2024年度共同利用研究報告書

2024年08月16日

所属・職名 鹿児島大学大学院理工学研究科・准教授
吉田 拓真

	整理番号	2024a004

1.研究計画題目	大規模クラスターデータに対する極値統計モデリングの開発			
2.新規・継続	新規			
3.種別	一般研究			
4.種目	短期研究員			
5.開催方法	対面開催			
6.研究代表者	氏名	吉田 拓真		
	所属 部局名	鹿児島大学大学院理工学研究科	職名	准教授
7.研究実施期間	2024年05月13日(月曜日)～2024年05月17日(金曜日)			
8.キーワード	極値統計学, スパースモデリング, 大規模クラスターデータ			
9.参加者人数	2人			

10.本研究で得られた成果の概要

本研究は大規模クラスターデータにおけるリスク予測のための極値統計モデルの構築がテーマである.クラスターデータとは地域などの属性情報が付与されたデータで, 例えば日本全土の約1300の雨量の観測地点があり, 各観測所で降雨量がデータとして蓄積されている.この場合, 観測地点がクラスターに相当する. 豪雨災害など降雨量が多い場合, その定量的な災害リスク評価は防災行動の基準を与えるために必要なデータ科学である. このようなクラスターデータに対して極値統計モデリングによるリスクを予測する際にクラスターの関連を考慮したモデリングを行いたい.

本研究は各クラスターの極値モデルに含まれる形状パラメータをfused lasso由来のスパース法で推定し, いくつかのクラスターの分布が統合され同一視できることに着目したクラスター統合解析法を提案するものである.
本研究について, 現状は以下の成果が得られた:
(1) 極値従属性が高いクラスター同士にスパース法を適用し解釈性の高い統合が実現できた
(2) fused lassoの派生であるAdaptive fused lassoを利用するとより実情に合った統合が可能となり, 予測性能も向上した
(3) 簡便で高速なアルゴリズムを構成できた
(4) クラスター数が大きい大規模な気象データに有用であることが確認できた

今後はスパース項に含まれる調整パラメータの効率的な決定, 統計理論の構築, 提案モデリングのアルゴリズムの一般公開を目指し, 継続して研究を行う.

本研究で得られた成果を2024年度統計関連学会連合大会で報告予定である. そこで得られたコメント等を元に研究を改善し, 論文にまとめる予定である.

研究課題「大規模クラスターデータに対する極値統計モデリングの開発」 報告書

1 研究の目的

本研究の目的はクラスターデータに対する新しい極値統計モデリングを開発することである. クラスターデータとは地域や所属など, 特定の集団に分類されるデータを指す. 極値統計学は最大値など大きな値が生起する確率を予測するための方法である. 例えば, 各アメダス降雨量観測地点の降雨量や各株銘柄の株価の突発的な値をとるリスク評価を極値統計モデルで表現することになる. このとき, 降雨量データでは近い地域同士が, 株価データでは関連株同士が互いに影響を与えあうと考えるのが自然であろう. 本研究ではクラスター間の関連を考慮した極値統計モデリングの開発, 特に大規模クラスターを考えたときのクラスターの統合 (グループ化) 方法を確立する.

本報告書では, 1128 箇所のアメダス観測地点において, 2000 年 1 月 1 日〜2022 年 12 月 31 日までに得られた日降水量のデータの極値モデリングを例に, 得られた成果について報告する.

2 研究の方法

本研究は極値モデリングとして, 広く扱われている一般化パレートモデルを想定する. そして, モデルに含まれるパラメータを fused lasso (Tibshirani et al. 2005) 由来のスパース法で推定し, クラスター統合方法として確立する.

