方向データにおけるロバストなベイズ推定

中川 智之 [1]　　鶴田 靖人 [2]　　文 翔 [3]　　田畑 耕治 [3]

[1] 明星大学データサイエンス学環/理化学研究所
[2] 明治大学経営学部
[3] 東京理科大学創域理工学部

導入

$p-1$ 次元単位球面 $\mathcal{S}_p = \{\boldsymbol{x} \in \mathbb{R}^p \mid \|\boldsymbol{x}\| = 1\}$ 上の観測 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ は, さまざまな科学分野で現れる. ここで $\|\cdot\|$ はユークリッドノルムである. 例えば, 風向, 動物の方向性, 山火事の延焼方向などは $\mathcal{S}_2$ 上に分布する [5]. $p = 3$ の典型例としては, 岩石の古地磁気の方向や地球から星への方向が挙げられる [5]. テキストデータやゲノム配列の表現は, 高次元単位球面上の観測とみなされる [2]. 球面データの解析のために, 文献においていくつかの確率分布が提案されてきた. その中でも中心的な役割を果たしてきたモデルは, Langevin 分布とも呼ばれる von Mises-Fisher 分布である. 方向パラメータを $\boldsymbol{\mu}$, 集中パラメータを κ とすると von Mises-Fisher 分布の密度関数は次で与えられる.

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \kappa) = \frac{\kappa^{(p-2)/2}}{(2\pi)^{p/2} I_{(p-2)/2}(\kappa)} \exp\{\kappa \boldsymbol{\mu}^\top \boldsymbol{x}\}, \quad \boldsymbol{x} \in \mathcal{S}_p$$

ここで, $I_{(p-2)/2}(\cdot)$ は第一種の修正ベッセル関数である. 本研究では, von Mises-Fisher 分布における位置または集中パラメータのロバスト推定に焦点を当てる. Agostinelli[1] は, 円周上のデータに対して, 両方のパラメータを同時に推定するための別の 2 つの方法を提案している. また, Kato and Eguchi[4] は von Mises-Fisher 分布に対し, density power-divergence および γ-divergence を用いた M-推定量を導入し, そのロバスト性を検討した. しかしながら, これらの手法はカーネル密度推定などが必要で計算コストが高く p が大きい場合の推定が不安定になる. さらに, 信頼区間など不確実性の評価が難しい.

そこで本研究では, ベイズ統計の枠組みを用いて, 外れ値に対してロバストな事後分布を, density power-divergence および γ-divergence を用いて提案する. このアプローチは, ベイズ統計の枠組みを用いることで, 特に標本サイズが小さい場合にも推定の不確実性を容易に評価できる利点がある. さらに, 本研究では事後平均を推定するための重み付きベイズブートストラップ法を用いた近似的な事後分布からのサンプリングアルゴリズムを提示する.

方向データにおけるロバストなベイズ推定

方向データにおいて, Kato and Eguchi[4] は von Mises-Fisher 分布に対し, density power-divergence および γ-divergence に基づいた推定量を導入し, そのロバスト性を検討した. 一方で, ベイズ的枠組みにおいて, Ghosh and Basu[3] および Nakagawa and Hashimoto[6] は, それぞれ density power-divergence および γ-divergence に基づく一般化事後分布 $\pi^{(\alpha)}(\boldsymbol{\xi} \mid \boldsymbol{x}_{1:n})$

および $\pi^{(\gamma)}(\boldsymbol{\xi} \mid \boldsymbol{x}_{1:n})$ を提案している．ただし, $\boldsymbol{x}_{1:n} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ は観測データである．von Mises-Fisher 分布の場合, これらは次のように表される．

$$\pi^{(\alpha)}(\boldsymbol{\xi} \mid \boldsymbol{x}_{1:n}) = \frac{\pi(\boldsymbol{\xi}) \exp\left(-n d_\alpha(\bar{g}, f_{\boldsymbol{\xi}})\right)}{\int \pi(\boldsymbol{\xi}) \exp\left(-n d_\alpha(\bar{g}, f_{\boldsymbol{\xi}})\right) d\boldsymbol{\xi}}, \quad \pi^{(\gamma)}(\boldsymbol{\xi} \mid \boldsymbol{x}_{1:n}) = \frac{\pi(\boldsymbol{\xi}) \exp\left(-n \tilde{d}_\gamma(\bar{g}, f_{\boldsymbol{\xi}})\right)}{\int \pi(\boldsymbol{\xi}) \exp\left(-n \tilde{d}_\gamma(\bar{g}, f_{\boldsymbol{\xi}})\right) d\boldsymbol{\xi}}$$

