

Fused-MCP に基づく点事象集積領域検出における ハイパーパラメータ設定方法

竹本一至*・井上亮**

Selection of Hyperparameters of Point Event Cluster Detection Based on Fused-MCP

Kazushi Takemoto*, Ryo Inoue**

The method of detecting point event cluster based on L_1 regularization penalizing regional coefficient and the difference between adjacent regional coefficients is proposed. However, the method is known to output biased result. Therefore, a method using MCP function as regularizer is proposed to improve this bias, and confirmed it detects cluster with high accuracy depending on the hyperparameter setting. This study proposes the hyperparameter setting of point event cluster detection based on fused-MCP using information criteria. The most suited information criterion is confirmed by the experiments using simulation data. The experiments compare the result of detection using four types of information criteria in terms of power, false discovery rate, and false positive rate, and confirmed hyperparameter setting based on EBIC is appropriate.

Keywords: 点事象集積領域検出 (Point Event Cluster Detection), ポアソン回帰 (Poisson Regression), Fused-MCP, ハイパーパラメータ (Hyper-parameter), 情報量規準 (Information Criteria)

1. はじめに

点事象データは、落雷や犯罪発生、感染症発病、商業施設出退店など、時空間で起こる事象について、その発生位置や時点を記録したデータである。この点事象データから、自然現象や社会経済現象の地域的特徴を把握する地域分析が検討されてきた。

点事象データに基づく代表的な分析の1つは、点事象の集積領域検出である。尤度比に基づいて集積領域を評価する空間スキャン統計（例えば、Kulldorff and Nagarwalla, 1995）や False Discovery Rate (FDR) 制御法の活用 (Castro and 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ず、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という課題を有する。

一方、これらの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) を活用した点事象集積領域検出手法 (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 正則化を導入して推定する方法で、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各領域固有の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。Choi et al. (2018) は、Wang & Rodríguez (2014) のモデルに全域で共通の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

* 学生会員 東北大学工学部建築・社会環境工学科 (Tohoku University)

〒980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-06 E-mail : kazushi.takemoto.q7@dc.tohoku.ac.jp

** 正会員 東北大学大学院情報科学研究科 (Tohoku University)

出に影響を与える可能性がある。

Lasso の欠点を改善できる正則化関数の一つとして、Zhang (2010) は MCP を提案した。MCP は原点から離れるにつれ傾きが 0 に近づく関数で、係数値がある閾値以上では、罰則の大きさが一定となり、バイアスが生じない。Jing et al. (2018) は、隣接係数間に MCP の正則化項を加えた fused-MCP を提案し、その応用例の 1 つとして、二次元格子状に与えられたノイズ入りのデータから、各領域に固有の係数を推定してノイズを除去する問題への適用を示した。

井上・木元 (2019) は、Choi et al. (2018) のモデルの L_1 正則化項を、MCP 関数に置き換えた手法を提案した。シミュレーションデータを用いた実験を通して、正則化項の重みや MCP 関数の形状を定めるハイパーパラメータの設定次第で、検出力が高く、誤検出率が低い分析が可能だが、ハイパーパラメータ設定を定める方法は確立されていない。

そこで、情報量規準によるモデル選択を通して、fused-MCP に基づく集積領域検出手法のハイパーパラメータ設定を提案し、そのハイパーパラメータ設定による検出手法の性能評価を行うことを本研究の目的とする。本研究で検討する情報量規準は、代表的な情報量規準である AIC (Akaike, 1974), BIC (Schwarz, 1978) と、ポアソン回帰モデルへの適用時に AIC が有するバイアスを修正した Bias-Corrected AIC (CAIC) (Kamo et al., 2013), および、モデル候補数が非常に大きい場合に生じるバイアスを修正する Extended BIC (EBIC) (Chen et al., 2008) の利用を検討する。性能評価は、シミュレーションデータを用いた実験で行う。

