統計数理(2024)
第 72 巻 第 2 号 245–260

公的統計における外れ値への対処とソフトウェア
―R の MSD 法パッケージの実装について―

和田 かず美†

（受付 2024 年 1 月 9 日；改訂 2 月 10 日；採択 2 月 14 日）

要　　旨

公的統計の分野において，連続値の外れ値検出法で最も多く使用されているのはレンジチェックと呼ばれる単変量の方法であるが，調査統計の成果物が従来の統計表に加えて個別データの提供も進むにつれて，多変量な外れ値への対処の重要性も徐々に認識されつつある．本稿では，単変量の外れ値と多変量の外れ値の違いについて解説し，多変量の外れ値検出法のうち単峰で対称な楕円分布を前提とする手法の 1 つである MSD(Modified Stahel-Donoho)推定量による方法を実装した R のパッケージ RMSD 及び RMSDp とその利用方法について紹介する．

キーワード：データクリーニング，楕円分布．

1. はじめに

公的部門の統計調査データの集計処理においては，何らかのエラーや，間違いではないが他の大部分のデータとは傾向が違うデータの発生を避けることが難しい．そのような観測値は「外れ値」と呼ばれ，広義には分布の裾にあり他の大多数のデータとは異なる傾向を持つものと定義される（Eurostat, 2014)が，厳密には統計調査の各集計プロセスに応じて様々に定義される．

調査票情報を電子化し，記入ミスや誤りを検出するデータチェックのプロセスにおいては，検出すべき外れ値は誤りの可能性がある観測値が対象となる．また，例えば欠測を回帰モデルに基づき補完するプロセスにおいては，検出すべき外れ値は梃子比と Cook の距離が共に大きいデータが回帰推定及びそれに基づく補完値を歪める可能性がある観測値となるだろう．データのエラーを除外・修正し，必要に応じて欠測を補完した後の，集計値に乗率と呼ばれる標本抽出率の逆数を乗じて統計表の数値を作成する母集団推定プロセスでは，誤りではなくとも，特異な値が大きな乗率を持ち，集計結果に影響するようなデータが検出すべき外れ値となり，そのようなデータについては乗率の調整等の措置を検討することになる．統計作成プロセス毎の外れ値の処理については，Wada (2020)などに詳しい．

公的統計における伝統的な外れ値検出法は，何らかの方法で正常な値の範囲を決めるレンジチェックと呼ばれる単変量の方法である．この正常値範囲をデータ自体から導き出す方法については，野呂・和田 (2015)を参照されたいが，複数の相関のある数量項目を含む調査統計において，単変量の外れ値検出法を個々の項目に適用する場合，特定の変数で極端な値をとらない

† 総務省 統計研究研修所：〒 185–0024 東京都国分寺市泉町 2–11–16

が調査項目間の相関関係が他の大部分の観測値と異なるタイプの外れ値を検出することができない．本稿では，そのような外れ値を「関係性の外れ値」と呼ぶことにする．公的統計の集計プロセスでは，そのような外れ値はデータを層化したり，相関の高い変数との比率についてレンジチェックを適用することにより疑似的に対応することが多く，実務上多変量の外れ値検出法はあまり普及はしていない．この理由としては以下のようなことが考えられる．

- 単変量の方法と比較して技術的に難易度が上がる．
- 実データでの正解（真の外れ値）はわからない．
- 多くの検出法が提案されているが，ベストな方法は状況により変わりうる．
- 検出された外れ値の審査や対処が単変量のものよりも困難である．

本稿では，楕円分布モデルを前提とした多変量な外れ値検出法の一つである MSD（Modified Stahel-Donoho）推定量に基づく方法とその実装パッケージについて解説し，公的統計への適用事例について併せて紹介する．これは，先に述べた多変量の方法が普及しない理由のうち，技術的難易度の緩和に貢献するものである．なお，p 次元確率変数 $\boldsymbol{x}$ が楕円分布に従うとき，その密度関数 $f(\boldsymbol{x})$ は次のように表現される．

$$f(\boldsymbol{x}) \propto g\left[(\boldsymbol{x}-\mu)^\top \Sigma^{-1}(\boldsymbol{x}-\mu)\right]$$

ここで，μ は $\boldsymbol{x}$ が従う楕円分布の平均値ベクトル，Σ は分散共分散行列とする．また，g は非負値関数で，確率変数の定義域全体での積分が有界となる．公的統計分野において，家計における所得と消費や事業所・企業における従業者数と売上高などは，対数変換などの変数変換によって対称化することで楕円分布と見なすことができる．探索的データ解析に基づく変数変換等により多変量正規分布に近づくデータも，典型的な楕円分布モデルに基づく外れ値検出法の適用範囲となる．

まず，第 2 節では本稿で取り上げる MSD 法に関する先行研究について紹介し，第 3 節でその実装アルゴリズムと並列化の方法についても解説する．第 4 節ではこれらのソフトウェアのパッケージ化とその仕様を実例を交えて紹介する．

2. 先行研究

本稿で取り上げる MSD 法は，Patak（1990）が考案した，Stahel-Donoho（SD）推定量と射影追跡（PP: Projection Pursuit）法という二つの手法の組み合わせである．SD 推定量は Stahel（1981）と Donoho（1982）がそれぞれ考案したもので，観測値を様々な方向の基底ベクトルに射影し基底ベクトル上の中心からの乖離度に応じて各観測値に付与したウェイトを用いて，平均値ベクトルと分散共分散行列のロバスト推定を行う．SD 推定量は，変数の数によらず高い破局点を持つことが知られている（Maronna and Yohai, 1995）．続く PP 法のステップでは，SD 推定量である分散共分散行列から固有値分解による主成分分析（PCA: Principal Component Analysis）を行う．PCA は，相関のある変数を含むデータセットを観測値の持つ多様性を極力保持しつつ主成分得点に変換するが，主成分同士は相関を持たないために前節で述べた「関係性の外れ値」を想定する必要がない．このため，個々の主成分に対しては単変量の外れ値検出法の考え方を適用することができる．