ここで, $\boldsymbol{\xi} = \kappa \boldsymbol{\mu}$ であり, $\alpha, \gamma > 0$ はロバスト性を制御するためのチューニングパラメータであり, $d_\alpha(\bar{g}, f_{\boldsymbol{\xi}}) = n^{-1} \sum_{i=1}^n \ell_\alpha(\boldsymbol{x}_i, \boldsymbol{\xi})$, および $\tilde{d}_\gamma(\bar{g}, f_{\boldsymbol{\xi}}) = -\gamma^{-1} \left\{\exp\left(-\gamma d_\gamma(\bar{g}, f_{\boldsymbol{\xi}})\right) - 1\right\} = n^{-1} \sum_{i=1}^n \ell_\gamma(\boldsymbol{x}_i, \boldsymbol{\xi})$ はそれぞれ次のように定義される．

$$\ell_\alpha(\boldsymbol{x}_i, \boldsymbol{\xi}) = -\frac{1}{\alpha} \frac{\exp\left(\alpha \boldsymbol{\xi}^\top \boldsymbol{x}_i\right)}{K_p(\boldsymbol{\xi})^\alpha} + \frac{1}{\alpha + 1} \frac{K_p((1+\alpha)\boldsymbol{\xi})}{K_p(\boldsymbol{\xi})^{\alpha+1}},$$
$$\ell_\gamma(\boldsymbol{x}_i, \boldsymbol{\xi}) = -\frac{1}{\gamma} \frac{\exp\left(\gamma \boldsymbol{\xi}^\top \boldsymbol{x}_i\right)}{K_p((1+\gamma)\boldsymbol{\xi})^{\gamma/(1+\gamma)}} + \frac{1}{\gamma}$$

ロバスト性

影響関数 (Influence Function; IF) は, ロバスト統計学において推定量のロバスト性を評価するために用いられる概念であり, データに対する微小な変化に対する感度を定量化するものである．すなわち, ある推定量 $T_n(G)$ が外れ値や小さな摂動にどの程度影響を受けるかを評価する尺度として次のように定義される．

$$\mathrm{IF}(\boldsymbol{y}, T_n, G) = \lim_{\varepsilon \to 0} \frac{T_n\big((1-\varepsilon)G + \varepsilon \Delta_{\boldsymbol{y}}\big) - T_n(G)}{\varepsilon}$$

ここで, ε は汚染比率, $\Delta_{\boldsymbol{y}}$ は点 $\boldsymbol{y} \in \mathcal{S}_p$ に集中する退化分布である．方向データの場合, データ空間がコンパクト集合上に存在するため, 影響関数は常に有界となる．このような場合には, スケールおよびパラメータ化の影響を排除するために, 標準化影響関数 (Standardized Influence Function; SIF) を考慮するのが有用である．このとき, SIF のノルムは次式で定義される．

$$\mathrm{SIF}(\boldsymbol{y}, T_n, G) = \sqrt{\mathrm{IF}(\boldsymbol{y}, T_n, G)^\top S(G)^{-1} \mathrm{IF}(\boldsymbol{y}, T_n, G)}$$

ここで, $S(G)$ は漸近分散共分散行列を表す．この指標は推定量の外れ値に対する感度を統一的なスケールで評価できるため, ロバストベイズ推定法の比較や性能評価において重要である．当日は数値実験の結果を含めて詳細を報告する．

参考文献

[1] C. Agostinelli. Robust estimation for circular data. *Computational Statistics & Data Analysis*, 51(12):5867–5875, 2007.

[2] A. Banerjee, I. S. Dhillon, J. Ghosh, and Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9):1345–1382, 2005.

[3] A. Ghosh and A. Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

[4] S. Kato and S. Eguchi. Robust estimation of location and concentration parameters for the von Mises–Fisher distribution. *Statistical Papers*, 57:205–234, 2016.

[5] K. V. Mardia, P. E. Jupp, and K. Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000.

[6] T. Nakagawa and S. Hashimoto. Robust Bayesian inference via γ-divergence. *Communications in Statistics-Theory and Methods*, 49(2):343–360, 2020.