外れ値を検出する方法とその特性に関する研究

2017SS004 青木勇太
指導教員：松田眞一

1　はじめに

本研究では，外れ値を検出する手法に対してシミュレーションを行い，その外れ値検出手法の特性を理解し，性能を比較することを目的としている．

2　外れ値検出手法

本研究では R で利用可能な以下の手法を用いた．1 次元では，ホテリング理論，箱ひげ図，スミルノフ・グラブス検定．多次元では，ホテリング T^2 理論，One Class Support Vector Machine (SVM)，Local Outlier Factor (LOF) である．(Web[2]，高畠 [5] 参照)

なお，本研究では各手法の実現に R のパッケージを利用したが，ホテリング理論，多変量のホテリング T^2 理論は井出 [3] を，スミルノフ・グラブス検定は青木 [1] を参考にして作成した．

3　シミュレーションについて

本研究では，以下の 3 つのシミュレーションを行う．

1. 1 次元の手法の特性を見るシミュレーション
2. 1 次元の手法の性能比較のミュレーション
3. 多次元の手法の特性比較のシミュレーション

以降はこれらをシミュレーション 1 等と呼ぶ．

4　シミュレーション 1

1 次元の各手法の特性を調べるためにシミュレーションを行った．外れ値を含みやすいという理由から自由度 3 の t 分布をデータに用いて，各手法がどのように外れ値を検出するかを確かめた．サンプルサイズ 100 個，試行回数 1000 回で行った．

それぞれの手法で検出された外れ値の総数は，ホテリング理論が 2610 個，箱ひげ図が 5617 個，スミルノフ・グラブス検定が 2304 個であった．

箱ひげ図が検出した外れ値の総数がほかの 2 倍以上あることから箱ひげ図は，計算段階で改善する必要があると考えられる．

4.1　箱ひげ図の改善について

まず，検出された総数の 5617 個という値が理論値に従ったものなのかを確かめる．自由度 3 の t 分布における外れる確率は 5.6 %であった．本シミュレーションは値を 100000 個発生させているから，その中の 5.6 %は 5600 個であり 5617 個は理論値に従っていることがわかった．

よって，他の手法の総数が約 2500 個だとすると，信頼区間は 97.5 %にすればよい．その場合に四分位範囲にかける値は 2.2 であった．そうしてもう一度シミュレーションした結果，外れ値の総数は 2640 個となった．よって改善できた．

5　シミュレーション 2

本シミュレーションは，用いた 1 次元の手法の性能比較のために行った．外れる個数を二項分布によって決定して，二項分布の外れる確率を指定することにより行うシミュレーションである．外れ値を含まない方の山を平均 0 分散 1 の正規分布に従うように，外れ値を含む方の山を平均 5 分散 0.3 の正規分布に従うように乱数で発生させた．サンプルサイズ 100 個，試行回数 1000 回で行った．

なお，箱ひげ図は 4.1 節で改善したものを用いた．

以下表 1 が結果であり，外れ値を含む確率それぞれに対して外れ値を正確に検出した回数を表している．

表 1　各手法が外れ値の数を正確に検出した回数

	ホテリング	箱ひげ図	スミルノフ
1 %	625	941	954
3 %	854	913	740
5 %	823	857	389
10 %	226	598	24

表 1 から，スミルノフ・グラブス検定とホテリング理論は，悪いときの値がかなり悪いが，箱ひげ図は，全体的にいい結果だとわかる．スミルノフ・グラブス検定は確率が小さいときに他より性能がいいと言える．ただし，それは確率が低いときのみに限ったことであり，1 %から 10 %に向けて箱ひげ図の落ち幅が他より少なく，全体を通して言えば箱ひげ図のほうがいいと言える．

また，ホテリング理論だけが外れ値を含む確率が 1 %の時に一番良い結果にならないことについて，外れ値を含まない山の分布が原因なのではないかと考え，外れ値を含まない平均 0 分散 1 の正規分布のシミュレーションを行った．その結果が図 1 であり，ホテリング理論は外れ値を含まない場合に他より外れ値を検出しやすいことがわかった．したがって，ホテリング理論が外れ値を含む確率が 1 %のときに一番良い結果にならない原因は，理論ベースである正規分布に完全に従う場合に限界値の設定から 1 %の外れ値を検出しようとしてしまうことにある．すなわち外れ値ではないのにサンプルサイズが 100 なので図 1 で 1 個見つける場合が一番多くなっている．

6　シミュレーション 3

多次元の各手法の性能を比較するするために行うシミュレーションである．データは，外れ値を検出するのが難しいデータとして Peña and Prieto[4]，和田 [6] が使用した

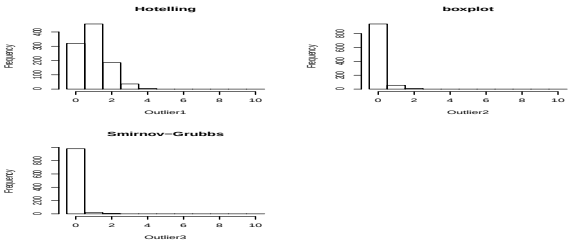

図 1　外れ値のない正規分布

多次元混合正規分布

$$(1-\alpha)N_M(\mathbf{0}, I) + \alpha N_M(\delta \boldsymbol{e}_1, \lambda I) \qquad (1)$$