周知のように，第一主成分は最大の固有値を持つ固有ベクトルに観測値を射影した結果であり，分散が最も大きくなるよう設定される．第二主成分は，データから第一主成分に直交する空間への射影方向の中で，同様に分散を最大化する方向を示す．第三主成分はデータから第一・第二主成分に直交するという制約の下で，同様に分散を最大化する方向を示すベクトルが

選択される．第一主成分は多変量データの共通性を代表する尺度，通常サイズ因子と呼ばれているが，第二主成分以降は共通性よりも徐々に観測値の特異性が集約されていくため，外れ値検出に関してはむしろ有用性が高い．

このように，Patak (1990)が考案したMSD法は，SD推定量として得られたロバストな分散共分散行列にPP法を適用してさらに良い平均値ベクトルと分散共分散行列を推定し，これに基づき各観測値のマハラノビス平方距離を算出し，その数値が大きなものがF検定統計量を目安に外れ値として検出される．このMSD法は，大標本での破局点が約0.5と高く，直交変換不変(orthogonally equivariant)である (Franklin and Brodeur, 1997)．

伝統的な外れ値への対応策としては，母集団が正規分布などの特定の分布に従うことを前提に仮説検定を行う方法論や，回帰など異なるモデルを前提とするものも存在するが，ここでは観測値が平均値ベクトルと分散共分散行列で決まる楕円分布族に従うことを前提とする検出法に焦点を当てる．統計調査の集計実務への適用可能性を重視し，国連欧州経済委員会(UNECE)やEU統計局(Eurostat)などが行っている公的統計分野の取り組みであるUNSC/UNECE刊行のStatistical Data Editingシリーズや，2001〜2003年にEU域内の政府統計部局・大学及び企業の統計専門家が参加したEUREDITプロジェクト[https://www.cs.york.ac.uk/euredit/euredit-main.html]の報告書などを調査し，多変量外れ値検出法の先駆的な実用化事例であるカナダ統計局の年次卸売・小売業調査(AWRTS)に使用されたMSD法に着目した．

カナダ統計局での適用方法とその手法の詳細については，Franklin and Brodeur (1997)に詳しい．EUREDITプロジェクト報告書の一部であるBéguin and Hulliger (2003)においてこれが紹介され，改善点についていくつか指摘がなされている．和田 (2010)は，カナダ統計局の方法とEUREDITでの指摘に基づいた改良手法を統計ソフトRで実装し，実際にBéguin and Hulliger (2003)による改良が外れ値検出の性能を向上させることを確認している．

3. 実装

3.1 Modified Stahel-Donoho 推定量のアルゴリズム

大きさnでp次元のデータセット$\boldsymbol{X}$について，そのi番目の観測値を$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$とすると，Béguin and Hulliger (2003)に基づくMSD法のアルゴリズムは以下のとおり．

(1)直交基底の作成

ランダムなb個の直交基底$\boldsymbol{\nu}^{(k)}$ $(1 \leq k \leq b)$を作成するために，$b \times p \times p$個の一様乱数を生成してb個のp行p列の行列を作成，これにGram-Schmidtの直交化法を適用する．Béguin and Hulliger (2003)は，pが大きくなっても検出性能を維持するために必要な基底数bについて，pの増加に伴い指数関数的に増加すると述べたMaronna and Yohai (1995)を根拠として1変数当たり$\exp(2.1328 + 0.8023p)/p$としているので，$b = \exp(2.1328 + 0.8023p)$となる．

(2)射影と残差の標準化

直交基底行列$\boldsymbol{\nu}$のj番目の基底ベクトル$\boldsymbol{\nu}_j$ $(1 \leq j \leq p)$が張る直線上に観測値$\boldsymbol{x}_i$を射影した座標$\boldsymbol{\nu}_j^\top \boldsymbol{x}_i$を中央値(median)と中央絶対偏差(MAD: median absolute deviation)を用いて

$$r_{ij} = \frac{|\boldsymbol{\nu}_j^\top \boldsymbol{x}_i - \mathrm{med}(\boldsymbol{\nu}_j^\top \boldsymbol{x})|}{\mathrm{mad}(\boldsymbol{\nu}_j^\top \boldsymbol{x})/0.674} \tag{3.1}$$

によりロバストに標準化し，標準化残差r_{ij}を求める．ここで，medは中央値，madはMADを示す．データが標準正規分布に従うと仮定すると，MADは0.674になるため，MADを0.674で割ることにより標準偏差と基準を揃える．

(3)残差に基づく次元別ウェイトの算出

i 番目の観測値 $\boldsymbol{x}_i$ を j 番目の基底ベクトル $\boldsymbol{\nu}_j$ に射影して得られる標準化残差 r_{ij} から，

$$\tilde{r}_{ij} = \begin{cases} r_{ij} & \text{if} \quad 0 \le r_{ij} \le c, \\ c^2/r_{ij} & \text{if} \quad c \le r_{ij}, \end{cases} \qquad c = \sqrt{\chi^2_{p,0.95}} \tag{3.2}$$

により刈り込み残差 $\tilde{r}_{ij}$ を算出し，

$$w_{ij} = \tilde{r}_{ij}/r_{ij} \tag{3.3}$$

により次元別のウェイト w_{ij} を得る．

Maronna and Yohai (1995)は，定数 c について，$\boldsymbol{X}$ が p 次元の正規分布に従う場合に射影残差の二乗が漸近的に自由度 p のカイ二乗分布に従うことを根拠としている．

