外れ値を検出する方法の特徴比較

青木 勇太 *　松田 眞一 *
E-Mail: matsu@nanzan-u.ac.jp

外れ値を検出する方法は現在，異常検知の観点などから幅広い分野で重要視されている．本論文では，1 次元のデータおよび多次元のデータに対する外れ値を検出する方法のうち R で取り扱いやすいものを取り上げて，その特性や性能を比較するためにシミュレーションを行った．1 次元データに対する外れ値検出方法として，ホテリング理論，箱ひげ図，スミルノフ・グラブス検定の 3 つを，多次元データに対する外れ値検出方法として，多変量のホテリング T^2 理論，One Class SVM，LOF の 3 つを比較した．

シミュレーションの結果，それぞれの方法の特性が分かり，特に箱ひげ図による外れ値検出方法は手順の改善を行うとかなり外れ値を検出する精度が向上することが分かった．性能に関しては，1 次元データに対する方法では改善した箱ひげ図による方法が 3 つの方法の中で一番良い性能を示し，多次元のデータに対する方法では LOF が 3 つの方法の中で一番良い性能を示すことが分かった．

1　はじめに

計測のミスや標本が無作為にとられるために偶然出現する外れ値や異常値は，様々な統計量に影響を与える．データ解析を行う際にその外れ値を適切に対処することが重要であるが，外れ値を検出する方法は数多く提案されているため，実際に外れ値を含むデータの解析を行う際にどれを用いるのが適切かという問題が生じる．

また，外れ値検出方法や異常検知方法は，車をはじめ機械の異常検知や人の健康異常の検知，また機械学習の学問分野など幅広く応用されている．このことも外れ値検知方法の特徴を知ることが重要であることを示している．

本論文では誰もが比較的簡単に使用することができる方法を念頭に，R を用いてシミュレーションを行い，その特徴比較と考察を行う．

2　外れ値検出方法

外れ値を検出する方法は，簡単に調べただけでも 40 ぐらいの提案が存在する．(Web[3][10] 参照) 本論文ではそのうち R で比較的簡単に用いることのできる以下の方法を扱う．1 次元データに対する方法は，ホテリング理論，箱ひげ図，スミルノフ・グラブス検定であり，多次元データに対する方法は，多変量のホテリング T^2 理論，One class SVM (Support Vector Machine)，LOF (Local Outlier Factor) である．

本章を通して 1 次元のデータは，データの大きさを N とし，データは $\{x_1, \ldots, x_N\}$ で表す．多次元のデータは，データの大きさを N，データを $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ とし，次元を M 次元として表す．

*南山大学理工学部システム数理学科

なお，本論文で用いた外れ値検出方法のうち，ホテリング理論，スミルノフ・グラブス検定，多変量のホテリング T^2 理論は，それに関する R のパッケージがなかったため下記に示す手順に従って R の関数を自作したが，他の 3 つに関しては R のパッケージを用いてシミュレーションを行った．

2.1 ホテリング理論

ホテリング理論とは，以下に示すような手順に基づいて各データに対する異常度を算出し閾値と比べ，その異常度が閾値を超えていたら，それを外れ値とする方法である．

1. 異常度を計算するために，標本平均 $\hat{\mu}$ と標本分散 $\hat{\sigma}^2$ を計算する．

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{1}$$

2. 観測値を x' とし，x' に対する異常度 $a(x')$ を下記のように計算する．

$$a(x') = \frac{(x' - \hat{\mu})^2}{\hat{\sigma}^2} \tag{2}$$

3. 閾値を求め，2 で求めた $a(x')$ と比較し，外れ値かどうかを判定する．なお，元のデータが正規分布のとき，N が大きければホテリング理論の異常度は自由度 1 のカイ二乗分布で近似できる．閾値を求める式は以下で表される．

$$\alpha = \int_{\chi_1^2(\alpha)}^{\infty} f\chi(x|1)dx = 1 - \int_{0}^{\chi_1^2(\alpha)} f\chi(x|1)dx \tag{3}$$

式 (3) の α は分布の上側確率値であり，α を与えることによって求められる $\chi_1^2(\alpha)$ が閾値を表す．一般的に α は，0.01 とする．なお，式 (3) の $f\chi(x|m)$ は自由度 m のカイ二乗分布の密度関数である．

(井出 [4] 参照)

2.2 箱ひげ図

箱ひげ図の書き方は複数存在するが，本論文では R のパッケージに合わせて次で説明する Tukey が提案した方法を用いることとする．

1. データから第 1 四分位点，第 3 四分位点と中央値を求める．
2. 第 1 四分位点から第 3 四分位点まで箱を描き，中央値に当たるところに線を入れる．
3. 第 3 四分位点から第 1 四分位点を引いた値である四分位範囲を計算する．
4. 3 で求めた四分位範囲を 1.5 倍し，その数値分第 3 四分位点に足した値と，第 1 四分位点から引いた値の 2 つの値を上下の閾値とする．

5. 閾値までに存在するデータの最も遠いところまで箱からひげを引く.

6. 各データと閾値を比較し，4 で求めた閾値の外にある値をすべて外れ値とし，ドットで個別に表す.

本論文での箱ひげ図による外れ値検出方法とは，この手順による方法を指す．一方，高校で学習するひげの端点を最大値，最小値とする方法は，Tukey の箱ひげ図の簡略版であり，外れ値を示すようにはなっていない．(白旗 [11]，BellCurve[2] 参照)

なお，本論文では R の grDevices パッケージ中の boxplot.stats 関数を利用した．(Web[9] 参照)

2.3　スミルノフ・グラブス検定

スミルノフ・グラブス検定は，データの中の最大値もしくは，最小値が外れ値であるかを検定するものである．すなわち，この方法は，データが正規分布に従うと仮定して，以下の帰無仮説，対立仮説の下で検定を行う.

