シミュレーション実験】

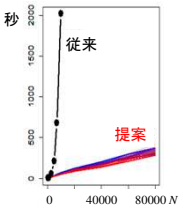
・$N\in\{500, 1000, 1500, 2,000, 3,000, 5,000, 10,000, 20,000, 80,000\}$
・下記から生成した擬似データに繰り返しあてはめることで
(各ケース200回反復)、従来型GWRと提案型GWRを比較

$$\mathbf{y} = \boldsymbol{\beta}_0 + \mathbf{x}_1 \circ \boldsymbol{\beta}_1 + \mathbf{x}_2 \circ \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}) \qquad \mathbf{x}_k \sim N(\mathbf{0}, \mathbf{I})$$



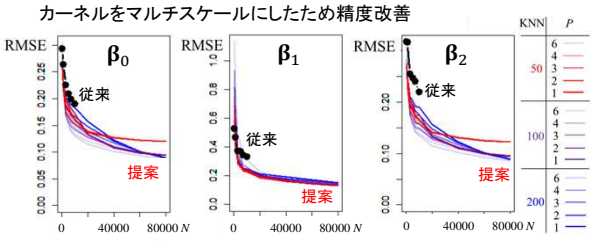
結果: 計算時間

高速化したことを確認



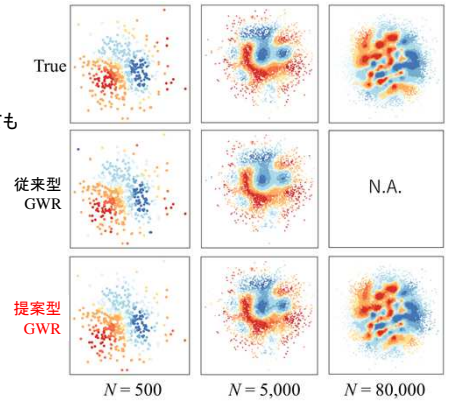
結果: SVC推定精度

カーネルをマルチスケールにしたため精度改善



推定されたSVCの例

標本が大きくなっても縮退せずに精度良くSVCが推定されたことを確認

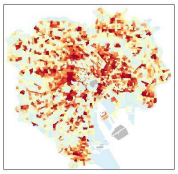


【実データへの応用】

東京23区内の非侵入窃盗件数の要因分析に応用

非侵入窃盗件数
(2012; 町丁目別)

すり, 自動車盗, 万引き, 置き引き,...

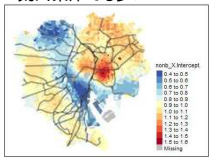


説明変数
- 先月の犯罪件数
- 昼間人口密度(国勢調査)
- 夜間人口密度(国勢調査)

出典:大東京防犯ネットワーク
(https://www.bouhan.metro.tokyo.lg.jp/)

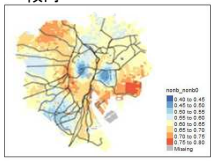
場所毎の定数項

- 都心・下町で多い
- 荒川東岸でも多い



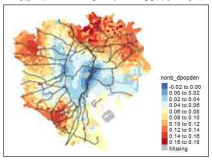
先月の犯罪件数の影響

- 山手線沿線で繰り返される傾向



昼間人口密度の影響

- 郊外:人口密集地で発生
- 都心:人が少ない場所で発生



夜間人口密度の影響

- 沿岸部の人が少ない場所は夜危険