統計研究彙報　第 67 号　2010 年 3 月　（89～157）

多変量外れ値の検出　～MSD 法とその改良手法について～

和田かず美†

Detection of Multivariate Outliers　– Modified Stahel-Donoho Estimators –

WADA Kazumi

外れ値とは、データの大部分の傾向と異なるもので、必ずしも誤りとは限らないが、データ集計や分析の際にその存在が結果を歪(ゆが)めてしまう可能性がある。外れ値の存在は標本平均や標本標準偏差の算出に影響を与えるため、これらの値を用いた外れ値検出は検出漏れを起こす危険性が高い。

外れ値の影響を受けにくいロバストな単変量の外れ値検出法の代表が、順序統計量である四分位数を用いる箱ひげ図だが、単変量の手法は対象となる変数が他の変数と関係がある場合の外れ値検出には適さない。多変量データには、MSD（Modified Stahel-Donoho）法のような多変量でロバストな外れ値検出法が必要となる。

本稿は、統計調査の製表業務における多変量外れ値検出法の適用を目的として、1993 年からカナダの年次卸売・小売業調査（AWRTS）で実務に適用された MSD 法と、2001～2003 年に Eurostat が中心となってデータエディティング及び補定のための新手法の開発・評価を行った EUREDIT プロジェクトにおいて提案された MSD 法の改良版について比較評価し、実用化に向けた考察を行うものである。

キーワード：多変量外れ値検出、射影追跡法、MSD、Stahel-Donoho 法（SDE）

An outlier is a data point which has a different trend from the majority. Though it does not always mean error, its existence may distort the statistics such as the sample mean and the sample standard deviation. Those statistics should not be used as the estimators for the outlier detection since they are not robust.

The most commonly used robust method for the univariate data is the box plot which uses order statistics, however, it does not suitable for the correlated multivariate data. As for the multivariate data, the robust and multivariate method, such as Modified Stahel-Donoho (MSD) estimators, is necessary.

Statistics Canada has adopted the MSD method for the Annual Wholesale and Retail Trade Survey (AWRTS) since 1993. The EUREDIT project, funded by EUROSTAT, proposed a refinement of the method on its report in 2003.

This paper evaluates the MSD method of Statistics Canada and the refined version by EUREDIT aiming for implementation of the multivariate outlier detection for statistical survey data processing at National Statistics Center in Japan. Several topics for practical use are also considered.

Key words: Multivariate Outlier Detection, Projection Pursuit, MSD, SDE (Stahel-Donoho Estimator)

原稿受理日　平成 21 年 12 月 18 日

†　独立行政法人統計センター情報技術部研究主幹室

はじめに

独立行政法人統計センターにおける統計調査の製表業務では、量的変数に関しては、過去の調査結果などから想定される正常値の範囲を設定し、その範囲から外れる極端な値を外れ値として検出している。これは単変量の外れ値検出法であり、その量的変数が他の変数と密接な関係がある場合に、単変量で見た場合には極端な値ではないが他の変量との関係性に関して大多数のデータと傾向が異なる外れ値を検出することができない。このため、属性によりデータを細分化し、グループごとに正常値の範囲を調整することにより、擬似的にそのような多変量外れ値への対応を行っている。

一方、多変量外れ値検出法は、複数の量的変数を同時に取り扱い数学的な処理を行うことにより、極端な値をとる単変量の外れ値に加えて関係性についての外れ値を検出するものである。

統計調査の製表業務において、このような多変量外れ値検出法はこれまであまり普及していなかったが、コンピュータの処理能力が飛躍的に向上してこのような手法が実用に耐えるようになってきたことから、本稿では業務への適用可能性を探ることを目的に、カナダ統計局で実用化されている MSD 法とその改良手法について統計ソフト R によりプログラム開発を行い、比較評価を行うものである。

第Ⅰ章では様々な外れ値検出法を取り上げ、単変量の手法と多変量の手法の違いやなぜロバストでない手法を使ってはいけないのかを解説する。第Ⅱ章では、多変量でロバストな外れ値検出法である MSD 法とその改良手法について概説し、第Ⅲ章でシミュレーション及びデータテストによる比較評価の枠組みを示す。第Ⅳ章では比較評価の結果について述べ、第Ⅴ章でこのような外れ値検出法を統計調査データの製表業務に利用するための課題について考察している。

Ⅰ　多変量外れ値検出の必要性

外れ値検出法には、大きく分けて単変量でロバストではない手法、単変量でロバストな手法、多変量でロバストではない手法及び多変量でロバストな手法の４種類に分類することができる。それぞれの特徴は、以下のとおり。

表１ 様々な外れ値検出法

種類		手法の例
単変量の外れ値検出法		
	ロバストではない手法	標本平均と標本標準偏差による方法
	ロバストな手法	箱ひげ図（四分位数による方法）
多変量の外れ値検出法		
	ロバストではない手法	標本平均と標本分散・共分散行列から算出したマハラノビス平方距離による方法
	ロバストな手法	MSD 法などによりデータの中心とデータの広がりをロバストに推定しマハラノビス平方距離に相当する統計量を用いる方法

１．単変量でロバストではない手法：標本平均と標本標準偏差による方法

この種類の外れ値検出法で最も一般的なのが、標本平均をデータの中心とし、そこからおおむね標本標準偏差の３倍以上離れた値を外れ値とする経験的な判定方法である。

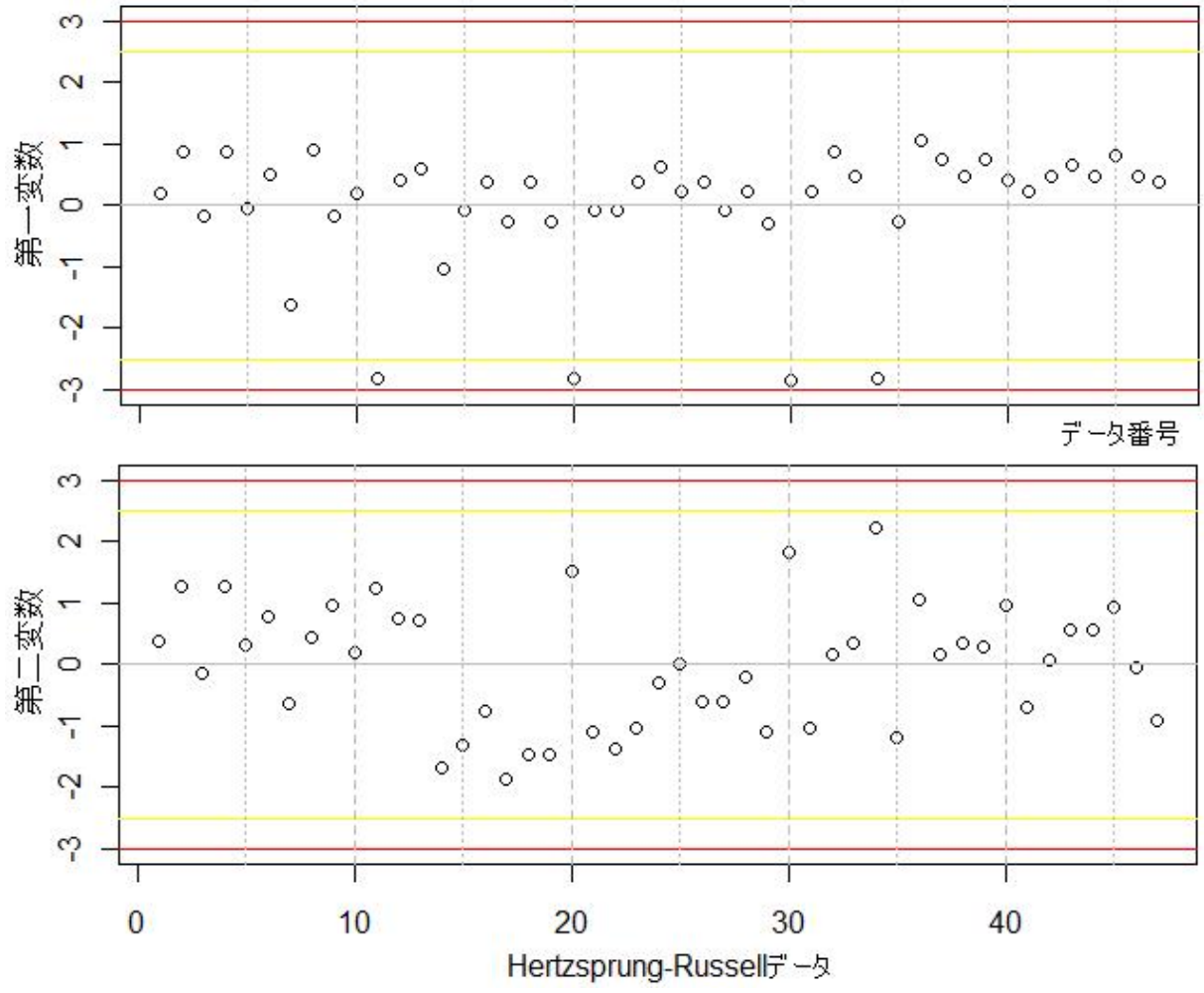
図 1.1.1 では、外れ値検出法のテスト用によく使用される Rousseeuw and Leroy（1987）の星の表面温度（対数）と光密度（対数）のデータセットである Hertzsprung-Russell データを、標準化して変数ごとにプロットしている。

データを標準化すれば平均が 0、標準偏差は 1 になる。データが正規分布に従うと仮定すると、赤線で示した値の-3 から 3 までの範囲に入る確率は 99.9973%であるが、この基準では外れ値は存在しないことになる。

ただし、標本平均はすべてのデータの値を用いて算出するため、少数でも極端な値が混入すれば大きな影響を受ける。標本標準偏差は標本平均よりも更に大きく外れ値の影響を受けるため、外れ値の検出漏れを起こしやすいことが知られている。