- 帰無仮説 H_0 :「すべてのデータは，同じ正規母集団から抽出されたものである.」
- 対立仮説 H_1 :「データの中の最大値もしくは，最小値は外れ値である.」

本論文では，有意水準 $\alpha = 0.01$ とし，最大値または最小値に対して，片側検定で行った．統計量は，U を不偏分散としたとき，i 番目のデータが最も平均から離れているとして

$$T_i = \frac{|X_i - \overline{X}|}{\sqrt{U}} \tag{4}$$

で求められ，上側 $100\alpha\%$の近似的な有意点 t は，$t_{\alpha/N}$ を自由度 $N-2$ の t 分布の上側 $100\alpha/N\%$点としたとき

$$(N-1)\left(\frac{t^2_{\alpha/N}}{N(N-2)+Nt^2_{\alpha/N}}\right)^{\frac{1}{2}} \tag{5}$$

によって求める.

実際に外れ値であるかは，スミルノフ・グラブス検定の統計量である T_i と有意点 t を比較して判断する．$T_i < t$ の場合は，対立仮説が棄却せれず外れ値とはみなされない．反対に，$T_i \geq t$ となった場合に，対立仮説が棄却され X_i が外れ値とみなされる．そして，仮に外れ値とみなされた場合，その値をデータから除いてもう一度検定を行い，外れ値とみなされる値がなくなるまで再帰的に繰り返し行う．(青木 [1] 参照)

2.4　多変量のホテリング T^2 理論

先程 1 次元のホテリング理論について説明したが，多変量のホテリング T^2 理論は，1 次元のホテリング理論を多次元のデータに対応するよう拡張したものである．1 次元のホテリング理論で，データが正規分布に従うと仮定したようにこちらも，独立に同じ分布に従う

N 個の M 次元の観測値からなるデータが以下の確率密度の多次元正規分布に従うと仮定して行う．

$$N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{6}$$

用いた手順を以下に示す．

1. データから標本平均 $\hat{\boldsymbol{\mu}}$ と，標本分散共分散行列 $\hat{\boldsymbol{\Sigma}}$ を計算する．

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \tag{7}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^{\mathrm{T}} \tag{8}$$

2. 各観測値 $\boldsymbol{x}'$ に対して異常度 $a(\boldsymbol{x}')$ としてマラハノビス距離を計算する．

$$a(\boldsymbol{x}') = (\boldsymbol{x}'-\hat{\boldsymbol{\mu}})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}'-\hat{\boldsymbol{\mu}}) \tag{9}$$

3. 閾値を求め，2 で求めた $a(\boldsymbol{x}')$ と比較し，外れ値かどうかを判定する．なお，N が大きいとき，多変量のホテリング T^2 理論の異常度は自由度 M のカイ二乗分布で近似できる．閾値を求める式は以下で表される．

$$1-\alpha = \int_0^{\chi_M^2(\alpha)} f\chi(x|M)dx \tag{10}$$

式 (10) の α は分布の上側確率値であり，α と自由度 M を与えることによって求められる $\chi_M^2(\alpha)$ が閾値を表す．一般的に α は 0.01 を用いる．

(井出 [4] 参照)

2.5 One Class Support Vector Machine(SVM)

One Class Support Vector Machine（One Class SVM と略す） は，サポートベクターマシーンを学習データなしで 1 クラスで行う方法である．(高畠 [12] 参照)

2.5.1 サポートベクターマシーン（SVM）について

One Class SVM を説明するのに，まずサポートベクターマシーン（SVM と略す）について簡単に説明する．SVM とは，データを 2 つのクラスに分けるために使われる方法である．詳しく言うと，2 つに分けるために境界を引くための方法である．2 次元では線，3 次元では平面，M 次元では超平面が境界となる．

SVM では，実際に境界を引くのに「マージン最大化」という方法を用いている.「マージン最大化」という方法を説明する前に，まずサポートベクターを説明する．サポートベクターとは，引く境界から一番近い点のことをいう．そして,「マージン最大化」のマージン

とは，サポートベクターと境界の間の距離を示している．すなわち,「マージン最大化」とは，境界と 2 つのクラスそれぞれの一番近い点の距離が一番遠くなるところに境界を引く方法である（この説明はハードマージンという完全に分離できるときのものである．ソフトマージンの説明は省略する）．このようにして，分類したものを学習データとして学習し，新たなデータに対して 2 つのうちどちらのクラスに属するか判定する方法である．実際には分離が簡単な超平面で済むために，カーネルトリックというデータの空間を判別しやすいように写像する方法を用いて多様な判別を可能にしている．(高畠 [12] 参照)

2.5.2 One Class Support Vector Machine について

One Class SVM は，SVM と異なり学習データを用いないで 1 クラスで行う方法である．全てのデータのクラスを 1 とし，原点のみが -1 となるように与えられたデータをガウスカーネル等のカーネルを用いて，特徴空間に写像 ϕ によって写像する．写像した特徴空間上で原点との距離が最大となるところに境界を引き，境界を挟んで原点とその反対側とでクラスを分けることで外れ値を見つける方法である．また，境界を $< \boldsymbol{\omega}, \phi(\boldsymbol{x}) > -\rho = 0$（$\boldsymbol{\omega}$ は境界と直交する原点とは反対向きのベクトルであり，ρ は切片である）として判別を行う．データ $\boldsymbol{x}$ が高密度領域に属するかどうか，すなわち正常な値かどうかは

$$F(\boldsymbol{x}) = \mathrm{sign}(< \boldsymbol{\omega}, \phi(\boldsymbol{x}) > -\rho) \tag{11}$$

