Fused-MCP に基づく点事象集積領域検出手法

井上 亮・木元拓志

Point-Event Cluster Detection Based on Fused-MCP

Ryo INOUE and Hiroshi KIMOTO

Abstract: This study proposes a new point event cluster detection method based on the fused-MCP (Jing et al., 2018). The fused-MCP penalizes parameters themselves and the differences between adjacent parameters by the minimax concave penalty (MCP) (Zhang, 2010), and is proved that the estimators have the oracle property, which consists of variable selection consistency and estimation consistency. The parameter estimation is formulated as the maximization of log-likelihood and applied the MM algorithm. The application to the simulated data confirmed that the proposed method is better in estimating actual parameters than the methods by Wang and Rodríguez (2014) and Choi et al. (2018), which utilize the fused lasso. When the data is generated by setting the same parameters for adjacent regions, the proposed method is able to estimate them as the same value, indicating that it is useful for detecting multiple adjacent regions where point distribution is the same.

Keywords: 点事象（point event），集積検出（cluster detection），Fused-MCP

1. はじめに

近年，ICT や測位技術の普及や政府によるオープンデータの推進に伴い，詳細な位置を記録した多種の地理空間データが流通し，それらを活用した地域分析が可能になりつつある．本研究は，地理空間データの 1 形態である点事象データに着目し，点事象の集積領域検出手法について検討する．

点事象の集積領域検出に関する研究は数多く行われ，尤度比基準の空間スキャン統計 (例えば, Kulldorff & Nagarwalla, 1995) や False Discovery Rate (FDR) 制御法に基づく手法 (例えば，Castro & Singer, 2006) が提案されてきた．しかし，空間スキャン統計は，多重検定問題回避のため，同時に複数の集積領域を検出できない．また，FDR 制御法に基づく手法は，複数の集積領域の検出が可能だが，同等の集積性を持つ隣接地域を集約した検出はできない．

一方，近年，既存手法の課題を解決できる方法として，機械学習手法の Generalized fused lasso (Tibshirani & Taylor, 2011)に基づく手法 (Wang & Rodríguez, 2014; Choi et al., 2018)が提案されている．Generalized fused lasso は，係数自身と隣接係数の差に関する L_1 正則化を導入する方法で，係数と隣接係数の差が 0 に推定されやすい．空間分析において隣接地域の係数の差を正則化すると，共通の係数を取る地域の抽出ができる．Wang & Rodríguez (2014)は，分析対象を最小分析空間単位に領域分割し，各領域内の点事象数を，全領域共通の定数項と領域毎の集積性係数で表すポアソン回帰モデルを設定し，その対数尤度関数の最大化問題を，各領域の係数と隣接地域の係数の差に L_1 正則化を導入する推定方法を提案した．更に，Choi et al. (2018)は，共変量に対して全域で共通の

井上 亮
東北大学 大学院情報科学研究科
rinoue@tohoku.ac.jp

係数を導入したモデルに拡張した．

しかし, lasso 推定量は 0 に近づく方向のバイアスを有することが知られている (Fan & Li, 2001). Generalized fused lasso を用いた分析では，係数の絶対値を小さく推定すると共に，隣接係数の差も小さく推定するバイアスを持つため，それに基づく集積領域検出手法は，集積性係数の推定や集積領域の検出に影響を与える可能性がある．

MCP (Zhang, 2010) は，lasso の欠点を改善した正則化関数の一つである．MCP は原点から離れるにつれて傾きが 0 に近づく関数で，係数値がある閾値以上の領域では，罰則の大きさが一定であるため，バイアスが生じない．

MCP は信号の変化点抽出手法として，隣接パラメータ間に MCP の正則化項を加えた Fused-MCP (Jing et al., 2018) に拡張されている．この MCP を集積領域検出手法に応用できれば，バイアスのない集積性係数の推定に基づく集積領域検出を実行できる可能性がある．