EU 域内の国の統計部局が中心となり 2006～2007 年に実施された EDIMBUS プロジェクトにおいて、部門横断的な企業調査データのエディティング及び補定に関する推奨実践マニュアル（Recommended Practices Manual :RPM）が作成されたが、このマニュアルにも外れ値検出に標本平均や標本標準偏差は使用すべきではないと明記されている［Istat et al.(2007)，小林(2009)］。

図 1.1.1 標本平均と標本標準偏差による外れ値検出



２．単変量でロバストな手法：箱ひげ図

箱ひげ図は、順序統計量である四分位数（四分位値）を用いた検出法である。例えば 100 個のデータがあるとき、データを昇順にソートすると、標本中央値は 50 番目と 51 番目のデータの平均で、標本第一四分位数は 25 番目と 26 番目のデータ平均になり、標本平均や標本標準偏差と異なりすべてのデータの値を用いるわけではないために外れ値の影響を受けにくいという性質がある。図 1.2.1 が箱ひげ図の概要で、ひげ先は内堀の中で最も外側にあるデータの位置を示している。

図 1.2.2 は、図 1.1.1 に示すように標本平均から標本標準偏差の３倍という基準では明らかな外れ値が検出されなかった Hertzsprung-Russell データを、箱ひげ図にプロットしたもの。箱ひげ図では、47 個のデータのうち第１変数で５個（データ番号 7，11，20，30，34）の外れ値が検出された。

図 1.2.1 外れ値と箱ひげ図［出典：吉澤（1992)］

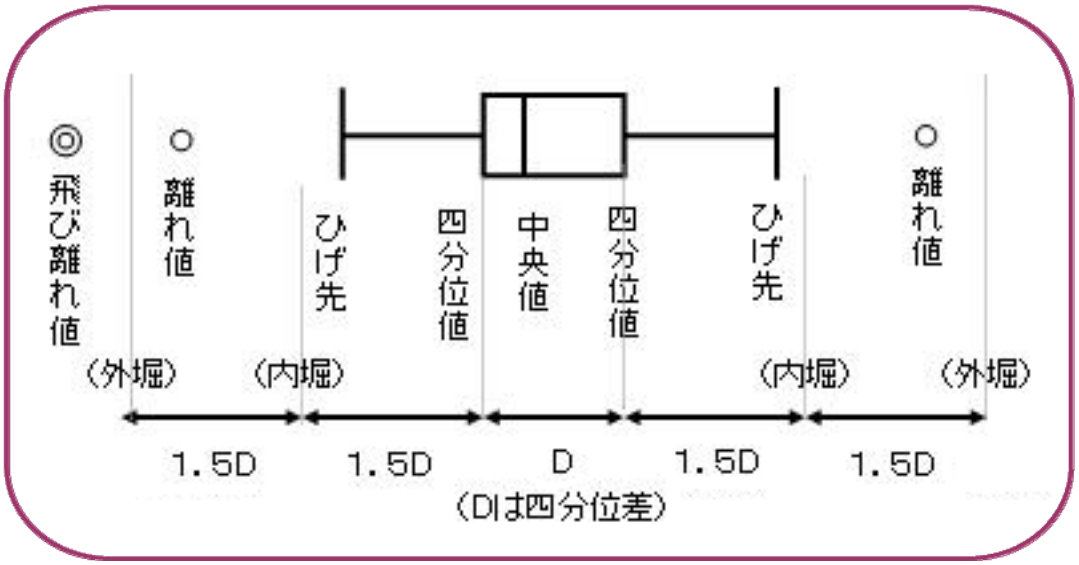
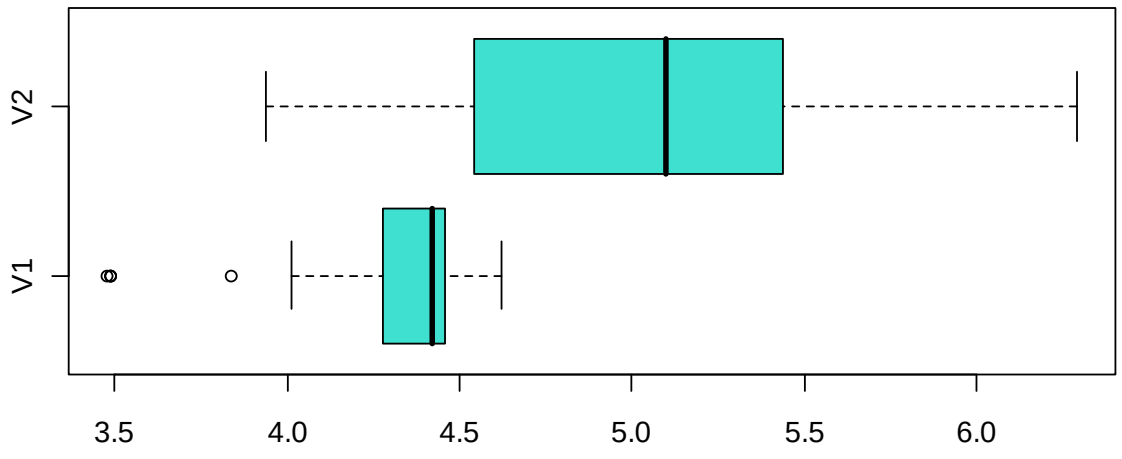


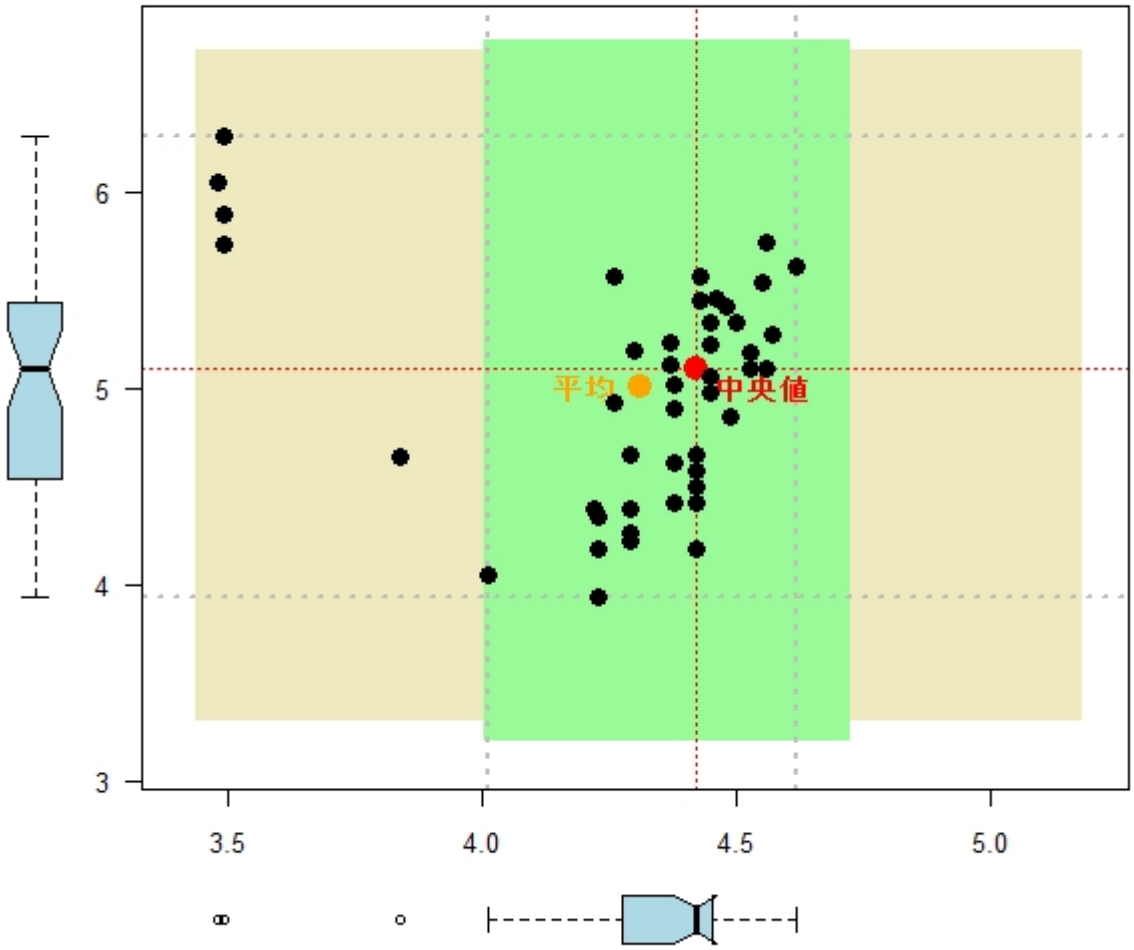
図 1.2.2 箱ひげ図：Hertzsprung-Russell データ［47×2 変数］



Hertzsprung-Russell データは、データポイントの固まりとそのクラスターから少し離れた幾つかのデータポイントで構成されている。このように正常値から離れたところにある外れ値が幾つもある場合、標本平均値も標本中央値もデータ中心を推計する指標であるが、標本平均値は標本中央値よりも外れ値の影響を受けて外れ値を除いたデータの中心より外れ値が多い方向にずれる傾向があり、またデータポイントの散らばりを示す指標である標本標準偏差も標本四分位差と比較すると外れ値を除いたデータよりも大きな値になりやすい。この影響により外れ値の検出漏れが起きやすく、この現象はマスキング効果と呼ばれる。

図 1.2.3 は、Hertzsprung-Russell データの標本平均を橙点、標本中央値を赤点、標本平均と標本標準偏差から計算される正常値の範囲をベージュの領域、標本中央値と標本四分位数により計算される正常値の範囲を緑の領域で示したもの。