2. Fused-MCP による点事象集積領域検出

本章は、井上・木元 (2019) の fused-MCP による点事象集積領域検出手法のモデルを示す。

分析対象領域 i の点事象数を $y_i (i=1, \dots, n)$ 、面積や人口など地域 i の大きさを表すオフセット項を e_i 、共変量ベクトルを $\mathbf{x}_i$ 、地域固有の係数を α_i 、共変量係数ベクトルを $\boldsymbol{\beta}$ とする。Fused lasso や fused-MCP に基づく点事象集積領域検出では、式(1)のポアソン回帰モデルに対して、地域係数に対する正則化項と、隣

接する地域係数の差に対する正則化項、定数項以外の共変量の係数に対する正則化項を導入した推定を行い、その推定値が 0 以外となる地域を集積として判断する。

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

式(1)の対数尤度関数を $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y})$ 、MCP 関数を

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

と記すと、井上・木元 (2019) の係数推定は、式(3)の罰則項付ポアソン対数尤度関数の最大化問題で表される。

$$\min_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^{1+p}}} \left[- \sum_{i=1}^n \ell(\boldsymbol{\alpha}, \boldsymbol{\beta}; y_i) + \sum_{i=1}^n \rho(|\alpha_i|; \lambda_1, \gamma_1) + \sum_{i \sim i', i < i'} \rho(|\alpha_i - \alpha_{i'}|; \lambda_2, \gamma_2) + \sum_{j=1}^p \rho(|\beta_j|; \lambda_3, \gamma_3) \right] \quad (3)$$

ただし、 $i \sim i'$ は地域 i と地域 i' の隣接を表し、 $\lambda_1, \lambda_2, \lambda_3, \gamma_1, \gamma_2, \gamma_3$ はハイパーパラメータを表す。

MCP 関数を用いた式(3)の係数推定を行うと、generalized fused lasso に基づく推定よりもバイアスが小さな結果が得られることが期待できる。なお、MCP 関数のハイパーパラメータ γ を $+\infty$ すると lasso の L_1 正則化に一致するが、井上・木元 (2019) のシミュレーションデータを用いた分析を通して、 γ を小さくすることにより、データ生成に用いた係数に近い結果が推定されることを確認している。

6 種類のハイパーパラメータ設定を調整すると、検出力が高く誤検出率の低い推定結果が得られることが確認されているが、真の係数が未知の状態において、ハイパーパラメータ設定をどのように定めるべきかはこれまでに議論されていない。

3. 情報量規準

情報量規準は統計モデルを予測性能に基づき評価する基準である。本研究では、情報量規準に基づくハイパーパラメータ設定の選定を提案するが、以下に利用を検討する 4 つの情報量規準を示す。

3.1. AIC

AIC (Akaike, 1974) は、 $l(\theta)$ はパラメータ θ を変数とするモデルの最大対数尤度、 p はパラメータ数とすると式(4)で定義される。

$$AIC = -2l(\theta) + 2p \quad (4)$$

AICは、データ数が小さい場合、パラメータ数がデータ数を上回る場合、推定されるモデルのなかに真のモデルが含まれない場合などに偏りが生じる。

3.2. BIC

BIC (Schwarz, 1978) は、ベイズ推定におけるモデルの事後確率に基づくモデルの評価基準で、 n をデータ数とすると、式(5)で表わされる。

$$BIC = -2l(\theta) + p \log n \quad (5)$$

3.3. Bias-Corrected AIC

Bias-Corrected AIC (Kamo et al., 2013) は、ポアソン回帰モデルの変数選択問題でデータ数が小さいときに AIC に生じる偏りを修正した基準で、観測されるデータ y_i ($i=1, \dots, n$) が、パラメータ λ_i のポアソン分布に従うポアソン回帰モデルのとき式(6)で表される。

$$CAIC = -2l(\theta) + 2p + \frac{1}{m} (\mathbf{1}_n)^T D_{\hat{\lambda}_p} \{ (\widehat{W}_{(3)} + D_{\widehat{w}} \widehat{W} D_{\widehat{w}}) D_{\hat{\lambda}_p} - D_{\widehat{w}}^2 \} \mathbf{1}_n \quad (6)$$