で判定し，$F(\boldsymbol{x})$ が $+1$ のとき，正常値であると判定される．したがって，集合 $\{\boldsymbol{x}|F(\boldsymbol{x}) = +1\}$ が，正常値の推定量である．

One Class SVM において，推定された高密度領域に属さない値すなわち外れ値の割合は，引数として設定するパラメータ $\nu \in (0, 1]$ に従う．本論文で ν は，0.005 を用いた．(高畠 [12] 参照)

なお，本論文では R の kernlab パッケージを利用した．(Karatzoglou[5] 参照)

2.6 Local Outlier Factor (LOF)

Local Outlier Factor（LOF と略す）は，局所外れ値因子法とも呼ばれる方法であり，局所密度の概念に基づいている方法である．局所密度とは，ある任意の 1 つの点がその点の近傍点 k 個とどれだけ密であるかを示している．これを式で表すと式 (12) になる．

$$(\text{局所密度}) = \frac{1}{(\text{近傍 } k \text{ 個の点からの近傍有効距離の平均})} \tag{12}$$

観測点 $\boldsymbol{x}'$ の k 近傍を $N_k(\boldsymbol{x}')$ とする．$N_k(\boldsymbol{x}')$ の要素をすべて含み $\boldsymbol{x}'$ を中心とする円の半径を $\varepsilon_k(\boldsymbol{x}')$ とする．これらから，近傍有効距離という量を以下のように定義する．

近傍有効距離の定義

距離 d が定義された M 次元空間において，$\boldsymbol{x}$ から $\boldsymbol{x}'$ への近傍有効距離 $l_k(\boldsymbol{x} \to \boldsymbol{x}')$ は次のように定義される．

$$l_k(\boldsymbol{x} \to \boldsymbol{x}') = \begin{cases} \varepsilon_k(\boldsymbol{x}) & (\boldsymbol{x}' \in N_k(\boldsymbol{x}) \text{ かつ } \boldsymbol{x} \in N_k(\boldsymbol{x}')) \\ d(\boldsymbol{x}, \boldsymbol{x}') & (\text{それ以外のすべての場合}) \end{cases} \tag{13}$$

点と点の距離が非常に近く，お互いがお互いの k 近傍となっている場合のみ $\boldsymbol{x}$ の k 近傍球の半径で値を置き換えるが，基本的に近傍有効距離は普通の距離と同じである．一般的には $\varepsilon_k(\boldsymbol{x}) \neq \varepsilon_k(\boldsymbol{x}')$ であるため，近傍有効距離は一般的な距離の定義を満たさないことに注意する．

LOF の異常度については，局所密度の比の平均として式 (14) で表される．

$$a_{LOF}(\boldsymbol{x}') = \frac{1}{|N_k(\boldsymbol{x}')|} \sum_{\boldsymbol{x} \in N_k(\boldsymbol{x}')} \frac{d_k(\boldsymbol{x}')}{d_k(\boldsymbol{x})} \tag{14}$$

なお，$|N_k(\boldsymbol{x}')|$ は集合 $N_k(\boldsymbol{x}')$ の大きさであり，k 近傍球の境界でタイとなるデータがなければ k と等しい．また，$d_k(\boldsymbol{x}')$ は，$\boldsymbol{x}'$ の k 近傍内の各点 $\boldsymbol{x}$ における $\boldsymbol{x}$ から $\boldsymbol{x}'$ への近傍有効距離を平均したもので，式 (15) で定義される．

$$d_k(\boldsymbol{x}') = \frac{1}{|N_k(\boldsymbol{x}')|} \sum_{\boldsymbol{x} \in N_k(\boldsymbol{x}')} l_k(\boldsymbol{x} \to \boldsymbol{x}') \tag{15}$$

式 (13) は，やや人工的な定義であるが，それは尺度変換を行ったことによって，分母が 0 になることを防ぐ工夫と考えられる．(井出 [4] 参照)

本論文では $k = 10$ とし，式 (14) に対する閾値は 2.57 を用いたが，詳細は第 5 章を参照のこと．

なお，本論文では R の DDoutlier パッケージを利用した．(Madsen[6] 参照)

3　シミュレーションの概要

本論文では，外れ値検出方法の特性を調べ，性能比較をするためにシミュレーションを行う．用いる方法は，1 次元データに対する方法がホテリング理論，箱ひげ図，スミルノフ・グラブス検定であり，多次元データに対する方法が多変量のホテリング T^2 理論，One Class SVM，LOF である．1 次元のデータと，多次元のデータでそれぞれに対していくつかのシミュレーションを行う．

1 次元のデータに対しては，以下の 3 つのシミュレーションを行った．

1. 各方法の特性を見るために行うシミュレーション
2. 性能を比較することを目的に外れ値の個数を固定して行うシミュレーション
3. 性能を比較することを目的に外れ値の出現確率を二項分布で決めて行うシミュレーション

多次元のデータに対しては，以下の 2 つのシミュレーションを行った．

4. 性能を比較するために行うシミュレーション

5. 外れ値を含まない山の分布を変えて行うシミュレーション

以上 5 つのシミュレーションは以後，シミュレーション 1 等と呼ぶことにする．

4　1 次元データに対するシミュレーション

1 次元のデータに対するシミュレーションは，第 3 章で示したように 3 つの方法に対して 3 つの設定で行う．なお，どのシミュレーションもサンプルサイズ 100，試行回数 1000 回とした．

4.1　シミュレーション 1

シミュレーション 1 は，各方法の特性を見るために行うシミュレーションである．ここでのデータは，外れ値を含みやすいという理由から自由度 3 の t 分布を用いる．