図 1.2.3 散布図：Hertzsprung-Russell データ［47×2 変数］



３．多変量でロバストではない手法：標本平均と標本分散・共分散行列から算出したマハラノビス平方距離による方法

多変量データは、１つのデータが幾つもの変数を持つために、どのデータが外れているのか比較判定するための単一の指標を作ることが必要になるが、ユークリッド距離あるいはマハラノビス距離を算出することにより、多変量データから単変量の指標を作り出すことができる。

ユークリッド距離は分かりやすく計算も簡単で広く利用されているが、マハラノビス距離は変数間の相関の影響も考慮される点でユークリッド距離よりも優れている。変数相関が 0 であればマハラノビス距離と標準化ユークリッド距離は同じ値である。マハラノビス距離の二乗値であるマハラノビス平方距離による外れ値の判定方法は、以下のとおり。

データ数ｎ、変数の数ｐの多変量データについて、ｉ番目のデータを $x_i^T=(x_{i1},x_{i2},...,x_{ip})$ とする。$x_i,...,x_n$ が平均値ベクトルｕ、分散・共分散行列Ｖの多変量正規分布の母集団からのランダムな標本であると仮定し、各観測値についてまず(1)式によりマハラノビス平方距離 $D^2(x_i)$ を算出する。

$$D^2(x_i)=(x_i-u)^T V^{-1}(x_i-u) \qquad (1)$$

$D^2(x_i)$ の検定統計量 F は p 及び n-p の自由度を持つ F 分布に従い、(2)式により求めることができる。

$$F_i=\frac{(n-p)n}{(n^2-1)p}D^2(x_i) \qquad (2)$$

多変量データの例として、ここでは Campbell（1989）が使用した、山火事の痕跡を分析することを目的として衛星から測定された５変数の Bushfire（山火事）データを使用する。

比較のため、まず個々の変数ごとに箱ひげ図による外れ値検出を行うと、図 1.3.1 に示すように、第４・第５変数で９個（データ番号 8，9，32，33，34，35，36，37，38）の単変量外れ値が検出される。これらの単変量外れ値を示す散布図行列が図 1.3.2、平行座標プロットが図 1.3.3 である。いずれも箱ひげ図で検出された単変量外れ値を赤で表示している。

図 1.3.1 箱ひげ図：Bushfire データ［38×5 変数］

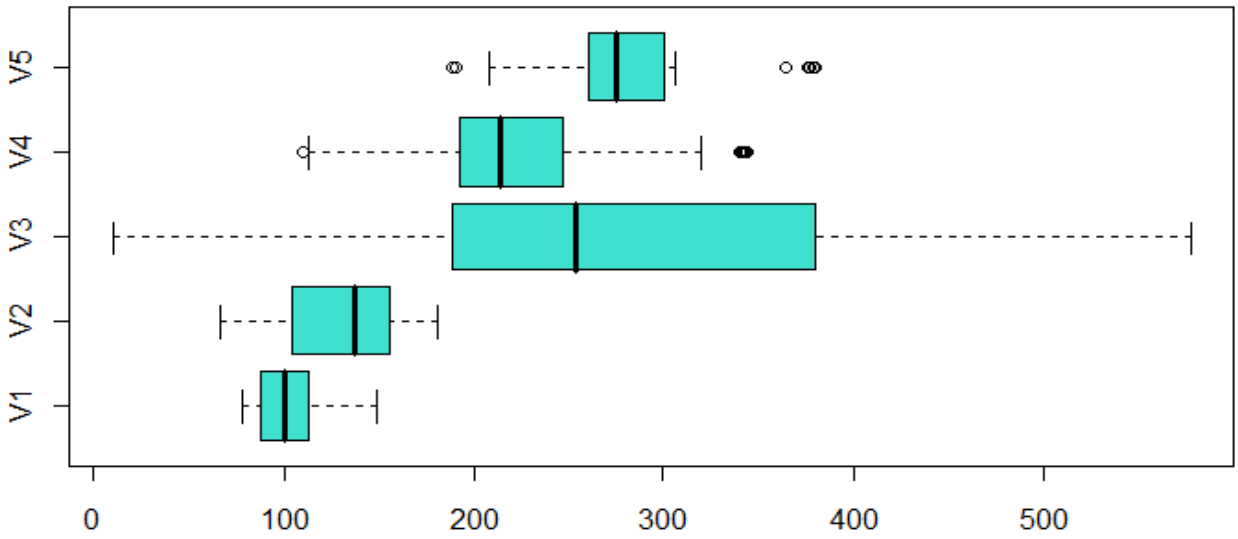


図 1. 3. 2 散布図行列：**Bushfire** データ［38×5 変数］

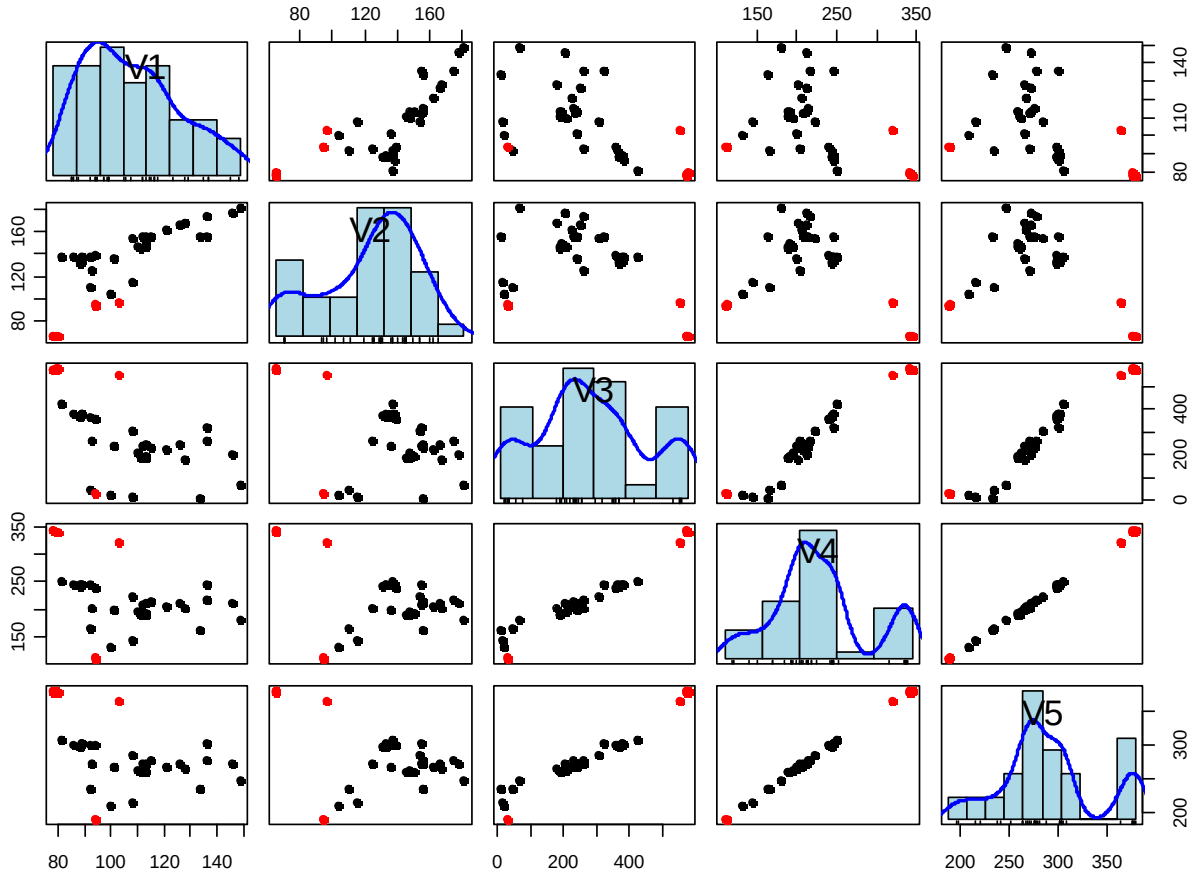
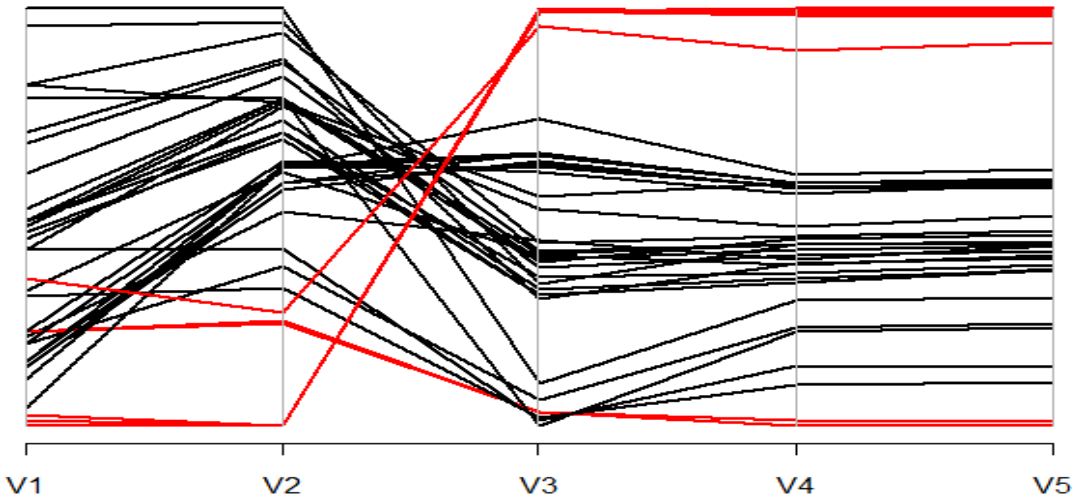
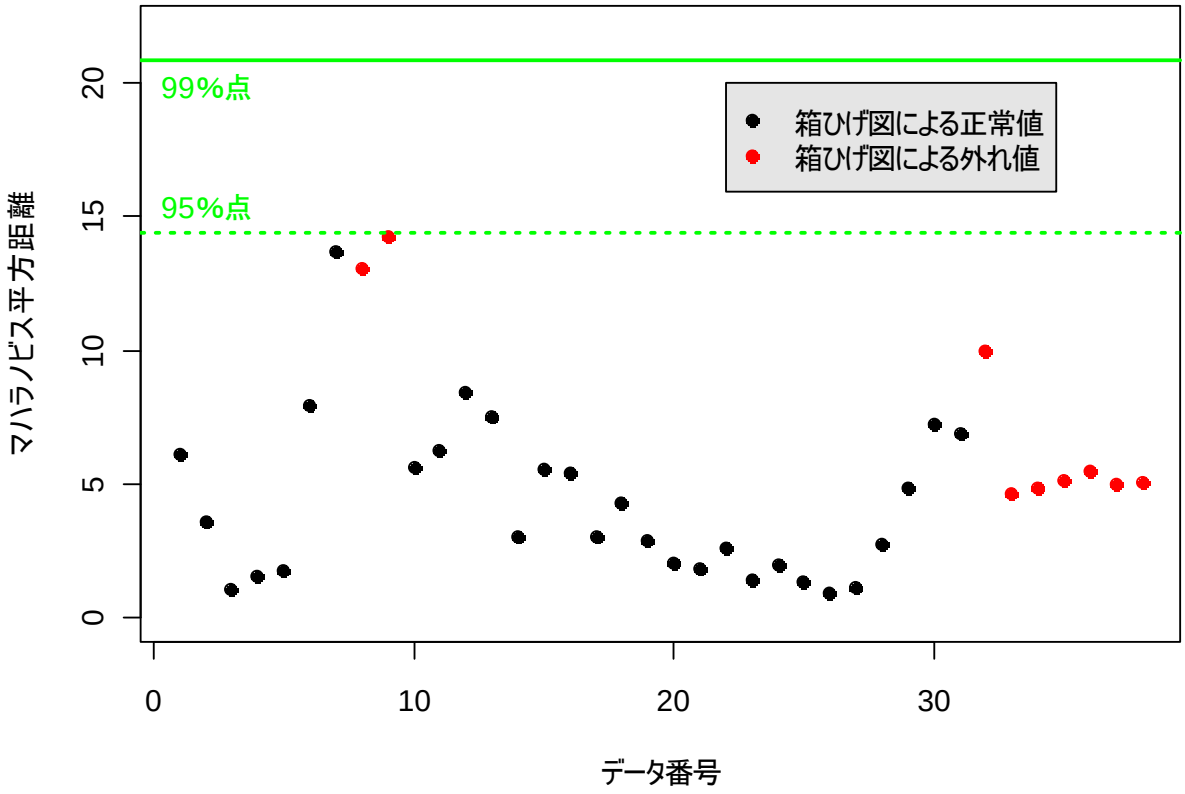