いま, $Z_{i,j}$, $i = 1, \ldots, N_j$, $j = 1, \ldots, J$ を j 番目のクラスターの i 番目のデータとする. ここで, J はクラスター数で, N_j は j 番目のクラスターで得られたデータ数である. ここでは簡単のため, 各 j について, $Z_{1,j}, \ldots, Z_{N_j,j}$ は独立であるとする. 各クラスター j について, 閾値 w_j を超過する部分の $Z_{i,j} - w_j$ は一般化パレート分布

$$P(Z_{ij} - w_j > y | Z_{i,j} > w_j) = \left(1 + \frac{\gamma_j}{\sigma_j} y\right)^{-1/\gamma_j}, \quad j = 1, \ldots, J$$

で表現される. ここで, γ_j は形状パラメータ, σ_j は尺度パラメータである. さて, 各 j で $Y_{ij} = Z_{ij} - w_j$ のうち 0 より大きいものを改めて $Y_{1,j}, \ldots, Y_{n_j,j}$ と書く. ここで, n_j は j 番目の地域のデータで閾値 w_j より大きいデータ数である.

ここではクラスター統合を目的とし, パラメータ $\{(\gamma_j, \sigma_j), j = 1, \ldots, J\}$ を

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \log f(Y_{i,j} | \gamma_j, \sigma_j) - \lambda \sum_{k<j} c_{k,j} |\gamma_k - \gamma_j| \tag{1}$$

の最大化によって推定する. ここで,

$$f(y|\gamma, \sigma) = \frac{1}{\sigma} \left(1 + \frac{\gamma y}{\sigma}\right)^{-1/\gamma - 1}$$

は一般化パレート分布の密度関数, $\lambda \geq 0$ は調整パラメータ, $c_{k,j} \in \{0, 1\}$ はどのペアにペナルティを付与するかを表す既知のパラメータである. 前半の項が対数尤度でデータへの適合を測るための基準, 後半の項がクラスター統合のためのペナルティ項である. このようなペナルティ項は Tibshirani et al. (2005), Ke et al. (2015) などで扱われているが極値モデルに応用した前例はない. 本研究では, $c_{k,j}$ の取り方, つまり, どのクラスター同士に統合化のスパース化するかを工夫する.

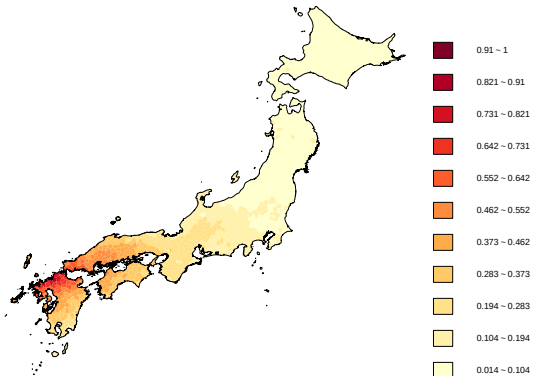

Figure 1: 福岡市とその他の観測所の降雨量の極値従属係数値.

どのクラスター同士が統合化できそうかを考えた時, 地理情報を含むデータであれば距離が近いクラスター同士を統合するのが自然である. そのような情報を含む特徴量としては例えば相関係数が挙げられる. しかし, 我々はいま大きな値である極値データのみに関連があるため, Coles et al. (1999) が提案した極値従属係数 (極値に相当するデータのみで計算した相関係数のような従属関係を表す指標) に着目し, この値が高いクラスター同士にペナルティを付ける. 例えば, Figure 1 は福岡県福岡市と全国各地の極値従属係数の値である. 図より, 福岡市の近隣エリアは従属性が高く, 離れるほど低くなっている. 大雑把に言うと, 福岡市に対しては, 図で赤色のエリアにペナルティ項を付け, クラスター統合を検討する. 逆に白いエリアはペナルティを付与せず, クラスター統合は最初から検討しない. ただし,fused lasso の特徴である “連鎖” で直接ペナルティがついてないクラスター同士も結果的に統合される可能性はある.