ただし、 $\mathbf{1}_n$ は全要素が 1 のベクトル、 $\rho = (\rho_1, \dots, \rho_n)^T = (m_1/m, \dots, m_n/m)^T$ 、 $\widehat{W} = [\widehat{w}_{ij}] = X(X^T D_{\hat{\lambda}_p} X)^{-1} X^T$ 、 $\widehat{w} = (\widehat{w}_{11}, \dots, \widehat{w}_{nn})^T$ 、 $\widehat{W}_{(3)} = [\widehat{w}_{ij}^3]$ 、 $D_{\hat{\lambda}_p} = D_{\hat{\lambda}} D_{\rho} = \text{diag}(\hat{\lambda}_1 \rho_1, \dots, \hat{\lambda}_n \rho_n)$ を表す。

3.4. Extended BIC

Extended BIC はモデルの共変量の候補の数が大きいときに BIC で生じる偏りを修正する基準で、 $\tau(S_j) = P$ C_j を P を地域パラメータの総数、 j を 0 以外の値で推定された地域パラメータ数として、 P と j の組み合わせの数とすると式(7)で表される。

$$EBIC = -2l(\theta) + p \log n + 2\gamma \log \tau(S_j), (0 \leq \gamma \leq 1) \quad (7)$$

4. シミュレーションデータによる実験

情報量基準に基づくハイパーパラメータ設定による集積領域検出性能を評価するため、シミュレーションデータを用いた実験を行う。

4.1. シミュレーションデータの生成

17×17 の格子領域を対象領域とし、中心に集積領

域を設定する。設定集積領域の大きさ、集積領域外平均点数、集積領域内外の平均点数の比を表す点密度比を設定して、パラメータを平均点数とするポアソン分布に従うデータを 40 回生成する。

なお本実験では、共変量は設定していない。定数項以外に全域で共通の共変量係数はないため、式(3)の第 4 項、共変量係数の正則化項は存在しないモデルを用いて推定する。

4.2. ハイパーパラメータ設定

ハイパーパラメータを $\lambda_1 = \{2^0, 2^1, \dots, 2^7\}$ 、 $\lambda_2 = \{2^0, 2^1, \dots, 2^5\}$ 、 $\gamma_1 = \gamma_2 = \{2^0, 2^2, 2^4, 2^6, 2^7, 2^8, 2^9, 2^{10}\}$ と設定し、情報量基準ごとに最良の結果を選択する。

4.3. 評価項目

検出力、誤検出率、偽陽性率や設定集積領域で検出される結合領域の数で評価する。

$$\text{検出力} = \frac{\text{検出された設定集積領域数}}{\text{設定集積領域数}}$$

$$\text{誤検出率} = \frac{\text{検出された設定集積領域外の領域数}}{\text{検出領域数}}$$

$$\text{偽陽性率} = \frac{\text{検出された設定集積領域外の領域数}}{\text{設定集積領域外の領域数}}$$

4.4. 実験

(1) 5×5 の集積領域設定

実験データは、設定集積領域内外で点数の差が明確なデータと不明確なデータを利用するため、点密度比を 1.5, 1.75, 2.0 の 3 通り、集積領域外平均点数を 10, 15, 20, 40, 60, 80 の 6 通りで設定し生成した。

集積領域 5×5 の 1 つの実験データでの検出領域を図-1 に示し、40 回の実験の集計結果を表-1 で示す。なお、AIC と CAIC は、全実験で同じハイパーパラメータ設定が選ばれたため合わせて表示する。

検出力は集積領域外平均点数と点密度比が小さくなると小さくなり、誤検出率と偽陽性率は情報量基準毎に同じような結果になる。

AIC と CAIC は、検出力は大きいと同時に誤検出率も大きくなる設定が選ばれ、偽陽性率も大きく、設定基準として適さないことを確認した。

BIC と EBIC の検出力は、集積領域内外の点事象数の差が明確になるデータで大きくなり、明確でないデータでは検出されない。誤検出率と偽陽性率は、

AICに比べBICとEBICでは低く抑えられ、EBICで最も小さく抑えられている。また、EBICとBICで隣接関係にある複数の集積領域をどれだけ、結合して検出するか比較すると、EBICのほうが、領域を結合し少ない領域数を検出する傾向にある。

