統計数理研究所 75 周年記念事業（2019）

ガンマ・ダイバージェンスに基づいた ロバスト統計

Robust Statistics via Gamma-Divergence

数理・推論研究系　藤澤 洋徳（Hironori Fujisawa）

1. はじめに

外れ値が存在するとき推定値にはバイアスが生じる．たとえば標本平均を考えよう．データの一つが非常に大きな値を取るとき，標本平均は非常に大きな値となってしまう．このバイアス問題を克服する単純な手法は中央値である．外れ値が一つであれば，中央値は大きく変わらない．ただし，外れ値の割合が大きい時には，中央値は大きくずれる．実は，外れ値の割合が大きい場合にも，バイアスを十分に小さくできる手法，というのは，長らくきちんと議論されていなかった．この問題はガンマ・ダイバージェンスによって解決される．**ロバスト統計で 25 年以上未解決だった重要問題が解決されたと考えている (Fujisawa and Eguchi, 2008).**

2. 本研究の特徴

過去の研究との大きな違いは二つある．一つはバイアスの議論の仕方であり，もう一つは外れ値への意識の仕方である．

上述したバイアスと言うのは，厳密には，統計科学で通常使われているバイアスではない．厳密に言えば「潜在バイアス」と呼ばれるものである．この潜在バイアスを小さくすることはロバスト統計の最大の目的の一つである．しかしながら，潜在バイアスを直接に議論することが難しかったため，理論的には，影響関数のような代替指標を使って議論することが主流であった．**これまで使われてきた代替指標を使わずに，潜在バイアスを直接に議論した点が，本研究の独創的な点である.**

もう一つは潜在バイアスを議論するときの前提条件に対する意識の違いである．過去のロバスト統計では，前述した影響関数という，強力で便利な代替指標を使って議論することが主流であった．しかし，この道具を使うためには，外れ値の割合が小さいということを暗に想定している．影響関数を使いながら，外れ値の外れ値らしさが軽く登場してくる．**本研究では意識レベルを逆にしている．外れ値の外れ値らしさの方を重視する．外れ値の割合が小さいという仮定は一切おいていない．この点も本研究の独創的な点である.**

結果的に，外れ値の割合が大きくても潜在バイアスが小さいロバスト推定，が可能になった．それはガンマ・ダイバージェンス $D_\gamma(g, f)$ に基づく手法である：

3. ガンマ・ダイバージェンス

ガンマ・ダイバージェンス $D_\gamma(g, f)$ は以下で与えられる：

$$d_\gamma(g,f) = -\frac{1}{\gamma}\log\int g(x)f(x)^\gamma dx + \frac{1}{1+\gamma}\log\int f(x)^{1+\gamma}dx,$$
$$D_\gamma(g,f) = d_\gamma(g,f) - d_\gamma(g,g).$$

ここで，g と f は密度関数であり，$\gamma > 0$ はロバスト性を調整するパラメータである．実際のパラメータ推定は，g に基づく期待値を経験密度関数 $\bar{g}$ で置き換え，f をパラメトリックモデル f_θ に置き換えて，$d_\gamma(\bar{g}, f_\theta)$ の最小化で行う．

図 1 はガンマ・ダイバージェンスに対して成立する近似的なピタリアン構造である．詳細は省くが，外れ値が外れ値らしいという仮定を置いている．**このピタゴリアン構造は，ガンマ・ダイバージェンスに基づくロバスト推定が，なぜ上手く働くかの直感的な説明を与える．**データ発生分布 g とパラメトリック分布 $h = f_\theta$ のダイバージェンス $D_\gamma(g, h)$ を小さくするのが普通の方法である．ここで次の二点に注意する：(i) 直交性が近似的に成り立っている．(ii) g と f は固定されていて自由に動けるのはパラメトリック分布 h だけである．そうすると，$D_\gamma(g, h)$ を小さくしようとすると，$D_\gamma(f, h)$ が小さくなり，結果的に，パラメトリック分布 h はターゲット分布 f に近づく．これがガンマ・ダイバージェンスに基づいたロバスト推定が上手く働く理由である．

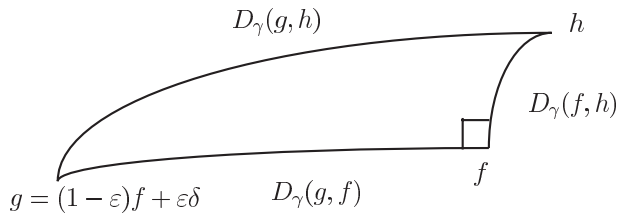

図 1. ピタゴリアン構造．g はデータ発生分布であり，ロバスト統計で慣習的に用いられている汚染分布 $g(x) = (1-\varepsilon)f(x) + \varepsilon\delta(x)$ を想定．ε は外れ値の割合．f はターゲット分布（外れ値に汚染されなかった場合の分布）．δ は外れ値の分布．h はパラメトリック分布 f_θ．

また，ある種の仮定の下では，外れ値の割合が大きくても潜在バイアスが小さくなる手法は，本質的に，ガンマ・ダイバージェンスに基づく手法だけであるという，ある種の唯一性が証明できる．加えて，パラメトリックモデルが指数型分布族に入っている場合には，ピタゴリアン構造を利用して，ロス関数の単調減少性をもつきれいなパラメータ推定アルゴリズムも提案できる．

4. 発展

唯一性の定理の仮定は弱く，ガンマ・ダイバージェンスはある種の決定打である．そのため，研究の発展は難しい状態が続いたが，最近になって研究が進むようになった．パラメトリック分布を，面積を変化させるパラメータ λ をわざと組み込んだ拡張モデル $\lambda f(x;\theta)$ に代えることで，ガンマ・ダイバージェンスでなくても，外れ値の割合が大きくても潜在バイアスが小さくなる手法を構築することができた (Kanamori and Fujisawa, 2015)．この研究と関連研究で，**日本統計学会研究業績賞**を頂いた．当初に提案したパラメータ推定アルゴリズムは，ピタゴリアン構造を利用していてきれいであったが，スパース罰則と組み合わせると，使いやすいパラメータ推定アルゴリズムの構築が容易でなく，ロバスト性とスパース性を同時に併せもつ手法の提案が難しかった．この問題は，Majorization-Minimization アルゴリズムの適用で克服されて，ガウシアン・グラフィカル・モデリングや回帰モデリングに適用されている (Hirose, Fujisawa and Sese, 2015; Kawashima and Fujisawa, 2017)．Google Scholar などで検索することで，その他にも様々に発展をして注目を浴びていることを見て取ることができる．