(1) で求まった $\gamma_1, \ldots, \gamma_J$ の推定値はいくつかのペアが同値となる. その際の最適化のポイントが調整パラメータ λ の値である. この λ がペナルティ項の強さを表現しており, すべてのペナルティで共通の強さに設定されている. もし事前情報から, クラスター同士の従属性も高く, かつ, パラメータの類似性も高いことがわかっているならば, その項のペナルティをより強く設定, つまり, Adaptive lasso ベースのペナルティに変更することができる. 幸い, 今回対象とするデータでは, 各クラスター毎の (γ_j, σ_j) の最尤推定量が求まる. いま, γ_j の最尤推定量を $\tilde{\gamma}_j$ と書き, (1) を

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} \log f(Y_{i,j}|\gamma_j, \sigma_j) - \lambda \sum_{k<j} \frac{c_{k,j}}{|\tilde{\gamma}_k - \tilde{\gamma}_j|}|\gamma_k - \gamma_j| \tag{2}$$

と変更するとこれは Adaptive fused lasso 由来のペナルティとなる.

本研究の主成果として, Ke et al. (2015) が提案した Homogeneity-Pursuit ペナルティを先行研究手法, 我々が新たに提案する方法を Dependence lasso (1), Dependence-Adaptive lasso (2) と比較し, 提案手法の性能を報告する.

Figure 2 には, 3 つの手法で推定された形状パラメータの分布を示している. 先行手法である Homogeneity Pursuit は中国地方はクラスター統合が顕著なものの, 東北地方はあまり統合されていない. また, この方法は近い地域同士を統合しているわけではないので, 例えば九州の地点と北海道の地点が統合されている場合もある. しかし, そのような統合は実学上の信憑性は高いとはいえない. 右図は, Dependence-lasso で推定された分布である. 同じ地方の特徴をかなり統合しているのが見て取れる. 非常に簡便な解釈が可能に見えるが, ペナルティの強さをすべての地域で同一としているため, 無理やり統合されている地域もあるように思える. この点についてはさらなる解

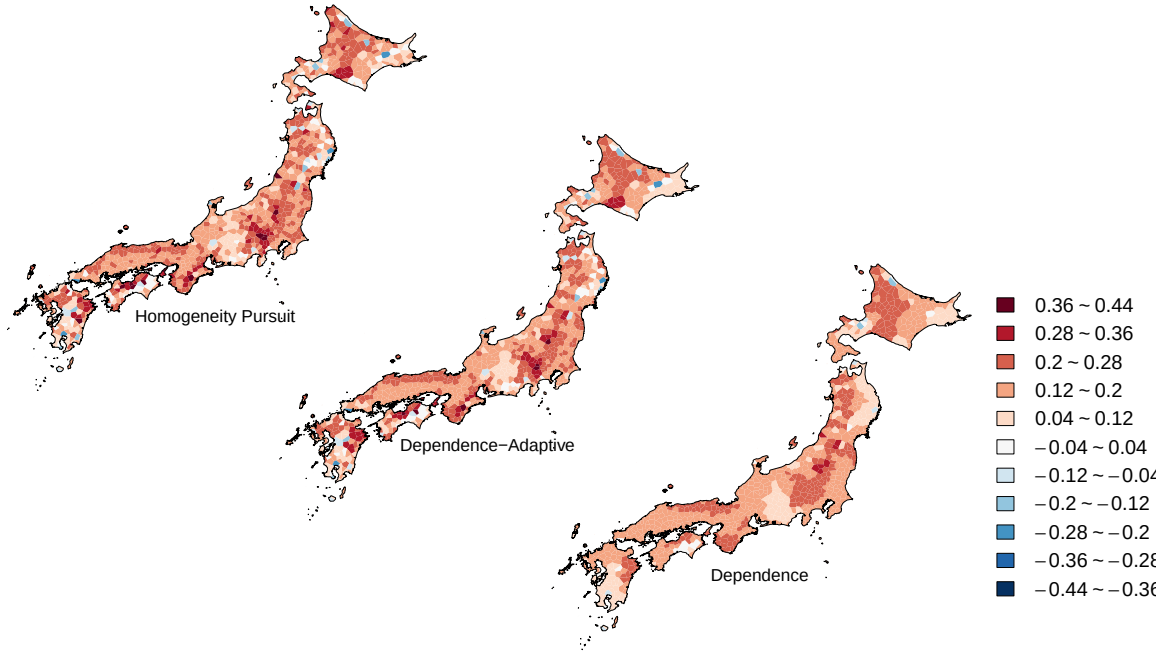

Figure 2: クラスター統合手法による形状パラメータ $\gamma_1, \ldots, \gamma_J$ の推定値．左: 先行手法 (Homogeneity Pusuit), 中央: 提案手法 (Dependence-Adaptive lasso), 右: 提案手法 (Dependence lasso)