(4)一次ウェイトの算出

各観測値に対し，直交基底 1 組毎に得られる p 個のウェイトについて，まず

$$w_i = \prod_{j=1}^{p} w_{ij} \tag{3.4}$$

により次元別ウェイトの積和をとると，各観測値毎に基底数と同じ b 個のウェイト $w_i^{(k)}$ が得られるので，$\breve{w}_i = \min\limits_{1\le k\le b} w_i^{(k)}$ として，観測値別に b 個の中から最小のウェイトを選択し，これを一次ウェイト $\breve{w}_i$ とする．

基底ベクトルへの射影後，単変量の位置と尺度の推定量に中央値と MAD を使用し，残差 $r_i = (r_{i1}, \ldots, r_{ip})$ としてウェイト $w_i = w(r_i)$ が正で連続かつ w_i と $r_i^2 w_i$ が $r_i \le 0$ で有界であれば，Donoho (1982)に基づき，SD 推定量は有限標本での破局点の最大値である $\varepsilon_0^* = [(n-p+1)/2]/2$ を達成する (Tyler, 1994)が，式(3.3)のウェイト関数はこの条件を満たす．

(5)重み付き主成分分析

一次ウェイト $\breve{w}_i$ を用いて，

$$\begin{aligned} \hat{\boldsymbol{u}} &= \frac{\sum_{i=1}^n \breve{w}_i \boldsymbol{x}_i}{\sum_{i=1}^n \breve{w}_i} \\ \hat{\boldsymbol{V}} &= \frac{\sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{u}})(\boldsymbol{x}_i - \hat{\boldsymbol{u}})^\top \breve{w}_i^2}{\sum_{i=1}^n \breve{w}_i^2} \end{aligned} \tag{3.5}$$

により平均値ベクトル $\boldsymbol{u}$ と分散共分散行列 $\boldsymbol{V}$ の推定値を求める．

(6)主成分射影と二次ウェイト算出

主成分分析により，p 次元の固有ベクトルが p セット得られるが，これも一組の直交基底とみなすことができる．一次ウェイト算出時と同じようにこれらの固有ベクトルに観測値を射影し，式(3.1)，(3.2)，(3.3)及び(3.4)により算出したウェイトの中から観測値別に最小の値を持つものを選び，これを二次ウェイトと呼ぶ．

(7)最終ウェイトの決定と平均値ベクトル・分散共分散行列の推定

各観測値別にさらに一次ウェイトと二次ウェイトを比較し，値が小さい方を最終ウェイトとする．この最終ウェイトを用いて再度式(3.5)により平均値ベクトルと分散共分散行列を推定する．

(8)マハラノビス距離の算出と外れ値の特定

ステップ(6)で推定した平均値ベクトルと分散共分散行列を用いて，

$$D^2(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \hat{\boldsymbol{u}})^\top \hat{\boldsymbol{V}}^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{u}})$$

によりマハラノビス平方距離 $D^2(\boldsymbol{x}_i)$ を算出する．$D^2(\boldsymbol{x}_i)$ の検定統計量 F_i は，自由度 p 及び $(n-p)$ の F 分布に従い，

$$F_i = \frac{(n-p)n}{(n^2-1)p} D^2(\boldsymbol{x}_i)$$

により得られる．外れ値と判定する基準は，Franklin and Brodeur（1997）に準じて F 検定統計量の 99.9% 点をデフォルトとする．

3.2 シングルコア版の性能評価

Maronna and Yohai（1995）及び Peña and Prieto（2001）は，外れ値検出法の評価に次のようなモデルに従う乱数データを使用している．

$$(1-\alpha)N_p(\boldsymbol{0}, \boldsymbol{I}) + \alpha N_p(\delta e_1, \lambda \boldsymbol{I}), \quad \boldsymbol{I} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

ここで，α は外れ値の割合，p は変数の数，λ は外れ値の分散，δ が第 1 変数上の正常値からの外れ値の距離を示す．外れ値ではない観測値の量はデータセットの大きさ $\times(1-\alpha)$，原点中心で分散共分散行列が I の p 次元標準正規分布乱数で，外れ値の量はデータセットの大きさ $\times\alpha$，平均は第一変数以外 0，分散 λ の正規分布に従う．この評価用データセットは外れ値もそれ以外の観測値もそれぞれ変数間に相関のない正規分布に従うため，関係性の外れ値は発生しない．このため，和田（2010）において実装された MSD 関数の評価には，上述のモデルの外れ値ではない観測値には相関関係を導入し，

$$(1-\alpha)N_p(\boldsymbol{0}, \boldsymbol{R}) + \alpha N_p(\delta e_1, \lambda \boldsymbol{I}), \quad \boldsymbol{R} = \begin{pmatrix} 1 & & r \\ & \ddots & \\ r & & 1 \end{pmatrix} \tag{3.6}$$

に従う乱数データセットを使用した．加えて，実データはより裾が重く歪みのある分布である場合も多いため，外れ値ではない観測値に Skew-T 分布，複合ポワソン分布及び対数正規分布に従う乱数を採用した．また，Howkins-Bradu-Kass（Hawkins et al., 1984），スイスのレストラン業（Béguin and Hulliger, 2003），Herzsprung-Russell（Rousseeuw and Leroy, 2003），Bushfire（Campbell, 1989），Stackloss（Rousseeuw and Leroy, 2003），Modified Wood Specific Gravity（Rousseeuw and Leroy, 2003）といった評価用データにより，EUREDIT 版がカナダ版より外れ値検出性能が高いことが確認された．

Wada et al.（2020）では，調査実務への適用を目的として，和田（2010）で実装された関数の EUREDIT 版と，BACON 法（Billor et al., 2000），Fast-MCD 法（Rousseeuw and Van Driessen, 1999）及び NNVE 法（Wang and Raftery, 2002）を比較し，上述の乱数データに加え，四乗根変換等を施した企業の経理項目データで性能比較を行い，裾が重く歪んだデータについて MSD 法の性能が高いことを確認している．比較に使用されたソフトウェアは表 1 のとおりで，オプション等は全てデフォルト設定を採用している．

