Robustness of Cramer-von Mises statistic under contiguous type contamination

(株) データサイエンスコンソーシアム, 慶應義塾大学　仲 真弓
(株) データサイエンスコンソーシアム, 慶應義塾大学, 早稲田大学　柴田 里程

1　Cramer-von Mises 統計量のロバスト性

データを解析して何かしらの知見を得ようとしたとき，得られたデータに対し，その背景を考慮したモデルを仮定し，その上で推定されたパラメータの違いを検証したり，そのモデルを用いて何かの予測をする，というアプローチがある．このような状況では，あくまで最後の「違いの検証」や「予測」が目標であり，そのためには仮定したモデルが妥当かどうかが判断できれば十分であることが多い．モデルの妥当性をチェックするためには，例えば Cramer-von Mises 統計量のような適合度を測る統計量を用いてチェックすることができる．

このとき，データ自体に少しの contamination（汚染）があっても適合度のチェックの際にあまり影響を与えない，というロバスト性は一つのメリットとして考えられる．この contamination が統計量に与える影響を明らかにするため，$X_1, X_2, \ldots, X_n$ が分布 $F(x, \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^\top$, の contaminate された分布

$$F^*(x, \boldsymbol{\theta}) = \left(1 - \frac{\varepsilon}{\sqrt{n}}\right) F(x, \boldsymbol{\theta}) + \frac{\varepsilon}{\sqrt{n}} G(x) \tag{1}$$

に従っているとき，統計量 $\sum_{j=1}^{n} \{F(X_{(j)}, \boldsymbol{\theta}) - \frac{j}{n+1}\}^2$ の漸近分布にどのような影響を与えるかを検証した．その結果，パラメータ推定を含まない場合，統計量は重み付き非心 χ_1^2 変数の和の分布に収束し，その重み λ_j, $j = 1, 2, \ldots$, は積分方程式

$$\lambda f(u) = \int_0^1 \{\min(u, v) - uv\} f(v) dv$$

の固有値で与えられ，非心度 μ_j^2, $j = 1, 2, \ldots$, は，

$$\mu_j = \varepsilon \lambda_j^{-\frac{1}{2}} \int_0^1 \int_0^1 f_j(u)(1_{u \geq v} - u) \left\{1 - \frac{g(F^{-1}(u, \boldsymbol{\theta}))}{f(F^{-1}(u, \boldsymbol{\theta}))}\right\} du dv$$

の二乗となることが明らかになり，contamination の影響は非心度のみにあらわれることが分かる．

2　検出力との関係

Cramer-von Mises 統計量を用いた検定において，帰無仮説「$X \sim F(x, \boldsymbol{\theta})$」に対して対立仮説「$X \sim F_\delta(x, \boldsymbol{\theta})$」を考え，その検出力を検証するため，統計量 $\sum_{j=1}^{n} \{F_\delta(X_{(j)}, \boldsymbol{\theta}) - \frac{j}{n+1}\}^2$ の漸近分布も既に知られている．ただし $F_\delta(x, \boldsymbol{\theta})$ は，その密度関数 $f_\delta(x, \boldsymbol{\theta})$ が

$$f_\delta(x, \boldsymbol{\theta}) = f(x, \boldsymbol{\theta}) \left\{1 + \frac{\delta}{\sqrt{n}} \eta(F(x, \boldsymbol{\theta}))\right\} \tag{2}$$

という形をとる，contiguous（$F(x, \boldsymbol{\theta})$ に隣接している）モデルについてであり，$\int_0^1 \eta^2(u) du = 1$ という条件のもとである．

結果として得られる漸近分布は似た形にはなるが，その近傍のとり方について，(1) と (2) では条件が異なっている．また，検出力においては，「モデルから少しでもずれていれば棄却してほしい」という立場であるが，実際に構築したモデルに基づいて何かを明らかにしたいときには，その先を進める前に「モデルが妥当かどうかチェックしたい」という立場であり，このときロバスト性は一つのメリットとなりえる．