各方法に対して行ったシミュレーションで各方法が検出した外れ値総数を表 1 に示した．

表 1: 各方法の外れ値総数

方法	ホテリング理論	箱ひげ図	スミルノフ・グラブス検定
個数（単位：個）			
外れ値総数	2610	5617	2304

表 1 からは，ホテリング理論とスミルノフ・グラブス検定の外れ値総数が 2500 個前後であるのに対して箱ひげ図のみが倍以上の 5617 個であることから，外れ値ではない値をかなり外れ値として検出する可能性が高い方法であると言える．そのため，箱ひげ図の改善が必要ではないかと考えた．

4.2　箱ひげ図の改善

2.2 節で述べたように箱ひげ図は，四分位範囲にかける係数によって閾値が決まる方法である．すなわち，精度を改善するには，この係数の値を変更すればよい．

まず，箱ひげ図によって検出された外れ値総数 5617 個というのが理論値に従った値なのかを調べる．自由度 3 の t 分布について四分位範囲を求めると 1.53 であると分かり，境界は ± 3.06 と計算される．よって，両側での外れ値が検出される理論値は，自由度 3 の t 分布では 5.6%となる．シミュレーションの結果はよく合っており理論通りであることが分かる．したがって，理論値に従って四分位範囲に掛ける適切な係数を求めていく．他の方法の外れ値を検出した総数が 2500 個前後であることから箱ひげ図も検出する外れ値総数が 2500 個前後になるように信頼区間が 97.5%となるような係数を求める．

自由度3のt分布の両側2.5%点は4.177であるので求める係数は$(4.177-0.765)/1.53 \approx 2.2$となる．すなわち，四分位範囲にかける値を 2.2 にすることで箱ひげ図が自由度 3 の t 分布において検出する外れ値総数を 2500 個前後にすることができる．改良後の箱ひげ図でシミュレーションを行うと検出する外れ値総数は，2640 個となった．箱ひげ図による方法の改善ができたと言える．

なお，平均 0 分散 1 の正規分布の場合，理論的に外れ値を検出する確率は改善前の 1.5 倍の場合が 0.7%であり，改善後の 2.2 倍の場合が 0.027%となる．

4.3 シミュレーション 2

シミュレーション 2 は，各外れ値検出方法の性能を比較するために行うシミュレーションである．ここでは，確実に外れ値と分かるデータを用いる必要があるため，二山分布で意図的に作成したものをデータに用いる．具体的には，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，外れ値の山を平均 5 分散 0.3 の正規分布に従うように乱数を発生させ，合計のサンプルサイズが 100 となるように合わせて，それを混合分布として用いる．試行回数は 1000 回である．

なお，外れ値の個数を 5 個から 1 個まで変化させて，その過程の変化やそれぞれの方法を比較する．また，箱ひげ図は，4.2 節での改善後のものを用いる．

紙面の都合上，図 1，図 2 で外れ値が 5 個の場合と外れ値が 1 個の場合の結果をそれぞれヒストグラムで示した．

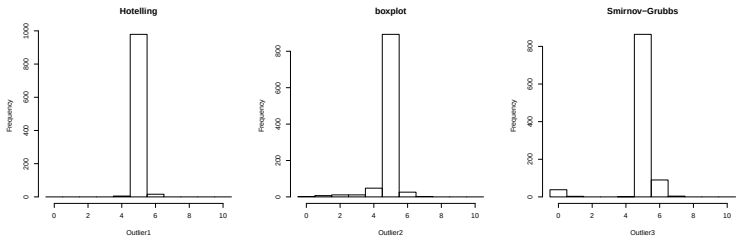

図 1: 外れ値が 5 個のときの結果

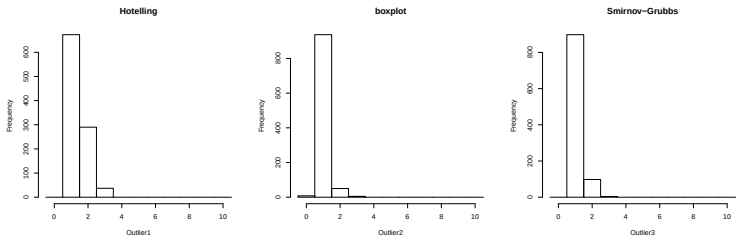

図 2: 外れ値が 1 個のときの結果

図 1，図 2 から次のことが読み取れる．ホテリング理論は，外れ値が 5 個のときほぼ完全に外れ値を検出しているが，外れ値の個数が少なくなるにつれて，本来外れ値ではない値

を外れ値として検出することが多くなっていると見て取れる．箱ひげ図は，他の方法に比べて意図した個数ではない個数で返していることも見て取れるが，その頻度は少なく外れ値の個数が少なくなっても精度は安定していることから，3 つの方法の中で一番安定して設定した外れ値を検出してると考えられる．最後にスミルノフ・グラブス検定は，外れ値が 5 個の場合から外れ値の個数が少ない場合にかけて外れ値がないと返すことがなくなった．これは，外れ値の個数が少ない方が，正常値と比べて外れ値が異常であることが際立つからだと考えられる．もしくは，サンプルサイズ 100 に対して，外れ値が 5 個というのが少し多いということも考えられるが，基本的に含まれる外れ値の個数が少ない方が良い結果を示す方法だと考えられる．

また，本シミュレーションでは，外れ値の個数を 5 個から 1 個まで変化させて行ったが，その 5 つのシミュレーションの合計で意図した外れ値の個数を正確に検出した割合は，表 2 に示したようになった．

表 2: 意図した外れ値の個数を正確に検出した割合

方法	ホテリング理論	箱ひげ図	スミルノフ・グラブス検定
割合（単位：%）			
正しく検出した割合	88.3	92.1	89

この結果からも箱ひげ図の性能が一番良いと言える．

なお，改善前の箱ひげ図の方法でもシミュレーションを行った結果，意図した外れ値の個数を正しく検出した割合は 56.1%であった．この結果から改善後の方がかなり良いと分かった．