MSD 法は，特に変数間の相関が高い場合に検出力が高いが，一方で計算負荷が上述の他の手法と比較しても高く，特に変数の数 p の増加に伴い計算負荷は指数関数的に増大する．和田（2010）で実装された MSD 関数について，大きさ 100 のデータセットについて，4GB メモリの 32bit 機では 11 変数を越えるとメモリ破綻する（Wada and Tsubaki, 2013）．

表 1．使用ソフトウェア．

方法	説明
BACON	Béguin and Hulliger (2003)で公開されている S-plus のコードを R に移植（詳細は `https://github.com/kazwd2008/BEM`）
Fast-MCD	R の rrcov パッケージ収録の covMcd 関数
NNVE	R の covRobust パッケージ収録の cov.nnve 関数

3.3 並列化

和田 (2010)による MSD 関数について，より高次元のデータセットへの適用を可能とするため，Wada and Tsubaki (2013)は和田 (2010)による MSD 関数の並列化実装を行った．本節以降では，和田 (2010)で実装された MSD 関数をシングルコア版と呼ぶ．並列化版は，for や apply 等の関数による繰り返し処理を並列化するための CRAN パッケージ foreach 及び doParallel を使用し，3.1 節のアルゴリズムのうち(1)から(4)までのステップの処理を小分けにして複数のスレッドに振り分けることによりメモリ破綻を防ぐ．

具体的には，シングルコア版では計算速度を重視し，b 個の基底を構成する一様乱数の直交化や $b \times p \times p$ 個の基底ベクトルへの射影と残差・ウェイトの計算をそれぞれ apply 関数を用いて一括で計算処理するが，並列化版はこれらをより小さなチャンクに分けて複数の子プロセスで実行し，1 セットの一次ウェイトに情報を集約した後に親プロセスに戻す．このため，同じデータセットにシングルコア版と並列化版関数を実行する場合，コア数が少なければ計算速度はシングルコア版の方が速い．

並列化版の性能評価は，シングルコア版と同様にモデル(3.6)に基づく正規分布に従う乱数データに適用し，大きさ 100，20 変数のデータセットにおいて，シングルコア版による少ない変数のデータセットでの結果と比較して外れ値検出性能が落ちないことが確認された．加えて，独立行政法人統計センターが公開していた 2004 年全国消費実態調査を模した合成データである疑似ミクロデータ(現在は疑似ミクロデータの公開は終了し，一般用ミクロデータが公開されている)への適用事例も紹介された．テストに使用されたのは EPSON Pro4700 で，CPU は Core™ i5(3.33GHz)，メモリ 4GB で 32bit 版の Windows 7 Pro である．約 2 万レコードの疑似ミクロデータについて，並列化版の処理は 8 変数では 55 秒，10 変数で 150 秒を要した．

4. パッケージ化

シングルコア版も並列化版も，平均値ベクトルと分散共分散行列をロバスト推定する関数コードの形でそれぞれ和田 (2010)及び Wada and Tsubaki (2013)で公開されているが，これに外れ値判定プロセスを追加してパッケージ化し，2023 年 11 月に CRAN の審査を経て公開された．本節ではその仕様と利用方法について解説する．

4.1 シングルコア版

4.1.1 仕様

パッケージ RMSD に収録された RMSD 関数は，Béguin and Hulliger (2003)に基づく MSD 法による多変量外れ値検出の関数で，並列化版の RMSDp::RMSDp() と比較すると高速だが，高次元データへの適用はメモリ破綻を起こしやすい．引数と戻り値をそれぞれ表 2 及び表 3 に示す．

引数は，inp に行が各観測値で列が変数の行列形式で外れ値を検出する欠測のないデータを指定する．nb は基底数で，$b = \exp(2.1328 + 0.8023p)$ を変更したい場合に指定する．変数の数

表 2. RMSD 関数の引数.

引数	説明
inp	入力データ（行列形式）
nb	射影を行う基底の数（デフォルト値推奨）
sd	再現性が必要な場合に指定する乱数シード (任意の整数値)
pt	外れ値検出の閾値（デフォルト 99.9% 値）

表 3. RMSD 関数の戻り値.

変数名	説明
u	平均値ベクトル
V	分散共分散行列
wt	最終ウェイト
mah	マハラノビス平方距離
FF	F 検定統計量
cf	外れ値判定の閾値（pt の値に対応するマハラノビス平方距離）
ot	外れ値判定結果（1: 外れ値ではない 2: 外れ値）

が多くメモリ破綻を起こす場合は，デフォルトよりも小さな数を指定することで破綻を防げる可能性がある一方，外れ値検出力が低下する．この関数は直交基底の作成に一様乱数を使用しているために，同じデータについて複数回実行した場合，結果が動く可能性がある．sd の指定は，基底を構成する乱数のシードに使用されるため，関数の実行結果を固定したい場合に使用する．pt は外れ値検出の閾値でデフォルトは 99.9% 値が設定されているがこれはあくまでも目安であり，正規分布よりも裾の重いデータセットの場合は外れ値判定するマハラノビス二乗距離の閾値 cf をデータに応じて調整し，その調整値よりも大きなマハラノビス平方距離 mah のものを外れ値として検出し直す．また，企業の経理項目や世帯収入等の歪みの大きなデータは，Wada et al. (2020)での事例のように事前に適切な変換を施す．

戻り値は u がロバスト推定された観測値の平均値ベクトル，V が分散共分散行列である．この二つの推定値により楕円分布の形が決まる．wt は各観測値の最終ウェイトで，この値が 0 に近いほど観測値はデータ中心から外れている．mah は各観測値のデータ中心からのマハラノビス二乗距離で，これが F 検定統計量に対応する閾値 cf の値よりも大きい場合に外れ値として ot の値に 2 がセットされる．