図 1. 3. 3 平行座標プロット：**Bushfire** データ［38×5 変数］



Bushfire データについて、算出したマハラノビス平方距離をプロットしたものが図 1. 3. 4 である。検定統計量 F の 95%値を点線、99%値を実線で表示しているが、95%値を基準としても、マスキング効果のために外れ値は 1 つも検出されない。

図 1.3.4 マハラノビス平方距離プロット：**Bushfire** データ［38×5 変数］



４．多変量でロバストな手法：MSD 法

多変量の検出手法を用いても、それがロバストなものでなければ、マスキング効果による外れ値の検出漏れが起こりやすい。このため、多変量でロバストな外れ値検出法が必要であり、近年様々な手法が提案されているが、ここでは第Ⅱ章で述べる Modified Stahel-Donoho（MSD）法を取り上げる。

図 1.4.1 は、Bushfire データについて、MSD 法により外れ値検出を行った結果を平行座標プロットで示したもの。赤線で示す箱ひげ図による外れ値は、個々の変数別で極端な値をとるデータである。一方、MSD 法で検出されるのは、箱ひげ図で検出される単変量外れ値に加え、１変数で見たときに必ずしも極端な値はとらないが変数間の関係性が他のデータの大部分と違う傾向を持つような緑線で示す外れ値であることが特徴である。

図 1.4.2 は、MSD 法によりロバストに推計されたマハラノビス平方距離のプロットで、箱ひげ図の外れ値と箱ひげ図で検出されない外れ値を同じように色分け表示している。

図 1. 4. 1 平行座標プロット：**Bushfire** データ［38×5 変数］

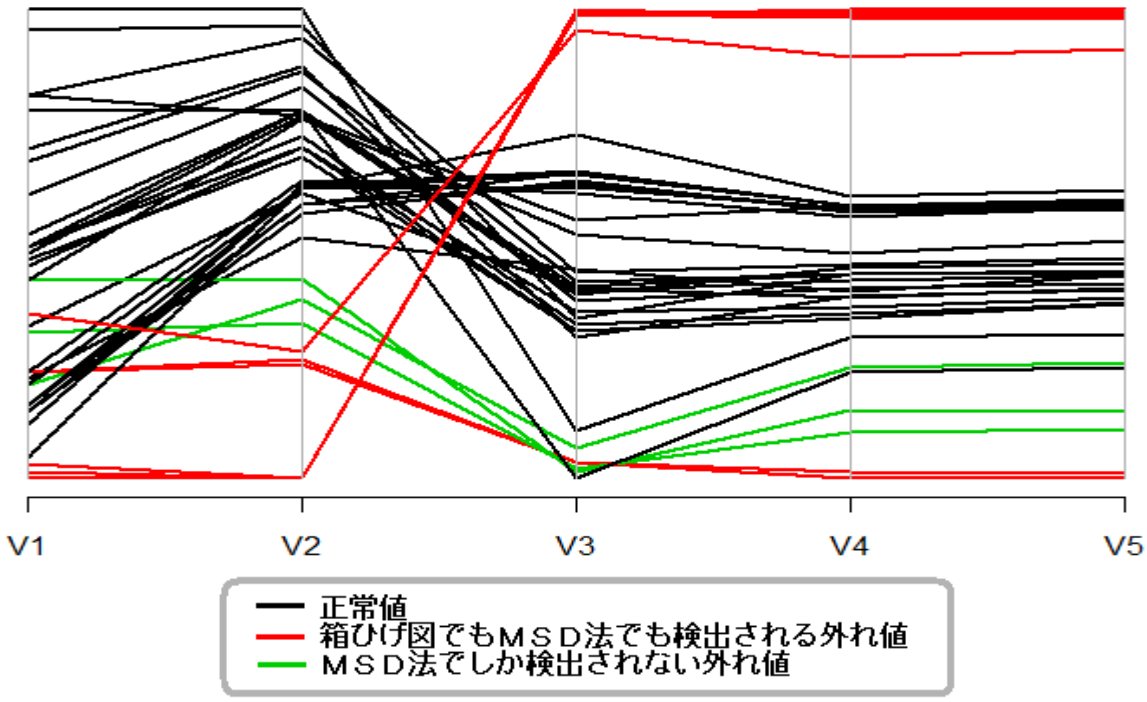
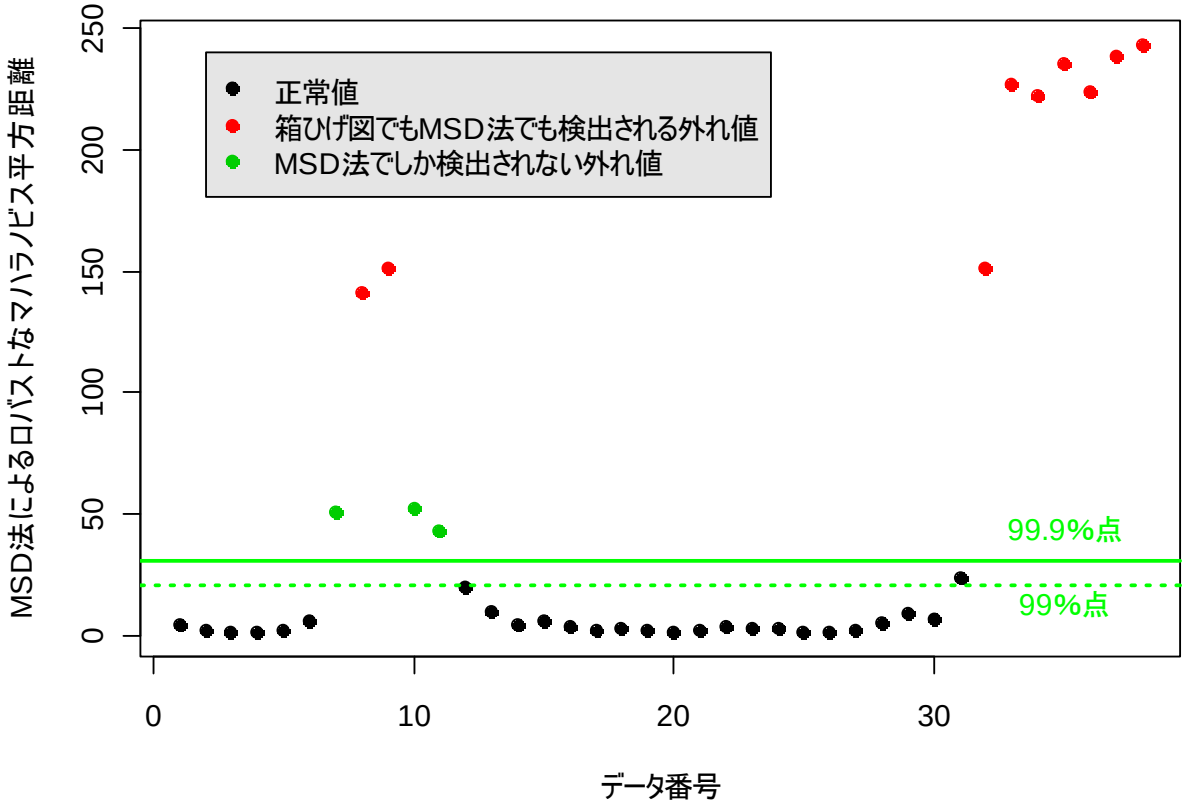