先程も述べたが，ホテリング理論が，外れ値の個数が少なくなるほど，外れ値ではない値を外れ値として検出する場合が増えることについて，外れ値を含まない平均 0 分散 1 の正規分布でシミュレーションした結果で，各方法の外れ値を検出しなかった割合を表したのが表 3 である．

表 3: 外れ値を含まない正規分布で外れ値を検出しなかった割合

方法	ホテリング理論	箱ひげ図	スミルノフ・グラブス検定
割合（単位：%）			
外れ値の未検出割合	32	94	90

表 3 から，外れ値が存在しない場合にホテリング理論が他 2 つの方法よりもかなり外れ値を検出しやすいことが分かった．

4.4　シミュレーション 3

シミュレーション 3 は，シミュレーション 2 と同様に各外れ値検出方法の性能を比較するために行うシミュレーションである．シミュレーション 2 では，外れ値の個数を 5 個から 1 個まで変化させ，その 5 つのシミュレーションの合計で，意図した外れ値の個数を正確に検出した割合で性能を比較したが，5 つのシミュレーションの合計で比較したことで，

それぞれのシミュレーションが平等であるかが懸念点であった．そのため，外れる個数を二項分布によって決定し，二項分布の外れる確率を指定することによりシミュレーションを行う．

なお，用いるデータは，シミュレーション 2 で用いた混合分布と同様で，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，外れ値の山を平均 5 分散 0.3 の正規分布に従うように発生させたものを混合分布として用いる．

各方法が外れ値を正確に検出したか，少なく検出したか，多く検出したかの回数を表したものを表 4 に示した．

表 4: 外れ値の個数を当てた回数

方法	ホテリング理論			箱ひげ図			SG 検定		
	少なく	正しい	多く	少なく	正しい	多く	少なく	正しい	多く
回数 (単位:回)									
1%	0	625	375	9	941	50	22	954	24
2%	2	778	220	26	928	46	94	883	23
3%	16	854	130	47	913	40	240	740	20
4%	61	860	79	82	883	35	426	560	14
5%	130	823	47	113	857	30	600	389	11
6%	261	711	28	161	815	24	742	249	9
7%	403	581	16	206	774	20	845	150	5
8%	538	451	11	268	715	17	914	82	4
9%	657	336	7	327	658	15	952	45	3
10%	770	226	4	390	598	12	974	24	2

それぞれについて見ていく．まず，ホテリング理論については，外れ値が 1%のときから 4%のときまで精度が良くなっていくが，そこからは悪化する．また，1%のときは，意図した外れ値の個数より多く外れ値を検出することが多かったが，これも 4%を境に反対になりそれ以降は意図した外れ値の個数より少なく返すことが多くなる．次に箱ひげ図については，外れ値の個数が少なければ少ないほど精度が良いと分かるが，外れ値の含まれる確率が高くなってもある程度よく検出していることが分かる．そして，箱ひげ図も外れ値が含まれる確率が増えるほど意図した外れ値の個数より少ない個数を返すことが多いと分かる．最後にスミルノフ・グラブス検定についてであるが，箱ひげ図と同様に外れ値が含まれる確率が少ないほど精度が良い．しかし，箱ひげ図と違い外れ値が含まれる確率が高いとき，ひどく精度が悪いことが分かる．

まず，3 つすべての方法で見て取れた外れ値が含まれる確率が増えるほど，含まれる外れ値の個数より少ない個数で返されることについて考察する．これはサンプルサイズに対して外れ値の個数が多すぎることが原因として考えられる．データのうちの 10%もが異常な値だと正常値に近い値が含まれてもおかしくはない．

ホテリング理論が他の 2 つの方法と違い外れ値が 1%の確率で含まれるときが一番精度が良いというわけではないのは，4.3 節の平均 0 分散 1 の正規分布に対して行ったシミュションで明らかになった分布による影響が関係していることが考えられる．また，外れ値が含まれる確率が低いとき，意図した外れ値の個数より多い個数で返すことが多いのも同様な理由が考えられる．

箱ひげ図の結果で外れ値が含まれる確率が増えると精度が悪化するのは，予想していたことではあったが，5%くらいまではもう少し落ち幅が少ないと予想していた．これは当然

ではあるが，外れ値の個数が少ない方が性能が良いことを表していて，外れ値の含まれる確率が 5% でも少し多いことが考えられる．

スミルノフ・グラブス検定で，正確に外れ値の個数を検出する場合の精度の悪化がこれほどまでにひどいとは，予想外であった．箱ひげ図の精度とはいかなくとも，もう少し緩やかに悪化していくと予想していた．また，外れ値の個数よりも少なく返されるのは良いとして，多く返されることの少なさにも驚いた．スミルノフ・グラブス検定は，再帰的に検定を繰り返す方法であるから 1 つでも外れ値を検出すれば，外れ値を検出しやすくなる方法である．そのため，もう少し多く，意図した外れ値の個数を超えて外れ値を返すことがあると思っていたが，こういう結果になったのは，外れ値を検出しなかったことが多かったと考えられる．このことから，外れ値が含まれたのにも関わらず，外れ値を検出しなかった回数を各確率で調べてみた．表 5 がその結果である．

表 5: SG 検定において，外れ値は含まれているのに外れ値を検出しなかった回数

確率	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
回数（単位：回）										
検出しなかった回数	17	86	228	415	588	738	843	913	958	971