(2) 集積領域数を変えた実験(EBIC)

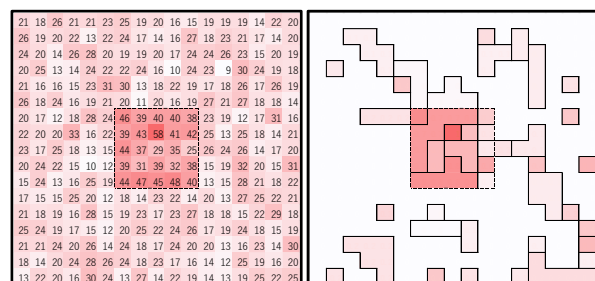
点密度比と集積領域外平均点数の設定を変えずに、設定集積領域数9(3×3)、25(5×5)、49(7×7)で実験を行う。各領域数のEBICの結果をまとめて表-2で示す。

集積領域数が大きくなると、検出力が大きくなる反面、誤検出率は大きくなる傾向にある。精度の高い検出が可能なのは、集積領域数が小さく、設定集積領域内外の点数差が明確になるデータである。

(3) まとめ

EBICは、第3項の0以外に推定されるパラメータが少いデータを選ぶ働きにより、保守的であるが誤検出を低く抑える結果を選択しやすいと考える。

また、EBICにおいて、集積領域数が小さいデータでは検出手法の精度が高くなる。設定集積領域と隣接する領域数の変化が影響していると考えられる。



(a) 分析データ

(b) AIC

(c) BIC

(d) EBIC

図-1 集積領域 5×5、集積領域外平均点数 20 点、集積領域内平均点数 40 点の分析例

表-1 集積領域 5×5、40 個のデータの結果

(a) 検出力

検出力 集積領域外 平均点数	点密度比								
	AIC・CAIC			BIC			EBIC		
	1.5	1.75	2	1.5	1.75	2	1.5	1.75	2
10	0.583	0.757	0.909	0.008	0.198	0.570	0.002	0.021	0.294
15	0.711	0.854	0.940	0.037	0.435	0.775	0.000	0.093	0.566
20	0.748	0.914	0.967	0.087	0.595	0.850	0.009	0.274	0.725
40	0.925	0.984	0.998	0.547	0.900	0.975	0.208	0.868	0.960
60	0.975	1.000	1.000	0.793	0.976	0.998	0.690	0.959	0.990
80	0.984	0.999	1.000	0.884	0.990	0.999	0.767	0.979	0.998

(b) 誤検出率

誤検出率 集積領域外 平均点数	点密度比								
	AIC・CAIC			BIC			EBIC		
	1.5	1.75	2	1.5	1.75	2	1.5	1.75	2
10	0.758	0.723	0.684	0.273	0.227	0.181	0.000	0.250	0.145
15	0.658	0.651	0.653	0.431	0.294	0.238	0.000	0.238	0.264
20	0.675	0.641	0.640	0.336	0.330	0.316	0.000	0.237	0.302
40	0.649	0.657	0.653	0.211	0.277	0.229	0.137	0.129	0.119
60	0.690	0.621	0.651	0.224	0.242	0.246	0.217	0.232	0.190
80	0.606	0.627	0.622	0.293	0.286	0.355	0.258	0.278	0.270

(c) 偽陽性率

偽陽性率 集積領域外 平均点数	点密度比								
	AIC・CAIC			BIC			EBIC		
	1.5	1.75	2	1.5	1.75	2	1.5	1.75	2
10	0.173	0.187	0.186	0.000	0.005	0.012	0.000	0.001	0.005
15	0.129	0.151	0.168	0.003	0.017	0.023	0.000	0.003	0.019
20	0.147	0.154	0.163	0.004	0.028	0.037	0.000	0.008	0.030
40	0.162	0.178	0.178	0.014	0.033	0.027	0.003	0.012	0.012
60	0.205	0.155	0.177	0.022	0.030	0.031	0.018	0.027	0.022
80	0.143	0.159	0.156	0.035	0.038	0.052	0.025	0.036	0.035