4.1.2 利用例

ここでは，robustbase パッケージに収録された Herzsprung-Russell データセット(パッケージ内での名前は starsCYG)について RMSD 関数を適用し，結果を外れ値判定の閾値とともに散布図に図示するコード例を紹介する．

```
install.packages("RMSD")
library(RMSD)
data(starsCYG, package="robustbase")
o1.msd <- RMSD(starsCYG)

plot(starsCYG, pch=21, col="gray", bg=c("gray80", "black")[o1.msd$ot], cex=1.5)

# 確率楕円描画
n <- nrow(starsCYG);        d <- ncol(starsCYG)
eg <- eigen(o1.msd$V)                          # 共分散行列を固有値分解
P   <- eg$vectors                              # 固有ベクトル
D  <- matrix(rep(0, d*d), ncol=d)              # 対角成分が固有値、それ以外は 0
diag(D) <- eg$values
```

```
PP <- solve(P)                                          # 固有ベクトルの逆行列
cf <- o1.msd$cf*(n^2-1)*d / ((n-d)*n)                     # 検定統計量から距離の閾値を逆算
     # o1.msd$cf は、外れ値判定に使用したカットオフ値(デフォルトで 99.9% 点)
ax1 <- sqrt(cf * eg$values)                             # 楕円の長軸・短軸の計算

nb <- 0:200                                              # データの刻み
dw <- 2 * pi / 200;  w <- dw * nb
XA1 <- ax1[1] * cos(w);    XA2 <- ax1[2] * sin(w)
XX1 <- t(t(PP) %*% t(cbind(XA1, XA2)) + o1.msd$u)    # 有向楕円に
lines(XX1, col="gray60", lwd=3, lty=2)
points(o1.msd$u[1], o1.msd$u[2], pch=18, cex=2)  # データ中心
```

図 1 の菱形の点は推定されたデータ中心(平均値ベクトル)，外れ値の目安としている F 検定統計量の 99.9% 点を灰色の点線による楕円で示している．

4.2 並列化版

パッケージ RMSDp に収録された RMSDp 関数は，シングルコア版である RMSD::RMSD() を並列化したもので，シングルコア版よりも変数の多いデータに適用できるが，シングルコア版がメモリ破綻せず動く場合はシングルコア版の方が処理時間が短い．

関数 RNSDp の入力パラメータと出力はそれぞれ表 4 及び表 5 のとおり．RMSD::RMSD() の引数に cores と dv が追加されている．core は，関数が利用するコアあるいはスレッドの数を指定するもので，デフォルトでは関数を実行する PC で利用可能な全てのコアを使用する．dv は並列化のための処理のチャンクの大きさを指定するもので，デフォルト値の 10000 は一括で計算処理を行う行列の要素数を指定する．このデフォルト値は 4GB メモリの 32bit 機を基準に設定されているため，より大きなメモリのある PC の場合は値を増やすと処理速度が上がるが，大きくしすぎるとメモリ破綻する．戻り値はシングルコア版に準拠する．

4.2.1 利用例

ここでは，(独)統計センターが公開している一般用ミクロデータから平成 21 年全国消費実態調査の詳細品目データを使用して，RMSDp パッケージを利用した外れ値検出を行う．一般用ミクロデータは，集計表から調査の個別データを模して乱数により疑似的に作成されたもので，全国消費実態調査の詳細品目データは二人以上の世帯約 5 万世帯についての家計の収入・支出などの 430305 項目を収録し，調査の概要や符号表とともに公開されている．

データチェックを行う場合も，Wada (2020)で紹介されている個人企業経済調査のように補完を行う場合でも，全てのレコードについて全数量項目をまとめて外れ値検出するわけではな

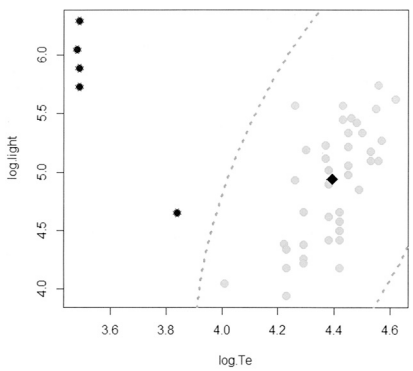

図 1. Herzsprung-Russell データセット．

表 4. RMSDp 関数の引数.

引数	説明
inp	入力データ（行列形式で 1 行が 1 つの観測値）
cores	利用するコア（スレッド）数
nb	射影を行う基底の数（デフォルト値推奨）
sd	再現性が必要な場合に指定する乱数シード (任意の整数値)
pt	外れ値検出の閾値（デフォルト 99.9% 値）
dv	チャンクの大きさ（デフォルトは最大要素数 10000）

表 5. RMSDp 関数の戻り値.

変数名	説明
u	平均値ベクトル
V	分散共分散行列
wt	最終ウェイト
mah	各観測地のマハラノビス平方距離
cf	外れ値判定の閾値（pt の値に対応するマハラノビス平方距離）
ot	外れ値判定結果（1: 外れ値ではない 2: 外れ値）

表 6. 飲料項目と変数名.