表 4 と表 5 より，スミルノフ・グラブス検定で，外れ値の個数より少ない結果を返す場合のうちのほとんどが，外れ値を含んでいないと返していることが分かった．

以上のことから，スミルノフ・グラブス検定において，実際の外れ値の個数より少なく返される原因は，外れ値を検出しないことが非常に多いということである．

4.3 節の平均 0 分散 1 の正規分布に対してのシミュレーションで分布による影響がみられたことで，外れ値を含まない山の分布を変更してシミュレーションを行うことにする．自由度 10 の t 分布に変更して行う．表 6 がその結果である．これは，意図した外れ値の個数を正確に返した回数を表している．

表 6: 外れ値を含まない山の分布が t 分布で意図した外れ値を検出した回数

方法	ホテリング理論	箱ひげ図	スミルノフ・グラブス検定
回数（単位：回）			
1%	276	561	730
2%	465	679	615
3%	562	671	466
4%	605	644	320
5%	566	605	201
6%	506	569	128
7%	395	527	72
8%	297	496	35
9%	211	415	18
10%	149	395	9

この結果からは，スミルノフ・グラブス検定の外れ値が少数の場合での検出力の高さが

うかがえる．箱ひげ図とホテリング理論は，同じような結果を返しているが，箱ひげ図の方が精度が高いことが分かる．

分布を変更することで，全体的に精度は落ちるが優劣の順位等の変化はないと考えていた．スミルノフ・グラブス検定の外れ値が含まれる確率が 1%のときの結果は，想定よりも高かったが，おおむね予想通りの結果となった．

スミルノフ・グラブス検定が 1%以外で性能が悪い原因は他の方法より厳しい基準で判定しており，1%の段階で多いのと少ないのとでつり合いが取れてしまい，そこからすでに精度の悪化が始まるからだと説明できる．したがって，それをカバーするように再帰的な方法を用いているが，カバーしきれていないことが示されたと言える．以上のことから考えると，$\alpha = 0.01$ より少し大きな値に変更することで改善することも考えられるが，値が大きすぎると再帰が強すぎて検出しすぎになることが考えられる．4.1 節での t 分布で行った考察より，現在の設定が基本的には他の方法と同等であると判断して α の修正は考えないこととする．

1 次元のデータに対するシミュレーションを通して，検出力や精度が一番高いのは改善した箱ひげ図であると言えるが，外れ値の個数がごくわずかであることが分かっているときは，スミルノフ・グラブス検定を行うことが最善である可能性もある．しかし，そのような事前情報がある場合はあまりないので，改善後の箱ひげ図が一番良いと言える．

5　多次元のデータに対するシミュレーション

多次元のデータに対するシミュレーションは，第 3 章で示したように 3 つの方法に対して 2 つの設定で行う．なお，どのシミュレーションもサンプルサイズ 100，試行回数 100 回とした（多次元データでは計算時間がかかるため試行回数はこれ以上増やせなかった）．

データに関しては，外れ値を検出することが困難なシミュレーションデータとして，多次元混合正規分布

$$(1-\alpha)N_M(\mathbf{0}, I) + \alpha N_M(\delta \boldsymbol{e}_1, \lambda I) \tag{16}$$

を用いる (Peña and Prieto[8]，和田 [13] 参照)．$N_M(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は M 次元の平均 $\boldsymbol{\mu}$，分散共分散行列 $\boldsymbol{\Sigma}$ に従う多次元正規分布を，α が外れ値の割合，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．なお，式 (16) は，(試行回数) $\times (1-\alpha)$ だけ $N_M(\mathbf{0}, I)$ に従う乱数を発生させ，(試行回数) $\times \alpha$ だけ $N_M(\delta \boldsymbol{e}_1, \lambda I)$ に従う乱数を発生させ，これらを混合分布とすることを意味している．

本論文で用いた基本のパラメータは，$\delta = 10$，$\lambda = 0.01$ である．このパラメータは，和田 [13] で用いられていたパラメータを参考にして決定した．

また，LOF の閾値であるが，独立に平均 0 分散 1 の正規分布に従う 2 次元データを大きさ 100 で 1000 回発生させて，$a_{LOF}(\boldsymbol{x}')$ の経験分布の上側 2.5%点を探した．すなわち，$a_{LOF}(\boldsymbol{x}')$ の全体での分布でそれより大きくなる値が 2500 個となるように定めたということである（2.5%という値は 4.1 節の結果を準用した）．その結果，閾値として 2.57 を用いることにした．

5.1 シミュレーション 4

式 (16) を用いて，外れ値を含まない山と外れ値の山の混合分布を発生させ各方法で検定を行う．

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 個まで変化させた．図 3，図 4 は，紙面の都合上，ぞれぞれ外れ値が 5 個のデータの結果と，1 個のデータの結果をヒストグラムで示したものである．

なお，$\delta = 3$ の設定で 2 次元で外れ値を 5 個含むデータでのシミュレーションも行ったが，試行回数 100 回のうち半数以上で外れ値を検出できなかった．このことから，$\delta = 3$ では外れ値とするのに不十分であったと考えられる．