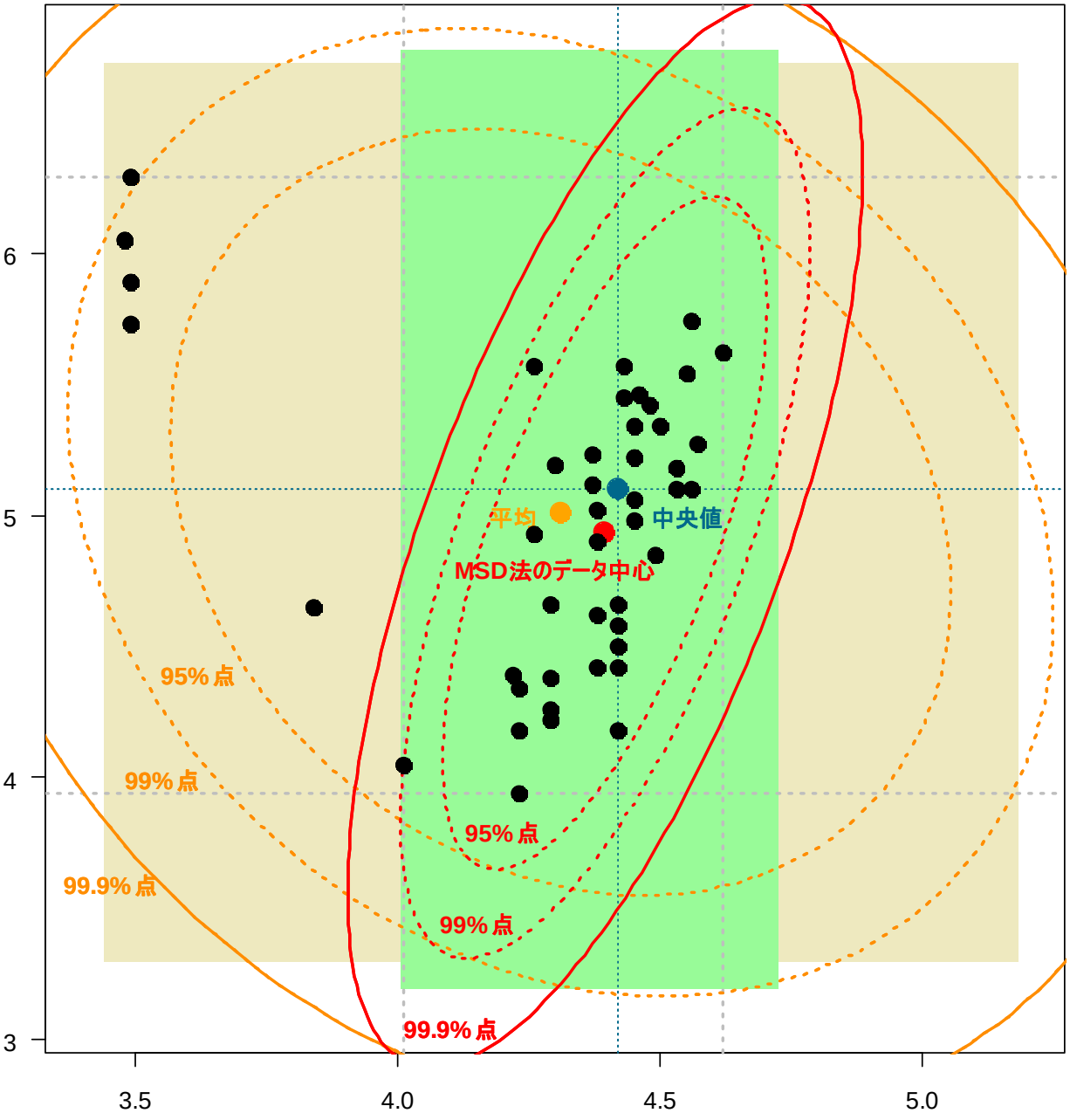
図 1. 4. 2 **MSD** 法によるマハラノビス平方距離プロット：**Bushfire** データ［38×5 変数］



また、単変量の手法について取り上げた第１節及び２節で使用した２変数の Hertzsprung-Russell データについて、多変量の手法を適用した結果を図 1. 4. 3 に示した。図 1. 2. 3 同様、標本平均と標本標準偏差から計算される正常値の範囲をベージュの領域、標本中央値と標本四分位数により計算される正常値の範囲を緑の領域で示し、さらにロバストではない通常の方法で推計したマハラノビス平方距離による正常値の範囲をオレンジ線、MSD 法によりロバストに推計したマハラノビス平方距離による正常値の範囲を赤線で示している。

２変数の場合、単変量の手法による正常値の範囲が矩形になるのに対し、多変量の手法による正常値の範囲は楕円形になる。単変量・多変量いずれの場合も、ロバストではない手法では外れ値が存在するとその影響で正常値と判定される領域が広がってしまう。

図 1. 4. 3 様々な外れ値検出法の比較：Hertzsprung-Russell データ［47×2 変数］



Ⅱ　Modified Stahel-Donoho（MSD）法とその改良手法について

１．基礎となる手法

Stahel（1981）及び Donoho（1982）が提案した SD 法は、データを様々な方向の直線上に射影し、その直線上における中心からの乖離度によって各データポイントにウエイトを付与することにより平均値ベクトルと分散・共分散行列をロバスト推定する、破壊点の高い、つまり多くの外れ値の混入に耐える手法である。

Patak（1990）は、SD 法によりロバスト推定された分散・共分散行列を用いた主成分分析を提案し、カナダ統計局はこれを用いて破壊点が約 0.5（理論的には 50%近い外れ値の混入に耐える。）と高く、直交変換不変な多変量外れ値検出法を実現した［Franklin and Brodeur（1997）］。さらに、EUREDIT プロジェクトで、カナダ統計局の手法の改良法が提案されている［Béguin and Hulliger（2003）］。

EUREDIT プロジェクト［http://www.cs.york.ac.uk/euredit/］は、Eurostat が資金提供し、データエディティング及び補定のための新手法の開発・評価を行うことなどを目的に、EU 域内の国の統計部局や大学の研究者たちが参加して 2001～2003 年に実施されたもの。

２．MSD 法の概要

まず、データを様々な方向の直線上に射影することにより、分布の端の方にある外れ値の候補を見付けて各データに一次ウエイトの付与を行う。分布の中心部に近いところにあるデータポイントのウエイトを１として、離れたところにあるデータポイントに１よりも小さいウエイトを付与することにより、このデータの影響を弱めたり、あるいはウエイトを 0 にして影響を排除したりすることができる。

この一次ウエイトを用いてデータの中心を示す平均値ベクトルと散らばりや相関を示す分散・共分散行列を計算することにより、外れ値の影響を受けにくい一次推計値を求める。

次に、こうして得られた分散・共分散行列を固有値分解することにより、主成分分析を行う。主成分分析自体はロバストな手法ではなく外れ値の影響を受けやすいが、データにウエイト付けして求めた分散・共分散行列を用いることによってロバストな分析が可能になる。

第一主成分は分散が最も大きくなるデータの射影方向を示すベクトルになるが、これは言わばデータの最大の類似成分を示す方向である。第二主成分は、データから第一主成分で表される成分を取り除いた後、同様に分散を最大化する方向を示すベクトルになり、第三主成分はデータから第一･第二主成分の要素を取り除いた後、同様に分散を最大化する方向を示すベクトルになる。つまり、第一主成分はデータの類似性を代表するが、第二主成分以降は類似性よりも非類似性が集約されていくため、外れ値検出に関して有用性が高い。このため、固有値の値にかかわらず算出されるすべての固有ベクトルを使用して再度射影を行い、ウエイトを精緻化することにより、更に精度の高い平均値ベクトルと分散・共分散行列の最終推計値を求める。

外れ値の判定には、この平均値ベクトルと分散・共分散行列の最終推計値を用いてマハラノビス平方距離を算出する。検定統計量 F を目安として、マハラノビス平方距離が大きなもの、つまりデータ中心から一定基準以上離れたところにあると判定されたデータポイントを外れ値として検出する。

以下は、Franklin and Brodeur（1997）に基づく MSD 法（カナダ版）及び Béguin and Hulliger（2003）

に基づく改良版（EUREDIT 版）との違いを検証するため、統計ソフト R で開発した MSD 法プログラムの処理概要である。

ソース及び実行コードは別紙１に示す。同じプログラムで制御パラメータによりカナダ版と EUREDIT 版の処理を行うことができる仕様になっている。

(1) データの中心化

最終的な分散・共分散行列と平均値ベクトルが、原点の取り方で結果が変わらないように、L_1 推定量を用いてデータの中心を原点に置く。

位置TのL_1 推定量：　　$\min_T \sum_{i=1}^{n} \| x_i - T \|$

通常、中心から各データポイントまでのユークリッド距離の二乗和が最小になるような推計量を用いるが、その場合、中心から遠いデータポイントほど中心の推定値に与える影響が大きくなる。これを避けるために、距離の絶対値の和が最小になるような推計値を使用している。

Béguin and Hulliger（2003）は、後の処理が原点不変なので中心化の必要はないと指摘しており、実際に検証を行って結果が変わらないことを確認した後、カナダ版についてもプログラムから該当ステップを削除した。

(2) 一次ウエイトの算出

一定の数のランダムな直交基底を作成し、基底を構成する各基底ベクトルが張る直線上にデータポイントを射影し、その線上で中心からの距離（残差）を求め、この残差の大きさにより一次ウエイトを設定する。