本研究では，Fused-MCP に基づく，点事象集積領域検出手法を提案する．点事象にポアソン点過程を仮定した上で，集積領域検出問題を，Fused-MCP の正則化項付ポアソン対数尤度関数最大化による二次元平面における変化点抽出問題として定式化する．また，集積性パラメータと共変量パラメータの 2 種類のパラメータを推定する方法として，ポアソン回帰モデルと MCP 関数を近似し MM アルゴリズム (Hunter & Li, 2005)で計算する方法を構築する．提案手法の性能を，点事象集積のシミュレーションデータの分析を通して評価し，有用性を確認する．

2. Fused-MCP に基づく点事象集積領域検出手法

本研究は，点事象にポアソン点過程を仮定した上で， Fused-MCP の罰則項付ポアソン対数尤度関数最大化による二次元平面における変化点抽出問題として集積領域検出問題を定式化する．Generalized fused lasso に基づく集積領域検出手法(Choi et al., 2018)と Fused-MCP を用いた変化点抽出手法 (Jing et al., 2018)で用いられた MM アルゴリズム(Hunter & Li, 2005)を活用し，集積性パラメータと共変量パラメータの 2 種類を推定する計算方法を提案する．

2.1 目的関数

地域i (= 1, …, n)の点事象数 y_i と，全域で共通の共変量 $\mathbf{x}_i = (1, x_{i1},\ldots, x_{in})^T$ の関係を共変量係数ベクトル $\boldsymbol{\beta} = (\beta_0, \beta_1,\ldots,\beta_p)^T$ (β_0は定数項)で表し，それで説明できない地域iに固有の影響を，集積性係数 α_i で表すモデルを考える．人口や面積など地域の大きさを表すオフセット項をe_iと記すと，ポアソン回帰モデルは式(1)で表される．

$$\ln E(y_i) = \ln e_i + \alpha_i + \mathbf{x}_i^T\boldsymbol{\beta} \tag{1}$$

本モデルで$\alpha_i > 0$となることは，地域 i が他地域より点事象数の多い領域であることを表す．

Fused-MCPに基づく推定を行う本研究は，各地域の集積性係数，隣接地域の集積性係数の差，全域共通の共変量係数にMCP

$$\rho(t;\lambda,\gamma) = \lambda\int_0^t\left(1-\frac{x}{\gamma\lambda}\right)_+ dx \tag{2}$$

による正則化を導入した，対数尤度最大化で係数を推定する．$\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)^T$，$C$を隣接地域対の集合，$\lambda_1, \lambda_2, \lambda_3, \gamma_1, \gamma_2, \gamma_3$ をハイパーパラメータとすると，本研究のパラメータ推定は式(3)の最適化問題として定式化できる．

$$\begin{aligned}\min_{\boldsymbol{\alpha},\boldsymbol{\beta}}\Big[&-\ell(\boldsymbol{\alpha},\boldsymbol{\beta}) + \textstyle\sum_{i=1}^n \rho(|\alpha_i|;\lambda_1,\gamma_1)\\ &+\textstyle\sum_{(k,l)\in C}\rho(|\alpha_k-\alpha_l|;\lambda_2,\gamma_2) + \sum_{j=1}^p\rho(|\beta_j|;\lambda_3,\gamma_3)\Big]\end{aligned} \tag{3}$$

ただし，ポアソン対数尤度関数 $\ell(\boldsymbol{\alpha},\boldsymbol{\beta})$ は定数部を省略した式(4)である．

$$\ell(\boldsymbol{\alpha},\boldsymbol{\beta}) = \textstyle\sum_{i=1}^n\left[y_i\left(\ln e_i + \alpha_i + \mathbf{x}_i^T\boldsymbol{\beta}\right) - \exp\left(\ln e_i + \alpha_i + \mathbf{x}_i^T\boldsymbol{\beta}\right)\right] \tag{4}$$

2.2 計算方法

本研究では，既往研究を参考に，式(3)を二次関数に近似し，MMアルゴリズムで解く．その際，Choi et al. (2018)と同様に，$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$を交互に固定した計算を繰り返す．