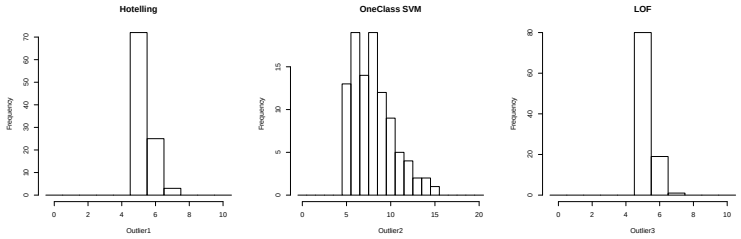

図 3: 2 次元データで外れ値 5 個

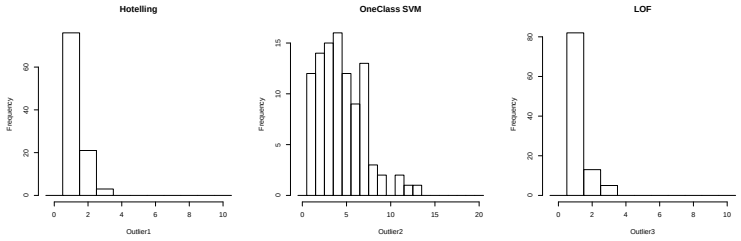

図 4: 2 次元データで外れ値 1 個

2 次元のシミュレーションでは，LOF が一番精度が良く，その次にホテリング T^2 理論となっている．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ため，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．しかし，One Class SVM のパラメータ ν を調整しても他の 2 つの方法のような形にはできずこれが限界であった．

One Class SVM の結果のヒストグラムの幅が広く，他の方法の結果のようにならないのは，One Class SVM 自身の外れ値を検出する境界が厳しく，用いた分布では外れ値を検出しすぎることが多いためと考えられる．また，ホテリング T^2 理論と LOF は，外れ値の個数を変化させた場合に LOF の方が少し精度が良いというくらいで，違いはあまり見られなかった．よって，違いを見るために次元をあげたシミュレーションを行う．3 次元から 5 次元で外れ値が 5 個の場合を検討したが，ここでは図 5 において 5 次元で外れ値が 5 個の結果のみ示す．

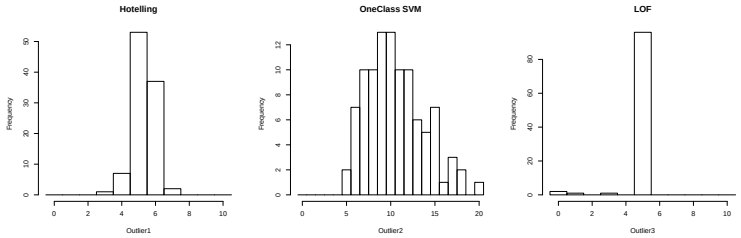

図 5: 5 次元データで外れ値 5 個

図 3，図 5 を見比べると，ホテリング T^2 理論は，次元が上がると横に広がる結果となり，反対に LOF は，精度が良くなっていることが分かる．これはホテリング T^2 理論が，M 次元の系の異常度を単一の指標で表すことで次元が高くなるにつれて，低い次元のみに生じる異常がかき消される傾向にあることが原因だと考えられる．LOF は，局所密度によって判定しているから次元が増えるとその分異常な値の周りの密度が顕著に空いて判断しやすくなっていると考えられる．ただし，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る．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．

この結果からも次元が高いと LOF の精度が良くなっていて，2 次元の時点でも精度としては高いことから，LOF が一番精度が良いと言える．

5.2 シミュレーション 5

本シミュレーションは，5.1 節の結果で One Class SVM の結果が横に広く，意図した外れ値の個数を正確に検出しなかったことが，外れ値を含まない山の分布が正規分布であることが影響しているのではないかと考え，外れ値を含まない山の分布を自由度 10 の t 分布に変えたものである．

示したそれぞれの図は，図 6 が 2 次元データでの結果，図 7 が 4 次元データでの結果である．外れ値の数は，2 つとも 5 個で行った．

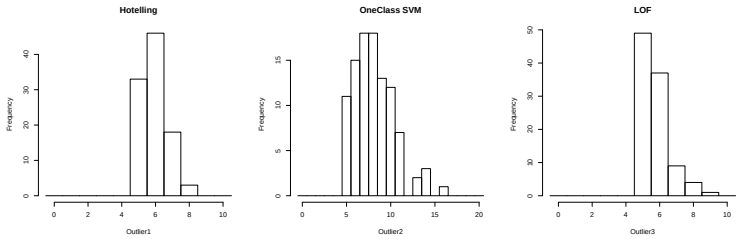

図 6: 2 次元データ，外れ値を含まない山が t 分布

図 6 からは，外れ値を含まない山の分布を t 分布に変えた結果，ホテリング T^2 理論の精度がかなり悪いという結果になった．2 次元データの場合は，5.1 節でホテリング T^2 理論と LOF の精度はほとんど同じであったことから同じような結果になることを予想していたが，ホテリング T^2 理論の結果から LOF の精度の高さが分かる結果になった．一方，One

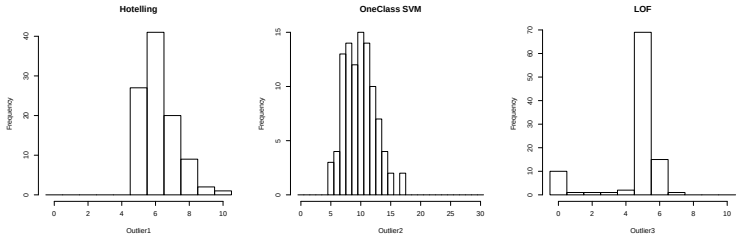

図 7: 4 次元データ，外れ値を含まない山が t 分布

Class SVM の結果のヒストグラムの形は 5.1 節のシミュレーション結果からは変化したが，横に広がっていることは変わらず意図した外れ値の個数を正確に検出しなかった．

そして図 7 からは，2 次元のときよりもホテリング T^2 理論のヒストグラムの形が横に広がり精度も悪化した．LOF については横には広いが精度は高いと考えられる．One Class SVM については 2 次元と同様に結果が良くなることはなかった．