項目名	変数名	項目名	変数名
緑茶	E080	紅茶	E081
他の茶葉	E082	茶飲料	E083
コーヒー	E085	コーヒー飲料	E086
ココア・ココア飲料	E087	果実・野菜ジュース	E089
炭酸飲料	E090	乳酸菌飲料	E091
乳飲料	E092	ミネラルウォーター	E093
他の飲料のその他	E094	清酒	E096
焼ちゅう	E097	ビール	E098
ウイスキー	E099	ワイン	E100
発泡酒・ビール風アルコール飲料	E101	他の酒	E102

い．世帯調査であれば世帯の属性によりできるだけグループ内の類似性が高くグループ間の相違が大きくなるようにグループ分けを行う．このデータセットの例では，調査の対象となる世帯の収入や支出に影響があると思われる属性でグループ化し，世帯主が就業者で年齢階級が 30 歳未満，かつ世帯人員が 2 人という属性を持つ世帯に着目する．また項目について，原則として相関関係があるものを選択するが，ここでは表 6 に示す飲料の 20 項目を対象とする．

これらの項目は全て金額なので，適切な変換を施して楕円分布に近づける必要がある．どの程度の変換とするかは，散布図行列の目視結果や Box-Cox 変換（Box and Cox, 1964）のパラメータ推定結果，検出外れ値数などを勘案して決定する．Box-Cox 変換のパラメータ推定値は，0 であれば対数，0.25 で四乗根，0.5 で平方根変換ということになるが，外れ値の影響を受けやすいパラメトリックな方法であるため，極端なデータが混入している場合により小さな推定値になる傾向がある．

Box-Cox 変換のパラメータ推定結果を表 7，RMSDp パッケージを用いた MSD 法による外れ値検出結果を表 8，QQ プロットを図 2 に示す．外れ値検出結果から対数変換が最も検出数が少ないが，Box-Cox 変換のパラメータ推定結果からは少なくとも対数変換は適切ではないことがわかる．次に，QQ プロットを見ると，変換なしあるいは平方根が適切に見えるが，変換な

表 7. Box-Cox 変換のパラメータ推定結果.

E080	E081	E082	E083	E085
0.12532751	0.07193657	0.11722085	0.20877778	0.13916805
E086	E087	E089	E090	E091
0.12729565	0.06312193	-0.04562682	0.04230969	-0.00431776
E092	E093	E094	E096	E097
0.12543596	0.15176793	0.16776192	0.08195253	0.15086917
E098	E099	E100	E101	E102
-0.06567368	0.14506019	0.02271221	-0.08031569	0.11825778

表 8. 検出された外れ値の数.

データ変換	正常値	外れ値
変換なし	247	58
二乗根変換	269	36
四乗根変換	280	25
対数変換	282	23

しの場合外れ値がかなり多く検出されることと，散布図行列を比較すると平方根変換を施した方がより楕円分布に近いため，平方根変換を採用する．ただし，裾の重い分布なので外れ値と判定するための閾値を広げるか，あるいは QQ プロットで検出したい個数を決めてマハラノビス距離の大きい順に外れ値判定する．

この事例についてのサンプルコードを以下に示す．コードを流す前に，(独)統計センターのサイト[https://www.nstac.go.jp/use/archives/ippan-microdata]からデータファイル ippan_2009zensho_z_dataset.csv と符号表 ippan_2009zensho_s.xls をダウンロードし，符号表を加工する．冒頭 7 行を削除し，行番号・項目名・階層・項目番号・変数名以外の列を削除，さらに変数名が空欄の行を削除して csv 形式で ippan_items.csv という名前で保存する．

全国消費実態調査の品目項目はレベル 5 までの階層構造で，内訳を持たない品目だけを選ぶために加工した符号表の情報を使用する．具体的には，次の行の品目レベルが同じであれば内訳を持たないと判定する．

この 305 レコード 20 変数の外れ値検出には，8 コア 32GB メモリの PC(VAIO SX12，モデル VJS123C12N，Intel Core™ i7-1065G7) を使用し，計算時間は約 3 時間半であった．

```
install.packages("RMSDp")
library(RMSDp)
#
# データファイルの読込み
dat <- read.table("ippan_2009zensho_s_dataset.csv",
                header=TRUE, sep="," ,        skip=8, colClasses=c(rep("factor",7),
                rep("numeric",423)))
attach(dat)
#
# 符号表情報の読込み
items <- read.csv("ippan_items.csv", header=TRUE)
# 内訳を持たない末端の品目特定
f.lv <- rep(0, length(items$Level))   # 末端フラグ
for (i in 1:(length(items$Level)-1)) if (items$Level[i] >= items$Level[i+1]) f.lv[i] <- 1
#
# 飲料関係の 20 項目一覧
(s.itm <- items[which(f.lv==1),][79:98,])
```