・ 直交基底の作成

変数の数を p とすると、p 個の要素を持つ直交ベクトル p 個で一組の直交基底を構成する。基底数を b とすると、まず b×p×p 個の一様乱数を作り、p 個の要素を持つベクトルを b×p 個発生させる。このベクトル p 個ごとにグラム・シュミットの直交化を行い、b 組の直交基底を作成する。

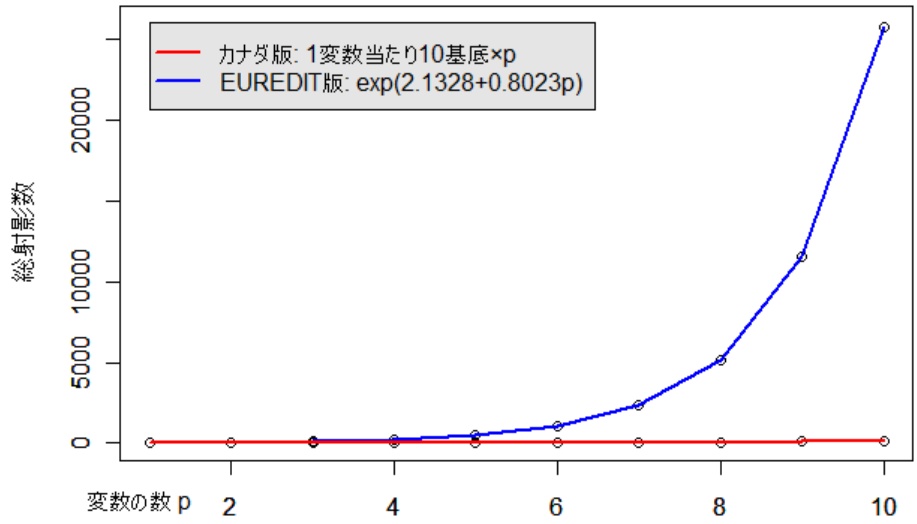
Franklin and Brodeur（1997）において１変数当たりの直交基底数は最低 10 としているため、カナダ版の基底数は１変数当たり 10 とした。

EUREDIT 版の基底数は、Maronna and Yohai（1995）に従い１変数当たり exp(2.1328+0.8023p)/p とした。このため基底数は次元数の増加に従い指数関数的に大きくなる。

表 2.2.1 基底数と射影数の違い

変数の数 p		2	3	4	5	6	7	8	9	10
1 変数当たりの基底数	カナダ版	10	10	10	10	10	10	10	10	10
	EUREDIT 版	41	93	208	466	1039	2319	5172	11539	25739
総射影数（基底ベクトル数）	カナダ版	20	30	40	50	60	70	80	90	100
	EUREDIT 版	82	279	832	2330	6234	16233	41376	103851	257390

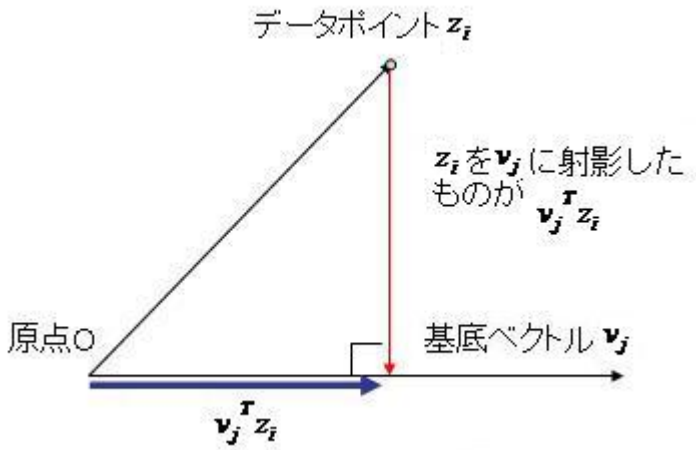
図 2.2.1 変数の数と総射影数



・　直交基底への射影

j 番目の基底ベクトル v_j （1≤j≤p）が張る直線上へ、各データポイント z_i を射影した（ベクトルに垂線を下ろした。）射影ベクトルの長さ $v_j^T z_i$ を計算する。

図 2.2.2 射影のイメージ図



・　残差の算出

射影ベクトルの長さ $v_j^T z_i$ を、標本中位数と標本中央絶対偏差によりロバストに標準化し、残差 r_{ij} を算出。正規分布を仮定すると、中央絶対偏差を 0.674 で割った値が標準偏差の推計値になる。中央絶対偏差とは、各データポイントから標本中位数を引いた値の中央値である。

$$r_{ij} = \frac{|v_j^T z_i - med(v_j^T z)|}{mad(v_j^T z)/0.674}$$

med：中位数（median）
mad：中央絶対偏差（median absolute deviation）

・　残差の刈り込み

カナダ版は、以下の基準で刈り込み残差$\tilde{r}_{ij}$を算出する。なお、元の分布が正規分布であると仮定した場合、刈り込み前残差r_{ij}は分母の符号を含めると平均 0、標準偏差 1 の正規分布に従うため、1.75 は 91.99%点、3.5 は 99.95%点に当たる。

$$\tilde{r}_{ij} = \begin{cases} r_{ij} & if\ r_{ij} \le 1.75 \\ 1.75 & if\ 1.75 < r_{ij} \le 3.5 \\ 0 & if\ \ 3.5 < r_{ij} \end{cases}$$

EUREDIT 版は、下式により変数の数 p によって刈り込み開始位置を変えて刈り込み残差$\tilde{r}_{ij}$を算出する。

$$\tilde{r}_{ij} = \begin{cases} r_{ij} & if\ 0 \le r_{ij} \le c \\ c^2 / r_{ij} & if\ c \le r_{ij} \end{cases} \quad \left(c = \sqrt{\chi^2_{p,0.95}} \right)$$

図 2.2.3　残差とウエイトの関係

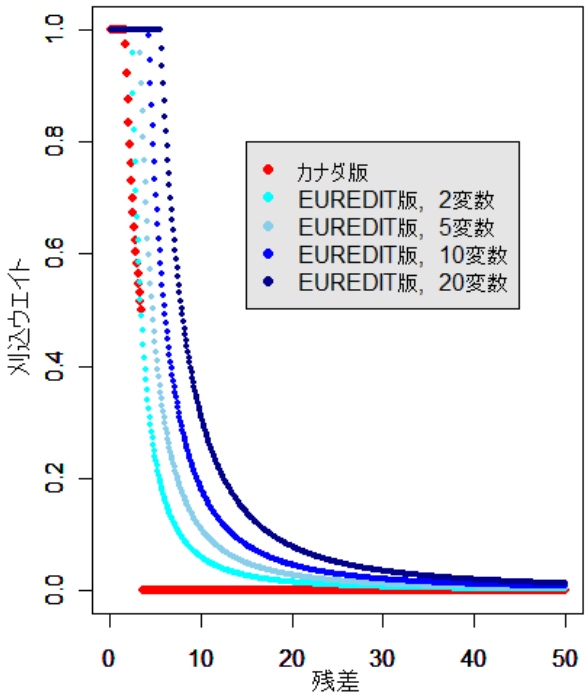
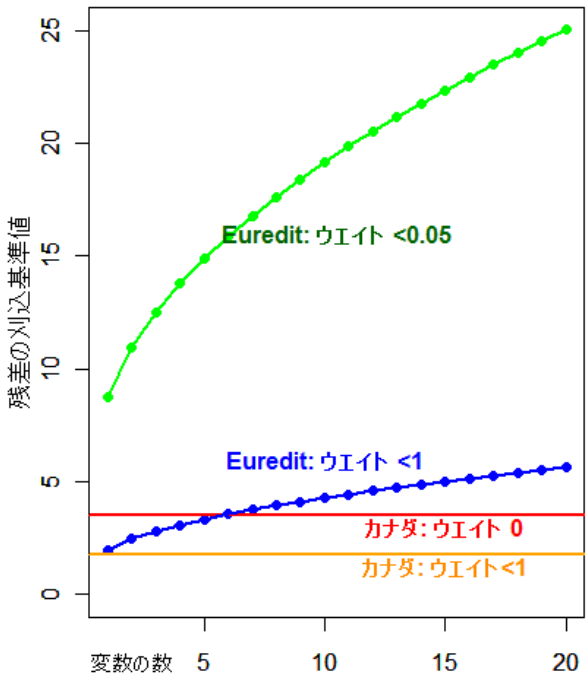