6　まとめ

1 次元データの外れ値検出方法と多次元データの外れ値検出方法それぞれについてその特性と性能を考察する．

6.1　1 次元データに対応する方法

特性に関してはまず，箱ひげ図の外れ値を検出する方法は改善することでかなり外れ値を検出する精度が向上することが分かった．また，ホテリング理論は，平均 0 分散 1 の正規分布に従って発生させた乱数に対して外れ値の検出を行うと外れ値を 1 つ 2 つ検出してしまい，ベースとなる正規分布にごくわずかの外れ値が含まれている分布に対しての外れ値検出の精度が悪いということが分かった．スミルノフ・グラブス検定は，外れ値の個数が少ないデータに対しての外れ値の検出力が高いと分かった．

性能に関しては，外れ値の個数が非常に少ないときはスミルノフ・グラブス検定の方が性能が良かったが，本論文のシミュレーションで性能が良かったのは外れ値が含まれる確率が 1%のときのみであったことと，数字の差もほとんどなかったこと，実際に用いるとき外れ値の個数や含まれている確率などは分からないことがほとんどであることを考慮すると，ほとんどの場合において精度が一番良かった改善後の箱ひげ図が 3 つの方法の中では一番性能が良いと言える．ただし，改善に用いた 2.2 倍という数値はサンプルサイズ 100 でのものであり，大きさが違えばこの倍率の調整が必要かもしれない．例えば，箱ひげ図の係数を変える試みは野呂・和田 [7] にも見られる．そこでは別の観点から 1.724 という数値を導いている．さらに，箱ひげ図が四分位点を基礎にした方法であるため，他の方法と比べて分布の変化に関してロバストであることにも大きな特徴と言える．

6.2 多次元データに対応する方法

特性に関してまず，多次元のホテリング T^2 理論は，次元数が低いときほど性能が良く，次元数が高くなるほど性能が段々と悪化していくと分かった．LOF は，反対に次元数が高くなっても性能が悪化するどころか良くなると分かったが，次元数が低くても性能は悪くない．また，サンプルサイズが検定に大きな影響を与えることが分かった．One Class SVM は，外れ値を検出したいデータのみで外れ値検出を行うと，外れ値を検出する個数がパラメータ ν に依存してしまうことが分かった．大きさ 100 のデータに対し，外れ値を 5 個と指定しても ν を 4%としたら，4 個と返すことが一番多くなるということである．これを改善するために外れ値を含まないデータを学習させて検定を行ったが，他の 2 つの方法のような結果にはならなかった．この方法は，異常検知で多く用いられており，それは今回のように正常データを学習させて外れ値を検出する設定と似たものであるが，単純に異常を見分けるのは困難な印象を受けた．シミュレーションの結果からは異常は見つけられるものの正常なものを異常と判定することが残る印象である．もっと複雑な形状の分布の場合は検討していないのでその場合は役に立つことが考えられる．また，今回検討しなかったパラメータを調整すれば今回のシミュレーション設定でも改善する可能性があるが，単純な分布に対してそのような調整が必要だとすると扱いが難しいと言える．

性能に関しては，外れ値を含まない山を正規分布に従うように発生させ，なおかつ次元数が低いときの精度はホテリング T^2 理論と LOF でほとんど変わらなったが，分布を変えたときや，次元数を高くしたときの結果は，LOF の方が良い結果を示したことから，多次元のシミュレーションで用いた 3 つの方法のうち 1 番性能が良いのは LOF と言える．ただし，次元のわりにサンプルサイズが小さい場合は，外れ値をまったく検出できないということが起きる可能性があることには注意すべきであろう．

7 おわりに

本論文を通して，比較的簡単に用いることのできる外れ値検出方法について，その特性や性能について知ることができた．しかし，多次元データのシミュレーションについては，時間的な制約と PC のスペックの問題で繰り返し回数を減らして低い次元でしか実験せざるを得なかったことは残念である．

本論文の 1 次元データのシミュレーションで用いた箱ひげ図であるが，当初は性能に関して期待していなかった．一般に知られた方法であり，昔から存在する方法ということで用いることにしたが，少し改善することでかなり性能が良くなることは予測しておらず，意外な結果であった．

外れ値を検出したうえで外れ値を除くことが必要なこともあるが，そうでない場合もある．今後は，そういった点に注意しつつ本論文で学んだことを活かしたい．

参考文献

[1] 青木繁伸 (2015) :「スミルノフ・グラブス検定」，
http://aoki2.si.gunma-u.ac.jp/lecture/Grubbs/Grubbs.html . (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] 「外れ値検出（知識）」:
https://sites.google.com/site/scriptofbioinformatics/mian-qiangmemo/waire-zhi-jian-chu-zhi-shi/ . (2020/6 閲覧)

[4] 井出剛 (2015) :「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, pp9-41,73-77.

[5] Karatzoglou, A (2019) :「Package ‘ kernlab ’」, CRAN,
https://cran.r-project.org/web/packages/kernlab/kernlab.pdf/ . (2020/6 閲覧)

[6] Madsen, J.M (2018) :「Package ‘ DDoutlier ’」, CRAN,
https://cran.r-project.org/web/packages/DDoutlier/DDoutlier.pdf/
(2020/6 閲覧)

[7] 野呂竜夫・和田かず美 (2015) :「統計実務におけるレンジチェックのための外れ値検出方法」, 統計研究彙報 第 72 号, pp.41-54.

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

[9] R Documentation :「Box Plot Statistics」,
https://stat.ethz.ch/R-manual/R-patched/library/grDevices/html/boxplot.stats.html/, (2020/6 閲覧)

[10] 「sfchaos's blog」:
http://sfchaos.hatenablog.com/entry/20140518/p1/ . (2020/6 閲覧)

[11] 白旗慎吾 (1992) :「統計解析入門」, 共立出版株式会社, pp.30-31.

[12] 高畠泰斗 (2007) :「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/ . (2020/11 閲覧)

[13] 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.