位置と尺度の推定に基づく多変量外れ値検出法の比較

独立行政法人 統計センター　和田 かず美

はじめに

本報告では、統計調査の実務への適用を目的に、観測変量が楕円体分布族に従うことを前提とする多変量で破局点（breakdown point）の高い外れ値検出法の比較を行う。なお、p 次元確率変数 $\boldsymbol{x}$ が楕円体分布族に従うとき、その密度関数は次の $f(\boldsymbol{x})$ で表される。

$$f(\boldsymbol{x}) \propto g\left\{\left((\boldsymbol{x}-\boldsymbol{u})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{u})\right)\right\}$$

ここで、$\boldsymbol{u}$ は $\boldsymbol{x}$ の位置母数のベクトル、$\boldsymbol{V}$ は尺度母数の行列である。また g は非負値関数であり、確率変数の定義域全体での積分が有界となる必要がある。

比較手法

比較を行った外れ値検出法は、和田（2010）及び Wada and Tsubaki (2013) が R で実装した MSD(Modified Stahel-Donoho)法の改良版、Béguin and Hulliger (2003) で S-plus コードが公開されている BACON(Blocked Adaptive Computationally Efficient Outlier Nominator)法、及び Wang and Raftery (2002) のノンパラメトリックな NNVE（Nearest-Neighbor Variance Estimation）法で、それぞれ下表に示す実装・移植コードあるいは CRAN パッケージを使用する。

MSD 法	https://github.com/kazwd2008/MSD https://github.com/kazwd2008/MSD.parallel/
BACON 法の移植について	https://github.com/kazwd2008/BEM
NNVE 法	cov.nnve 関数 [covRobust パッケージ]

統計調査データは分布の歪みや長い裾を持つ場合も想定されるため、性能評価には、相関のある多変量の正規分布または楕円体分布族ではない skew-t 分布に従う乱数データに正規分布に従う外れ値を分散と比率を変えて添加した混合分布データを用いて、シミュレーションによりそれぞれの外れ値検出法の特徴を明らかにする。

参考文献

Béguin, C. and B. Hulliger, B.　(2003). Robust multivariate outlier detection and imputation with incomplete survey data., EUREDIT Deliverable, D4/5.2.1/2 Part C.

Wada, K. and Tsubaki, H. (2013). Parallel computation of modified Stahel-Donoho estimators for multivariate outlier detection. Proceedings of 2013 IEEE International Conference on Cloud Computing and Big Data (CloudCom-Asia), 304-311, 16 -19, Dec. 2013, Fuzhou, China.

Wang, N. and Raftery, A. E. (2002), Nearest-Neighbor Variance Estimation (NNVE): Robust covariance estimation via nearest-neighbor cleaning, Journal of the American Statistical Association. 97, (460), 994-1019.

和田 かず美 (2010) 多変量外れ値の検出 ～MSD 法とその改良手法について～, 統計研究彙報第 67 号, pp.89-157, 総務省統計研修所.