(1) αの更新

βのm回目の繰り返し計算による近似解$\boldsymbol{\beta}^{(m}$が与えられた下では，式(3)第一項は

$$-\ell\left(\boldsymbol{\alpha},\boldsymbol{\beta}^{(m)}\right)=\sum_{i=1}^{n}\left[-y_i\left(\ln e_i+\alpha_i+\mathbf{x}_i^T\boldsymbol{\beta}^{(m)}\right)+\exp\left(\ln e_i+\alpha_i+\mathbf{x}_i^T\boldsymbol{\beta}^{(m)}\right)\right] \quad (5)$$

である．ここで，$\boldsymbol{\mu}_\alpha^{(m)}=\left[\exp\left(\ln e_i+\alpha_i^{(m)}+\mathbf{x}_i^T\boldsymbol{\beta}^{(m)}\right)\right]_i$，$\mathbf{y}=(y_1,\ldots,y_n)^T$ と記すと，勾配とヘッセ行列は

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\left(-\ell\left(\boldsymbol{\alpha},\boldsymbol{\beta}^{(m)}\right)\right)\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(m)}}\equiv-\left(\mathbf{y}-\boldsymbol{\mu}_\alpha^{(m)}\right) \quad (6)$$

$$\mathbf{H}_\alpha^{(m)}=\frac{\partial^2\left(-\ell\left(\boldsymbol{\alpha},\boldsymbol{\beta}^{(m+1)}\right)\right)}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\Bigg|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(m)}}=\operatorname{diag}\left(\boldsymbol{\mu}_\alpha^{(m)}\right) \quad (7)$$

と書ける．従って，式(3)第一項の$\boldsymbol{\alpha}^{(m)}$近傍の2次テイラー近似は，

$$-\left(\mathbf{y}-\boldsymbol{\mu}_\alpha^{(m)}\right)^T\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^{(m)}\right)+\frac{1}{2}\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^{(m)}\right)^T\mathbf{H}_\alpha^{(m)}\left(\boldsymbol{\alpha}-\boldsymbol{\alpha}^{(m)}\right) \quad (8)$$

となる．$\mathbf{H}_\alpha^{(m)}$の最大固有値を$\sigma_\alpha^{(m)}$とすると，

$$\mathbf{z}_\alpha^{(m)}=\boldsymbol{\alpha}^{(m)}+\left(\mathbf{y}-\boldsymbol{\mu}_\alpha^{(m)}\right)\Big/\sigma_\alpha^{(m)} \quad (9)$$

に対して以下の不等号が成立し，式(8)の近似の上界が得られる．

$$-\ell\left(\boldsymbol{\alpha},\boldsymbol{\beta}^{(m)}\right)\le\sigma_\alpha^{(m)}\left\|\mathbf{z}_\alpha^{(m)}-\boldsymbol{\alpha}\right\|_2^2\Big/2 \quad (10)$$

一方，MCPによる正則化項は，Jing et al. (2018)に従って式(11)の二次関数で近似する．

$$g_\varepsilon\left(t\,|\,t^{(m)}\right)=\rho\left(t^{(m)^2}\right)+\left(t^2-t^{(m)^2}\right)\rho'\left(\left|t^{(m)}\right|\right)\Big/2\left(\left|t^{(m)}\right|+\varepsilon\right) \quad (11)$$

ε は分母が0となるのを避ける微小な値である．

式(8)(11)を用いて，**α**の更新は次式で行う．

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\left[\frac{1}{2}\left\|\mathbf{z}_\alpha^{(m)}-\alpha\right\|_2^2+\frac{1}{\sigma_\alpha^{(m)}}\sum_{i=1}^{n}g_{\varepsilon_1}\left(\alpha_i\left|\alpha_i^{(m)};\lambda_1,\gamma_1\right.\right)+\frac{1}{\sigma_\alpha^{(m)}}\sum_{(k,l)\in C}g_{\varepsilon_2}\left(\alpha_k-\alpha_l\left|\alpha_k^{(m)}-\alpha_l^{(m)};\lambda_2,\gamma_2\right.\right)\right] \quad (12)$$