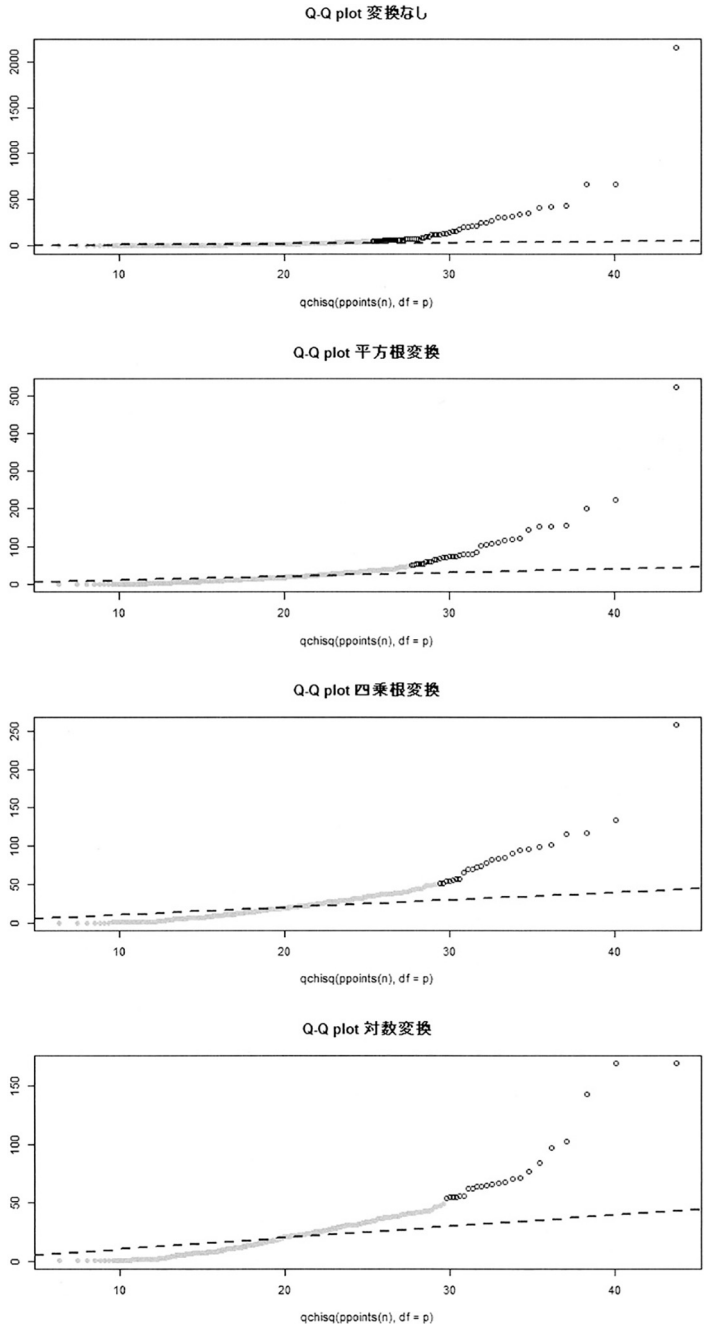

図 2. 一般用ミクロデータ(全国消費実態調査・飲料)QQ プロット.

```
#
# 世帯主が就業者で 30 歳未満 (T_Age_5s が 1) かつ世帯人員 2 人 (T_SeJinin が 2) の世帯
require(dplyr)
s1.dat <- dat |> filter(T_Age_5s == 1, T_SeJinin == 2) # レコード選択
d1 <- s1.dat[,which(f.lv==1)][,79:98]   # 項目選択
dim(d1)                # [1] 305  20
#
# 外れ値検出
(s1_time <- Sys.time())
ot1 <- RMSDp(d1)                  # 変換なし
(e1_time <- Sys.time())
#
(s2_time <- Sys.time())
ot2 <- RMSDp(sqrt(d1))           # 平方根変換
(e2_time <- Sys.time())
#
(s3_time <- Sys.time())
ot3 <- RMSDp(d1^(1/4))           # 四乗根変換
(e3_time <- Sys.time())
#
(s4_time <- Sys.time())
ot4 <- RMSDp(log10(d1))           # 常用対数変換
(e4_time <- Sys.time())
#
# [参考] 外れ値検出の進捗状況表示方法
# require(progress)  # プログレスバー表示
# pp <- 100      # 100% 刻み
# pb <- progress_bar$new(total = n)
# for (i in 1:pp) {
#    pb$tick()
#    ot1 <- RMSDp(d1)
#    Sys.sleep(1/pp)
# }
#
# 外れ値検出結果 (1:正常値; 2:外れ値)
table(ot1$ot)
table(ot2$ot)
table(ot3$ot)
table(ot4$ot)
#
# 並行座標プロット
require(MASS)
par(mfrow=c(4,1))
parcoord(d1, col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot1$ot])
parcoord(sqrt(d1), col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot2$ot])
parcoord(d1^(1/4), col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot3$ot])
parcoord(log10(d1), col=c(rgb(0,0,0, alpha=0.2), rgb(1,0,0, alpha=0.5))[ot4$ot])
#
# 散布図行列
pairs(d1, cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot1$ot])
pairs(sqrt(d1), cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot2$ot])
pairs(d1^(1/4), cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot3$ot])
pairs(log10(d1), cex=0.1, pch=20, col=c(rgb(0,0,0, alpha=0.2),
                rgb(1,0,0, alpha=0.3))[ot4$ot])
```

```
#
# QQ プロット
n <- nrow(d1);  p <- ncol(d1)
par(mfrow=c(2,2))
qqplot(qchisq(ppoints(n), df=p), ot1$mah, pch=19, col=ot1$ot[order(ot1$mah)],
        main="Q-Q plot 変換なし")
        abline(0, 1, col="green")
qqplot(qchisq(ppoints(n), df=p), ot2$mah, pch=19, col=ot2$ot[order(ot2$mah)],
   main="Q-Q plot 平方根変換")
        abline(0, 1, col="green")
qqplot(qchisq(ppoints(n), df=p), ot3$mah, pch=19, col=ot3$ot[order(ot3$mah)],
   main="Q-Q plot 四乗根変換")
        abline(0, 1, col="green")
qqplot(qchisq(ppoints(n), df=p), ot4$mah, pch=19, col=ot4$ot[order(ot4$mah)],
   main="Q-Q plot 対数変換")
        abline(0, 1, col="green")
#
# Box-Cox 変換のパラメータ推定
require(car)
(lambda1 <- powerTransform(d1))
#
# 外れ値判定の閾値を広げる
# F 検定統計量
FF <- ot2$mah * (n-p) * n /((n^2-1) * p)
ot2r <- rep(1, n)  # 外れ値フラグ

ot2r[which(FF > (ot2$fs*1.5))] <- 2  # 元の閾値を 1.5 倍に
table(ot2r)
```

4.3 処理時間や処理の限界について

シングルコア版は，すべての基底への射影計算を一度にメモリに展開するアルゴリズムのため，特に変数の数が増えればメモリの破綻を起こして処理不能となる．4GB メモリ搭載の 32bit 機の場合，大きさ 100 のデータセットによるシミュレーションで，11 変数が限界であった（Wada and Tsubaki, 2013），並列化版はデフォルト設定では同様の環境において処理変数を増やすために開発され，メモリ破綻を防止するために基底への射影計算を分割し，複数コアに割り振っている．