を用いる．α が外れ値の割合，M が次元数，δ が原点からの距離，$\boldsymbol{e}_1 = \{1, 0, \ldots, 0\}$，$\lambda$ が外れ値の分散を表している．パラメータは和田 [6] を参考に $\delta = 10$，$\lambda = 0.01$ に定めた．

シミュレーションは，2 次元データで外れ値の個数を 5 個から 1 個まで変化させシミュレーションを行い，その後次元数を 5 次元まで変化させて行う．紙面上の都合上，2 次元で外れ値 5 個，2 次元で外れ値 1 個，4 次元で外れ値 5 個の結果のみを図 2，3，4 で示す．

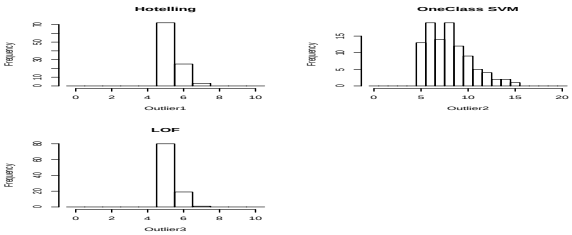

図 2　2 次元データで外れ値 5 個

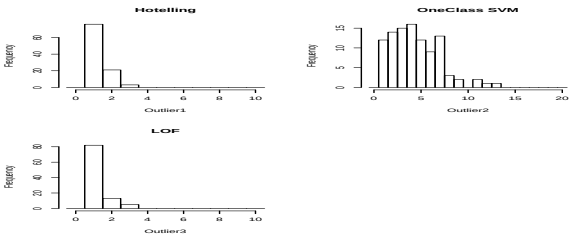

図 3　2 次元データで外れ値 1 個

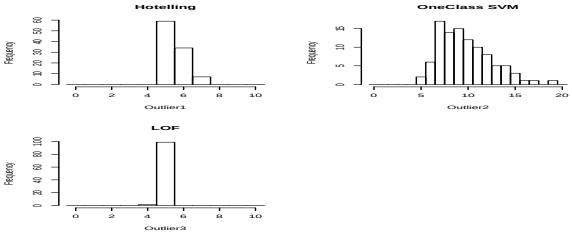

図 4　4 次元データで外れ値 5 個

One Class SVM は，与えた混合分布自身で検定したらうまくいかなかったので，外れ値を含まないデータを学習させて外れ値を含むデータの検定を行ったが，ほかの手法のような結果にはならなかった．

2 次元のシミュレーションでは，LOF が一番精度がよく，その次にホテリング理論というように見える．次に次元数を大きくしたときは，ホテリング理論は横に広がる結果となり，反対に LOF は，精度がよくなっていることがわかる．この結果からも次元が大きいと LOF の精度がよくなっていて，2 次元の時点でも精度としてが高いことから，LOF が一番いいと言える．

7　まとめ

1 次元データに対する手法に関しては，外れ値が含まる確率が低いときのスミルノフ・グラブス検定の精度の高さが目立ったが，実際に外れ値を検出する場面や，外れ値が含まれる確率が高いときのことを考えると改善後の箱ひげ図が一番性能がいいとわかった．多次元データに対する手法に関しては，次元数が大きくなっても結果がよかった LOF が一番いいとわかった．しかし，5 次元データで LOF が稀に全く外れ値を検出できていないことがあった．この理由は，データの間隔が広がりすぎてうまくいかないせいだと考えられる．サンプルサイズがもっと大きければ改善する可能性がある．

8　おわりに

本研究を通して，用いた外れ値検出手法の使用方法や，各手法の特性や性能について理解することができた．外れ値を検出した時，外れ値を除くことが必要なのかしっかりと判断したしながら，本研究で学んだことを活かしたい．

参考文献

[1] 青木繁伸：「スミルノフ・グラブス検定」．http://aoki2.si.gunma-u.ac.jp/lecture/Grubbs/Grubbs.html/，2015.（2020/6 閲覧）

[2] BellCurve：「4-3. 外れ値検出のある箱ひげ図」，https://bellcurve.jp/statistics/course/5222.html/，(2020/6 閲覧).

[3] 井出剛：「入門　機械学習による異常検知―R による実践ガイド」，コロナ社，2015.

[4] Peña,D. and F.J.Prieto：「Multivariate Outlier Detection andRobust Covariance Matrix Estimation」，*Technometrics*，Vol.43，pp.286-300，2001..

[5] 高畠泰斗：「密度推定法に基づくカーネル判別機械」，筑波大学大学院博士過程システム情報工学研究科修士論文 ，https://commons.sk.tsukuba.ac.jp/wp-content/uploads/sites/13/2016/08/200520847.pdf/，2007.（2020/11 閲覧）

[6] 和田かず美：「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.89-157，2010.