ただし，$\varepsilon_1=\dfrac{10^{-8}}{4\lambda_1}\times\min\left\{\left|\alpha_i^{(1)}\right|:\alpha_i^{(1)}\neq0\right\}$

$$\varepsilon_2=\frac{10^{-8}}{8\lambda_2}\times\min\left\{\left|\alpha_k^{(1)}-\alpha_l^{(1)}\right|:\alpha_k^{(1)}-\alpha_l^{(1)}\neq0\right\}$$

(2) βの更新

αのm+1回目の繰り返し計算による近似解$\boldsymbol{\alpha}^{(m+1)}$が与えられると，式(3)の第一項は式(13)となる．

$$-\ell\left(\boldsymbol{\alpha}^{(m+1)},\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\left[-y_i\left(\ln e_i+\alpha_i^{(m+1)}+\mathbf{x}_i^T\boldsymbol{\beta}\right)+\exp\left(\ln e_i+\alpha_i^{(m+1)}+\mathbf{x}_i^T\boldsymbol{\beta}\right)\right] \quad (13)$$

$\boldsymbol{\mu}_\beta^{(m)}=\left[\exp\left(\ln e_i+\alpha_i^{(m+1)}+\mathbf{x}_i^T\boldsymbol{\beta}^{(m)}\right)\right]_i$，　$\mathbf{X}=(x_1,\ldots,x_n)^T$

と記すと，勾配とヘッセ行列は

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(-\ell\left(\boldsymbol{\alpha}^{(m+1)},\boldsymbol{\beta}\right)\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}=-\sum_{i=1}^{n}\mathbf{x}_i\left[y_i-\exp\left(\ln e_i+\alpha_i^{(m+1)}+\mathbf{x}_i^T\boldsymbol{\beta}^{(m)}\right)\right]\equiv-\mathbf{X}^T\left(\mathbf{y}-\boldsymbol{\mu}_\beta^{(m)}\right) \quad (14)$$

$$\mathbf{H}_\beta^{(m)}=\frac{\partial^2\left(-\ell\left(\boldsymbol{\alpha}^{(m+1)},\boldsymbol{\beta}\right)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}\equiv\mathbf{X}^T\operatorname{diag}\left(\boldsymbol{\mu}_\beta^{(m)}\right)\mathbf{X}^T \quad (15)$$

となる．従って式(3)第一項の，$\boldsymbol{\beta}^{(m)}$近傍の二次テイラー近似は

$$-\left(\mathbf{y}-\boldsymbol{\mu}_\beta^{(m)}\right)^T\mathbf{X}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{(m)}\right)+\frac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{(m)}\right)^T\mathbf{H}_\beta^{(m)}\left(\boldsymbol{\beta}-\boldsymbol{\beta}^{(m)}\right) \quad (16)$$

と表せる．$\mathbf{H}_\beta^{(m)}$の最大固有値$\sigma_\beta^{(m)}$とすると，$\mathbf{z}_\beta^{(m)}=\boldsymbol{\beta}^{(m)}+\mathbf{X}^T\left(\mathbf{y}-\boldsymbol{\mu}_\beta^{(m)}\right)\Big/\sigma_\beta^{(m)}$ に対して以下の不等号が成立し，式(16)の近似の上界が得られる．

$$-\ell\left(\boldsymbol{\alpha}^{(m)},\boldsymbol{\beta}\right)\le\sigma_\beta^{(m)}\left\|\mathbf{z}_\beta^{(m)}-\boldsymbol{\beta}\right\|_2^2\Big/2 \quad (17)$$

正則化項の近似と合わせて，**β**は次式で更新する．

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{1+p}}\left[\frac{1}{2}\left\|\mathbf{z}_\beta^{(m)}-\boldsymbol{\beta}\right\|_2^2+\frac{1}{\sigma_\beta^{(m)}}\sum_{j=1}^{p}g_{\varepsilon_3}\left(\beta_j\left|\beta_j^{(m)};\lambda_3,\gamma_3\right.\right)\right] \quad (18)$$

ただし，$\varepsilon_3=\dfrac{10^{-8}}{4\lambda_3}\times\min\left\{\left|\beta_j^{(1)}\right|:\beta_j^{(1)}\neq0\right\}$ である．