データ分割の最小単位はデータセットのサイズになるため，並列化版については理論上はメモリ展開不能なサイズのデータでない限り，時間はかかるが処理は可能である．大容量のメモリが利用可能な場合は，関数 RMSDp の引数 dv を大きくすることで処理速度の向上を図ることができる．

処理速度は，コア数が 1 桁台の汎用 PC であればシングルコア版が速いが，コア数の多いワークステーションあるいはスパコン等では並列化版がより高速となる．

コア数 4 で 16GB のメモリを搭載した 64bit のノート PC において，シングルコア版の RMSD パッケージの試算を行った結果は表 9 のとおり．併せて並列化版のテストも行い，4 コア全てを使用して 16 変数テストを行ったところ，7 時間半で正常終了した．17 変数のテストについては，フル充電の上電源も接続していたが，演算終了前にバッテリー切れで電源が落ちて計算不能となった．ただしこれはテスト機(表 10)のハードウェアに起因するものである．

5. おわりに

本稿では，破局点の高い楕円分布モデルに基づく外れ値検出法の一つである MSD 法のパッ

表 9．シングルコア版の限界（コア数 4・16GB メモリ機の場合）．

変数の数	処理時間	実行可能性
12	54 分	○
13	2 時間	○
14	4 時間半	○
15	10 時間半	○
16	30 時間	○
17	メモリ破綻	×

表 10．テスト機について．

シリーズ・モデル	Vaio Z, VJZ13A
プロセッサ	Intel Core^{TM} i7-5557U プロセッサ（TDP 28W）
メインメモリ	16GB
プラットフォーム	x86_64-w64-mingw32/x64 (64-bit)
R バージョン	4.3.2 (2023-10-31 ucrt)

ケージを紹介した．現在のところ，外れ値を判定する閾値の調整は F 検定統計量のパーセンタイル値のみであるが，今後より裾の重い分布に対応できる引数の追加など，より使いやすいものになるよう改善していきたい．

ここで紹介した MSD 法のシングルコア版は，2019 年個人企業経済調査から，相関のある経理項目の補完プロセスについて，最近隣ホットデック法の補完値候補となるドナーデータセットから他の大部分と傾向の違う特異な観測値を除外するために使用されている．補完プロセスはデータチェック後に行われ，外れ値検出対象となるデータは誤り等を含まないクリーンデータであるため，検出された観測値は，審査や修正を行うことなくそのまま集計に使用されるが，欠測の補完値として使用されない．この実用化例では，冒頭で紹介した多変量外れ値検出法の利用が進まない理由の三番目への対応として，複数の候補からベストな方法をシミュレーションにより選択している．さらに四番目の理由に関連して，クリーンデータへの適用であるため検出された外れ値について審査をする必要がない．

謝　辞

本研究について，統計数理研究所の椿広計所長には前職の筑波大学ビジネス科学研究科から長年に亘るご指導をいただき，心より感謝の意を表します．

参 考 文 献

Béguin, C. and Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data, https://www.cs.york.ac.uk/euredit/results/Results/Robust/Part%20C.zip（最終アクセス日 2024 年 3 月 5 日）.

Billor, N., Hadi, A. S. and Velleman, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators, *Computational Statistics & Data Analysis*, **34**(3), 279–298.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **26**(2), 211–243.

Campbell, N. (1989). Bushfire mapping using noaa avhrr data, Technical Report, Commonwealth Scientific and Industrial Research Organisations (CSIRO).

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators, Ph.D. Qualifying Paper, Department of Statistics, Harvard University, Boston.

Eurostat (2014). Statistics Explained [Glossary:Outlier], http://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Outlier（最終アクセス日 2024 年 3 月 5 日）.

Franklin, S. and Brodeur, M. (1997). A practical application of a robust multivariate outlier detection method, *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, 186–191.

Hawkins, D. M., Bradu, D. and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets, *Technometrics*, **26**, 197–208.

Maronna, R. A. and Yohai, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator, *Journal of the American Statistical Association*, **90**(429), 330–341.

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, **72**, 41–54.

Patak, Z. (1990). Robust principal component analysis via projection pursuit, Ph.D. Thesis, University of British Columbia.

Peña, D. and Prieto, F. J. (2001). Multivariate outlier detection and robust covariance matrix estimation, *Technometrics*, **43**(3), 286–300.

Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*, John Wiley & sons, New Jersey, USA.

Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, **41**(3), 212–223.

Stahel, W. A. (1981). Breakdown of covariance estimators, Research Report, No.31, Fachgruppe für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich.

Tyler, D. E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics, *The Annals of Statistics*, **22**(2), 1024–1044.

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, **67**, 89–157.

Wada, K. (2020). Outliers in official statistics, *Japanese Journal of Statistics and Data Science*, **3**(2), 669–691.

Wada, K. and Tsubaki, H. (2013). Parallel computation of modified Stahel-Donoho estimators for multivariate outlier detection, *2013 IEEE International Conference on Cloud Computing and Big Data*, 304–311.

Wada, K., Kawano, M. and Tsubaki, H. (2020). Comparison of multivariate outlier detection methods for nearly elliptical distributions, *Austrian Journal of Statistics*, **49**(2), 1–17.

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.

Dealing with Outliers in Official Statistics —R Packages Implementing the MSD Estimators—

Kazumi Wada

Statistical Research and Training Institute,
Ministry of Internal Affairs and Communications (MIC)

In the field of official statistics, the univariate method known as range checking is still the predominant method for detecting outliers in continuous values; however, the importance of dealing with multivariate outliers is gradually being recognized because the products of statistical surveys increasingly include individual data, in addition to traditional statistical tables. This paper explains the difference between univariate and multivariate outliers. It also introduces the R packages RMSD and RMSDp, which implement the modified Stahel-Donoho (MSD) estimators as a multivariate outlier detection method that assumes a unimodal symmetric elliptic distribution.

Key words: Data cleaning, elliptical distribution.