図 2.2.4　変数による刈り込み基準の変化



・　一次ウエイト算出

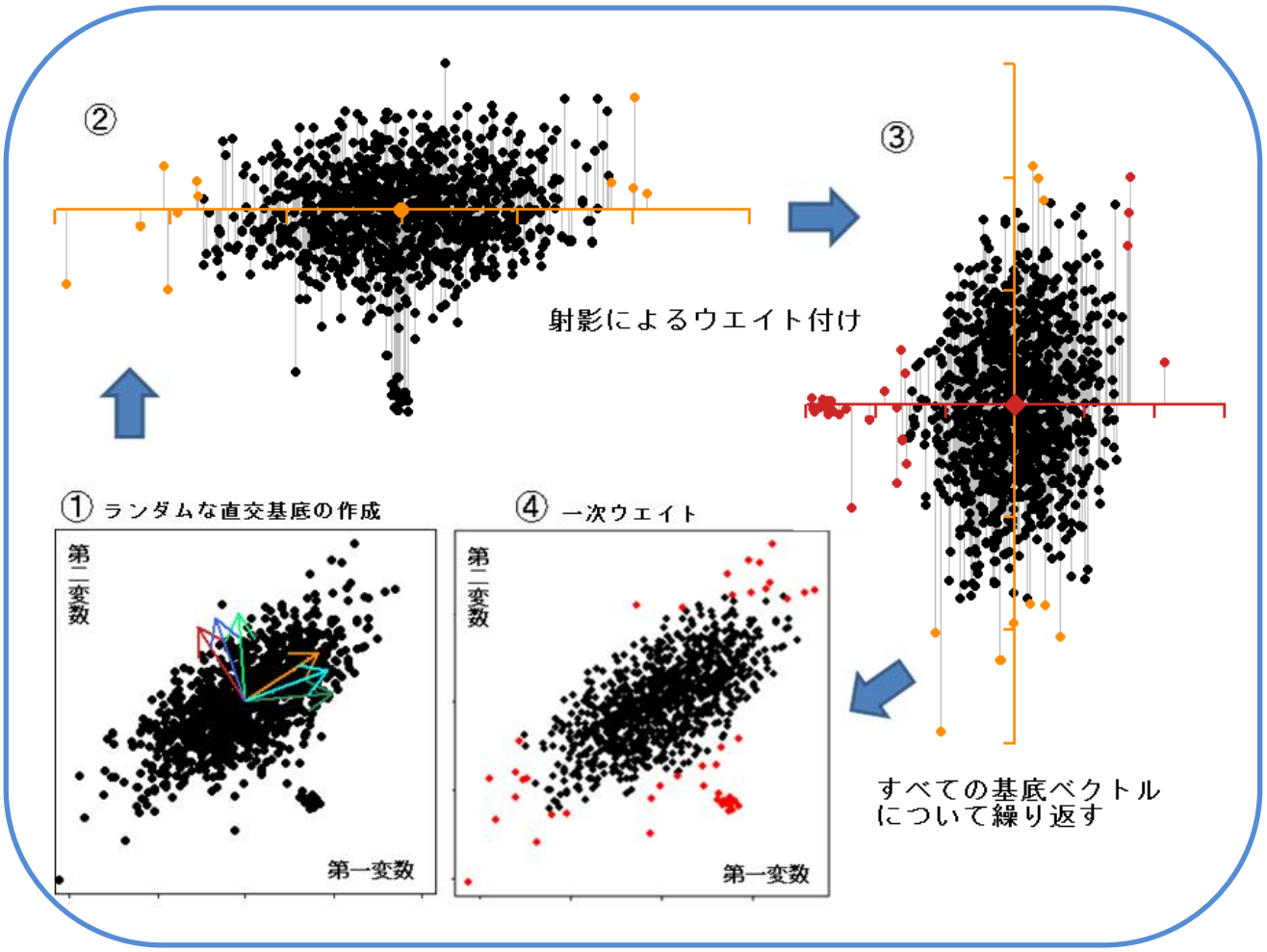
カナダ版、EUREDIT 版とも、下式のように刈り込み残差$\tilde{r}_{ij}$を刈り込み前残差r_{ij}で除して次元別ウエイトw_{ij}を算出する。これにより、刈り込まれないデータの中心部に近い位置にあるデータポイントのウエイトは 1 になるが、刈り込みの大きいデータの中心部から遠いデータポイントほどウエイトは小さくなる。

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

次に、下式のようにｂ組の直交基底ごとに次元別ウエイトの積和w_iを算出、さらに各データポイントごとに全基底を通じて最小のウエイトを選び、これを一次ウエイトとする。

$$w_i = \Pi_{j=1}^{p} \tilde{r}_{ij} / r_{ij}$$

図 2.2.5 ランダムな直交基底への射影による一次ウエイト算出



(3) ウエイト付き主成分分析

一次ウエイトを用いて平均値ベクトル$\hat{u}$と分散・共分散行列$\hat{V}$を推計し、得られた分散・共分散行列$\hat{V}$から固有値と固有ベクトルを求めることにより、ロバストな主成分分析を行う。

平均値ベクトル　$\hat{u} = \sum_i^n w_i x_i / \sum_i^n w_i$

分散・共分散行列　$\hat{V} = \sum_{i=1}^{n} (x_i - \hat{u})(x_i - \hat{u})^T w_i^2 / \sum_{i=1}^{n} w_i^2$

(4) 最終ウエイトの決定

主成分分析によりｐ変数データであればｐ個の要素を持つ固有ベクトルがｐ個算出されるが、これも先に作成したｂ組の直交基底と同様に一組の直交基底とみなすことができる。これらの

固有ベクトルにデータを射影し、残差の算出・標準化・刈り込み及び次元別ウエイトの積和算出により、二次ウエイトを作成する。

カナダ版の場合は、この二次ウエイトをそのまま最終ウエイトとして採用し、EUREDIT 版は一次ウエイトと二次ウエイトをデータポイントごとに比較して小さい方を最終ウエイトとする。

(5) マハラノビス平方距離の算出

最終ウエイトを用いて、再び分散・共分散行列と平均値ベクトルをロバスト推計し、これらの値から下式によりマハラノビス平方距離 $D^2(x_i)$ を算出する。

$$D^2(x_i) = (x_i - u)^T V^{-1} (x_i - u)$$

(6) 外れ値の特定

マハラノビス平方距離 $D^2(x_i)$ の検定統計量 F_i は、p 及び n-p の自由度を持つ F 分布に従い、下式により求めることができる。

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

外れ値と判定する検定統計量 F_i の基準は、Franklin and Brodeur（1997）に準じて、99.9％値とした。

Ⅲ　シミュレーションとデータテスト

１．比較方法

カナダの年次卸売・小売業調査（AWRTS）のデータ・エディティング業務に適用された MSD 法（カナダ版）と、EUREDIT プロジェクトで提案された MSD 法の改良版（EUREDIT 版）について、比較評価を行う。

これらの２つの版の詳細については第Ⅱ章に述べたが、相違点は射影のための１変数当たりの基底数及び射影の残差によるウエイトの付け方である。それぞれの効果を確認するために、この２点の順列組合せとなる以下の４通りの比較条件を設定し、シミュレーション及びデータテストを行った。

表 3.1.1 比較条件

	基底数	ウエイト付け
カナダ版	カナダ版	カナダ版
カナダ基底増加版	EUREDIT 版	カナダ版
EUREDIT 版	EUREDIT 版	EUREDIT 版
EUREDIT 基底減少版	カナダ版	EUREDIT 版

２．シミュレーション

Maronna and Yohai（1995）及び Peña and Prieto（2001）が、外れ値検出の難しいシミュレーションデータとして、ロバストな多変量外れ値検出法の評価に下式の形のデータを使用している。α が外れ値割合、p が変数の数、λ が外れ値の分散、δ が正常値からの外れ値の距離であり、正常値はデータ総数×(1－α)個の原点中心で分散・共分散行列 I の p 次元正規分布乱数で、外れ値はデータ総数×α 個、平均 0、分散 λ の正規分布に従う乱数を第一軸だけ原点から距離 δ だけ離して加えたものになる。

$$(1-\alpha)N_p(0, I)+\alpha N_p(\delta e_1, \lambda I)$$

このデータは、正常値も外れ値もそれぞれ正規分布に従うが、実際の統計調査データの場合は分布がきれいな形になるとは限らない。歪(ゆが)んだ分布あるいは裾(すそ)の厚い分布にどの程度対応できるかを確認するため、本研究では正常値には正規分布に加えて Skew-T 分布、複合ポワソン分布及び対数正規分布に従う擬似乱数データを使用した。図 3.2.1 及び 3.2.2 は、これらの分布の密度関数をプロットしたもの。各パラメータの値などのデータ設計の詳細は別紙２に示した。