以上の繰り返し計算をパラメータが収束するまで繰り返し，式(3)の解を求める．

3. 性能評価

シミュレーションデータを用いて，提案手法の性能評価を行った．シミュレーションデータは，17×17の格子領域に集積領域を設定し，集積領域内外にそれぞれ平均点数を与えて作成した．ま

た，初期パラメータはリッジ回帰で与えた．

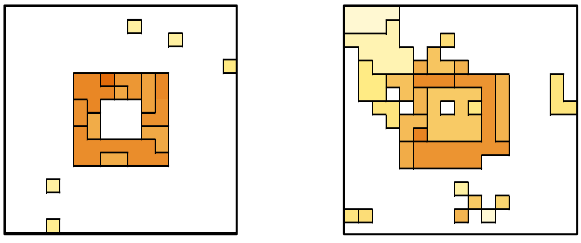
ここでは，対象領域中央の7×7の集積領域の内側に3×3の非集積領域がある，ドーナツ状の集積領域を設定した例を**図-2**に示す．ハイパーパラメータγ_2を調整し，その他は$\lambda_1 = 1,\ \lambda_2 = 5,\ \lambda_3 = 0.1,\ \gamma_1 = \gamma_3 = 1.0$に固定した．集積領域内外のメッシュ当たりの点数をそれぞれ25，5と設定した．

$\gamma_2 = 1$では，集積領域の周辺も非集積領域と推定されたが，$\gamma_2 = 10$では，集積領域近傍の非集積領域の集積性係数を大きく，非集積領域に隣接した集積領域の集積性係数を小さく推定し，集積領域の境界が不明瞭な結果が得られた．γ_2が大きいと，lassoに近い正則化となることが知られている．この結果は，MCPによる正則化によって，lassoに基づく手法よりもより適切な推定を得ることができる可能性を示している．

4. おわりに

本研究では，点事象分布にポアソン点過程を仮定した上で，Fused MCPによる罰則項付ポアソン対数尤度関数最大化による二次元平面における変化点抽出問題として点事象集積領域検出を定式化し，その推定計算手法を構築した．

シミュレーションデータに対する適用を通して，提案手法の性能を評価し，lassoを用いた既往手法よりも検出能力が高いことを示唆する結果を得た．現時点では初期の検討に留まっているおり，今後，適切なハイパーパラメータ設定を行うための判断基準となる，推定結果の評価基準について検討を進める．



(a)$\gamma_2 = 0.1$　　(b)$\gamma_2 = 10$

図-1　γ_2を変化させた場合の集積領域

謝辞

本研究はJSPS科研費18H01552の助成を受けた．

参考文献

Castro, M. C. & Singer, B. H. 2006. Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis*, **38**: 180–208.

Choi, H., Song, E., Hwang, S.S., & Lee, W. 2018. A modified generalized lasso algorithm to detect local spatial clusters for count data. *Advances in Statistical Analysis*, **102**(4):537–563.

Fan, J. & Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **95**:1348–1360.

Hunter, D. R. & Li, R. 2005. Variable selection using MM algorithms. *The Annals of statistics*, **33**(4):1617–1642.

Jing, B., Yang, G., Yu, X., & Zhang, C. 2018. Fused-MCP with Application to Signal Processing. *Journal of Computational and Graphical Statistics*, **27**(4):872–886.

Kulldorff, M. & Nagarwalla, N. 1995. Spatial disease clusters: Detection and inference. *Statistics in Medicine*, **15**: 707-715.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. 2005. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B*, **67**(1):91–108.

Tibshirani, R.J. & Taylor, J. 2011. The solution path of the generalized lasso. *The Annals of statistics*, **39**(3):1335–1371.

Wang, H. & Rodríguez, A. 2014. Identifying pediatric cancer clusters in Florida using log-linear models and generalized lasso penalties. *Statistics and Public policy*, **1**(1):86–96.

Zhang, C. H. 2010. Nearly unbiased variable selection under minimax concave penalty. The *Annals of statistics*, **38**(2):894–942.