析が必要である．最後に, 中央の図は Dependence-Adaptive lasso であり, その挙動はちょうど左右の結果の中間のような結果を返している．つまり, 近隣同士のクラスターを統合しつつ, パラメータの値がそこまで近くないところで別グループと判定している．このような微調整を施したような結果はペナルティ項の強さがそれぞれのペアで異なり, 強いところは引き付けあい, 遠いところは引き離す適度なバランスを持っていることによると考えられる．

3　まとめ

今回得られた結果により, 提案する手法はかねがねうまくいっているように思える．今後は Dependence-Adaptive lasso を中心により深く手法を理解し, 精度向上につなげたい．提案手法は今後パッケージ化し, GitHub で公開予定である．また, 今回の研究は 2024 年度統計関連学会連合大会で報告予定である．豪雨災害とその防災に詳しい土木学会の関係者らにも意見を伺い, 実用の際の微修正も行う予定である．その後も様々な国内外の会議で報告し, 得られたコメントを元に研究を深化させたい．

References

Coles,S., Heffernan,J. and Tawn,J. (1999). Dependence Measures for Extreme Value Analyses. *Extremes*. **2** 339–365.

Ke, Z.T., Fan, J. and Wu, Y. (2015). Homogeneity persuit. *Journal of American Statistical Association*. **110** 175–194.

Tibshirani S, Saunders M, Rosset S, Zhu J, Knight K. (2005). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society, Ser B*. **67** 91–108.

大規模クラスターデータに対する極値統計モデリングの開発

整理番号	2024a004
種別	一般研究-短期研究員
研究計画題目	大規模クラスターデータに対する極値統計モデリングの開発
研究代表者	吉田 拓真（鹿児島大学大学院理工学研究科・准教授）
研究実施期間	2024年5月13日（月）～ 2024年5月17日（金）
研究分野のキーワード	極値統計学, スパースモデリング, 大規模クラスターデータ
目的と期待される成果	本研究の目的はクラスターデータに対する新しい極値統計手法を開発することである。クラスターデータなど、特定の集団に分類されるデータを指す。極値統計学は最大値など、大きな値が生起する確率を予測法である。例えば、日本の降水量の観測所は約1300箇所あるが、各観測所が各クラスターとなり、観測所災害リスク予測のための極値モデリングを行うことになる。このとき、各観測所で得られたデータは近い響を受ける、もしくは関連があると考えるのが自然であろう。本研究ではクラスター間の関連を考慮したングの開発、特に大規模クラスターを考えたときのクラスターの統合(グループ化)方法を確立する。 具体的な方法としてスパース法を利用する。通常、各クラスターについて極値統計モデルは最尤法で構成lassoタイプのスパース法を導入することで、推定量の同一性が得られ、それがグループ化に繋がる。クラが高い箇所に優先的にスパース法を適用することで効率的な推定を行う。 本手法のメリットは統合後のグループ数と各グループに属するクラスターが自動で決定されるところにおいて、既存手法と一線を画することができる。また、グループ化されたクラスターにおけるデータは共通のれるため、解釈性が向上し、また、パラメータの高性能の推定が期待できる。本研究は地域別の豪雨災害頭に置いている。本研究でグループ化された地域は統一的な災害対策を打てるなど、リスクに応じた効立てることが可能となる。提案した手法はGitHubで公開し、関連研究の発展に貢献したい。
組織委員(研究集会) **参加者(短期共同利用)**	吉田 拓真（鹿児島大学 大学院理工学研究科・准教授） 川野 秀一（九州大学 大学院数理学研究院 ・教授）