(d) 結合領域の数

集積領域内結合領域数			
点密度比	集積領域外 平均点数	BIC	EBIC
1.5	10	0.125	0.05
	15	0.475	0
	20	1.125	0.225
	40	6.35	1.7
	60	8.125	5.975
	80	8.55	5.825
1.75	10	3.125	0.175
	15	4.85	0.75
	20	5.1	2.125
	40	9.975	9.875
	60	12.025	10.075
	80	10.475	8.775
2	10	6.85	2.525
	15	9.55	5.3
	20	9.15	6.15
	40	13.375	11.975
	60	10.5	9.525
	80	8.95	7.85

表-2 設定集積領域を変えた EBIC での結果

(a) 検出力

検出力 集積領域外 平均点数	点密度比								
	集積領域数:9			集積領域数:25			集積領域数:49		
	1.5	1.75	2	1.5	1.75	2	1.5	1.75	2
10	0.000	0.000	0.067	0.002	0.021	0.294	0.001	0.014	0.276
15	0.000	0.036	0.344	0.000	0.093	0.566	0.001	0.134	0.756
20	0.003	0.119	0.606	0.009	0.274	0.725	0.001	0.395	0.829
40	0.103	0.747	0.989	0.208	0.868	0.960	0.411	0.907	0.973
60	0.211	0.844	0.983	0.690	0.959	0.990	0.836	0.965	0.994
80	0.508	0.942	1.000	0.767	0.979	0.998	0.915	0.990	0.999

(b) 偽陽性率

偽陽性率 集積領域外 平均点数	点密度比								
	集積領域数:9			集積領域数:25			集積領域数:49		
	1.5	1.75	2	1.5	1.75	2	1.5	1.75	2
10	0.000	0.000	0.000	0.000	0.001	0.005	0.000	0.026	0.035
15	0.000	0.000	0.002	0.000	0.003	0.019	0.008	0.030	0.052
20	0.000	0.001	0.003	0.000	0.008	0.030	0.027	0.050	0.076
40	0.002	0.009	0.009	0.003	0.012	0.012	0.062	0.084	0.063
60	0.002	0.010	0.011	0.018	0.027	0.022	0.059	0.057	0.049
80	0.002	0.007	0.007	0.025	0.036	0.035	0.031	0.055	0.029

5. おわりに

本研究では Fused-MCP に基づく点事象集積領域検出のハイパーパラメータ設定に情報量規準を用いることを提案し、シミュレーションデータを用いた実験を通して、4つの情報量規準における最適なハイパーパラメータ設定を検出手法に採用して4つのモデルを生成し、検出手法の性能評価を行った。

結果から、AIC は、検出力は大きい偽陽性率も大きい設定を選ぶ傾向が強く、選定基準として適さないことを確認した。一方で、EBIC は保守的な検出を行うが、誤検出率・偽陽性率が低く、かつ、隣接領域を統合した検出をする設定が選ばれる傾向にあることを確認した。

謝辞

本研究は JSPS 科研費 18H01552 の助成を受けた。

参考文献

- Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6): 716–723.
- 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.

Chen, J., Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**(3): 759–771.

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

Fan, J., and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456): 1348–1360.

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.

Kamo, K., Yanagihara, H., & Satoh, K. (2013) Bias-Corrected AIC for Selecting Variables in Poisson Regression Models. *Communication in Statistics-Theory and Methods*, **42**: 1911–1921

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

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**(2): 461–464.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. *Journal of Royal Statistical Society B*, **58**(1): 267–288.

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.

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.

井上 亮, 木元 拓志 (2019) Fused-MCP に基づく点
事象集積領域検出手法. 地理情報システム学会 学
術研究発表大会講演集, 28: E-5-4. CD-ROM.