３．データテスト

データテストには、シミュレーションデータよりも実データに近く、多変量外れ値検出法の性能評価によく使用されるデータセットを中心に選択した。詳細は別紙３のとおり。

図 3.2.1 正規分布及び Skew-T 分布の密度関数

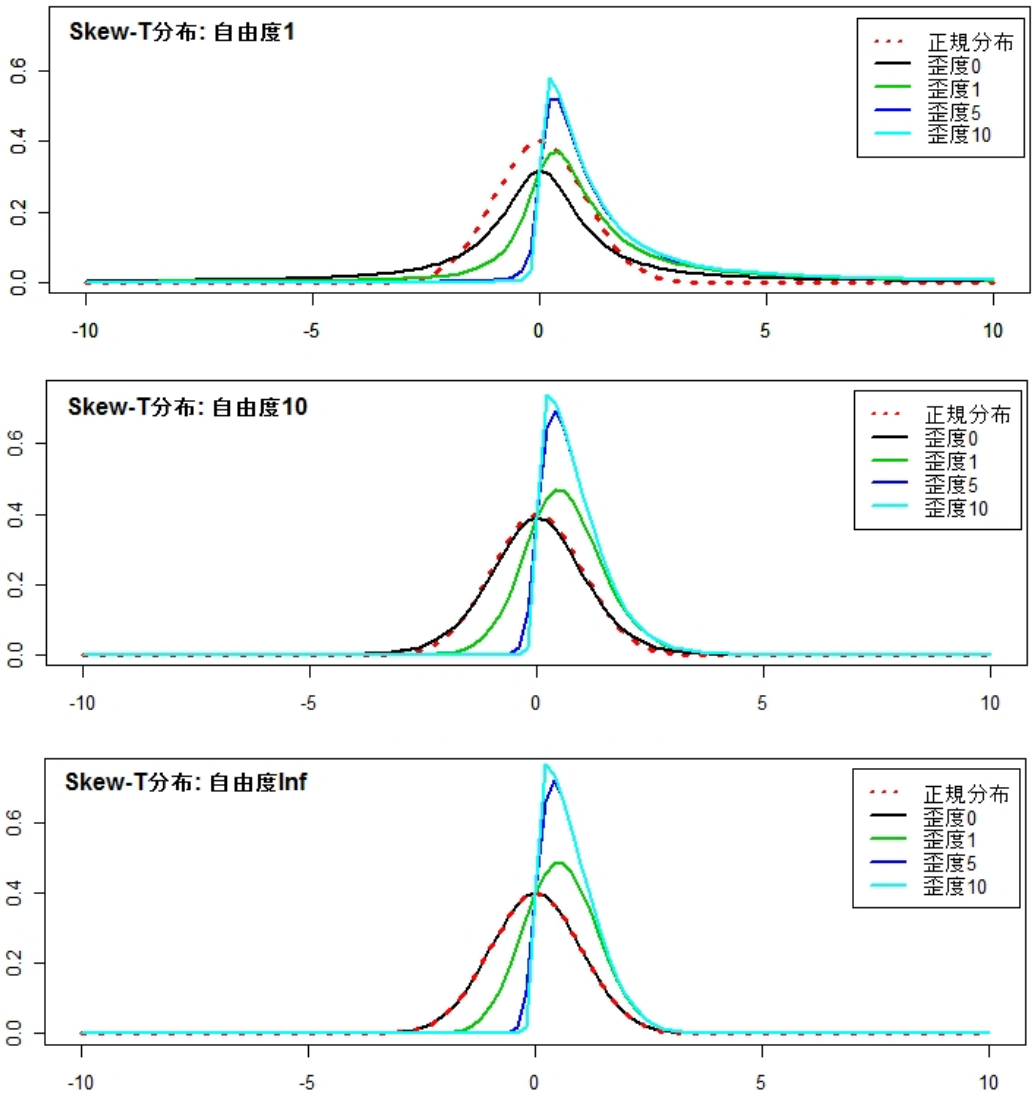
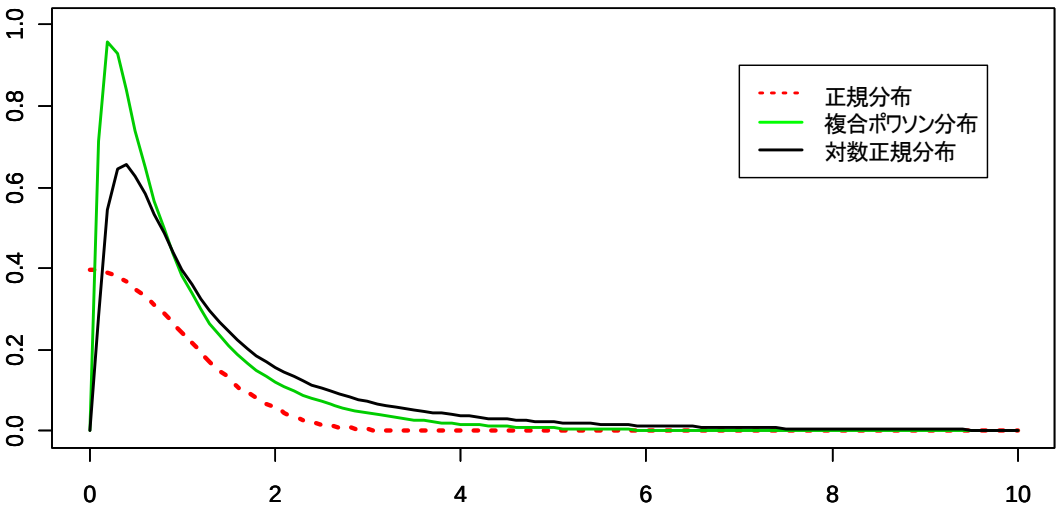


図 3.2.2 正規分布、複合ポワソン分布及び対数正規分布の密度関数



Ⅳ　結果

１．シミュレーションの結果

MSD 法の仕組み上は射影数が増えるほど検出率が高くなるはずだが、今回行ったシミュレーションでは、射影数の違いによる検出率の差は明確には見られない。これは、今回作成した擬似乱数によるシミュレーションデータが、射影数による検出力の差が明らかに出るほど十分に難易度が高いものではなかったためと考えられる。

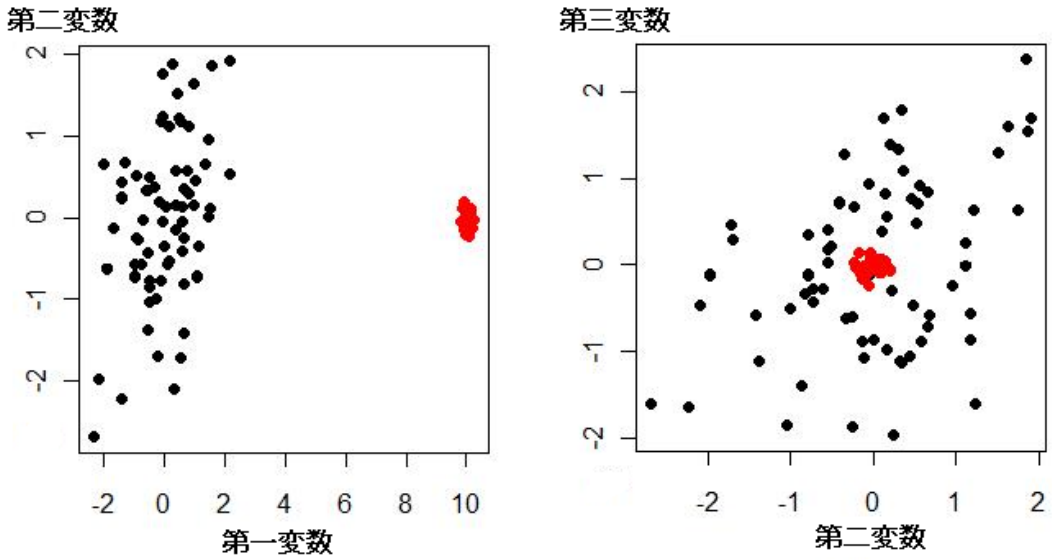
ウエイト付けについても検出率に明確な差異が現れていないが、カナダ版及びカナダ基底増加版については、途中で計算が不能になり外れ値検出ができない場合がある。計算不能となるデータ条件は表 4.1.1 のとおりで、10 変数で変数間相関が高くなく、標準偏差の小さい外れ値が 40%と多く、正常値と外れ値の間にデータポイントが全くない空間ができる図 4.1.1 のような状況で起きやすい。このとき、ほぼすべてのデータポイントの最終ウエイトが 0 か 0 に近い値となり、最終分散・共分散行列が正定値ではなくなる。マハラノビス平方距離算出には分散・共分散行列の逆行列が必要なため、外れ値検出ができない。計算が破綻しないという点では、EUREDIT 版のウエイト関数がカナダ版より優れているといえる。

表 4.1.1 外れ値が検出不能になるとき

条件	正常値の分布	自由度	相関	歪み	外れ値標準偏差	距離	変数	外れ値割合
カナダ基底増加版	正規分布	-	0	-	0.1	100	10	40
カナダ基底増加版	正規分布	-	0.4	-	0.1	100	10	40
カナダ基底増加版	対数正規分布	-	0	-	0.1	100	10	40
カナダ版･カナダ基底増加版	Skew-T 分布	10	0	5	0.1	10	10	40
カナダ版･カナダ基底増加版	Skew-T 分布	10	0	10	0.1	10	10	40
カナダ版･カナダ基底増加版	Skew-T 分布	Inf	0	1	0.1	10	10	40
カナダ版･カナダ基底増加版	Skew-T 分布	Inf	0.4	5	0.1	10	10	40
カナダ基底増加版	Skew-T 分布	10	0	1	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	10	0	5	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	10	0	10	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	Inf	0	0	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	Inf	0	1	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	Inf	0	10	0.1	10	10	40
カナダ基底増加版	Skew-T 分布	Inf	0	5	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	Inf	0	10	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	Inf	0.4	0	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	Inf	0.4	1	0.1	100	10	40
カナダ基底増加版	Skew-T 分布	Inf	0.4	5	0.1	100	10	40
カナダ版	Skew-T 分布	10	0.4	5	0.1	10	10	40
カナダ版	Skew-T 分布	Inf	0	5	0.1	10	10	40
カナダ版	Skew-T 分布	Inf	0.4	10	0.1	10	10	40

シミュレーションの結果の詳細は、正規分布データについて別表１、Skew-T 分布データについて別表２、複合ポワソン分布データについて別表３、対数正規分布データについて別表４に示した。これらの別表において計算不能が起きた箇所は、誤検出率・漏れ率・検出率すべてに「－」と表記している。

図 4.1.1 外れ値が検出不能になるシミュレーションデータ例



２．データテストの結果

(1) Hawkins-Bradu-Kaas データ

この３変数データの三次元プロットを図 4.2.1 に示す。図 4.2.2 は、３つの変数のうち特徴が分かりやすい第２変数と第３変数をプロットしている。どれが外れ値かは目視で明らかだが、ロバストではないマハラノビス平方距離で外れ値検出を試みると、検定統計量 F の 95%値を基準とした確率楕円（図 4.2.2 の緑の点線の一番内側）でも、赤で示す２つの外れ値しか検出できない。緑で示すデータ中心は実際よりもだいぶ外れ値側に寄り、外れ値の影響で分散も大きくなるために正常値範囲を示す楕円が広がってしまうのがその原因である。

一方 MSD 法については、図 4.2.2 で、カナダ版（赤）と EUREDIT 版（青）の違いはほとんどなく、両者とも正常値の平均値ベクトルと分散・共分散行列を推計できていることが分かる。

(2) スイスのレストラン業データ

２変数データなので、散布図にロバストではないマハラノビス平方距離、カナダ版及び EUREDIT 版による確率楕円を描いたものを図 4.2.3 に示す。検出される外れ値には若干の違いが出るが、実質的にはほとんど手法による差はないとみなすことができる。

外れ値が正常値と近接しており、正常値と比較して外れ値の数が少ない場合は、ロバストではない手法でも平均値ベクトルや分散・共分散行列の推計値の偏りがあまり大きくならない。

図 4.2.1　Hawkins-Bradu-Kaas データ
３Ｄプロット図

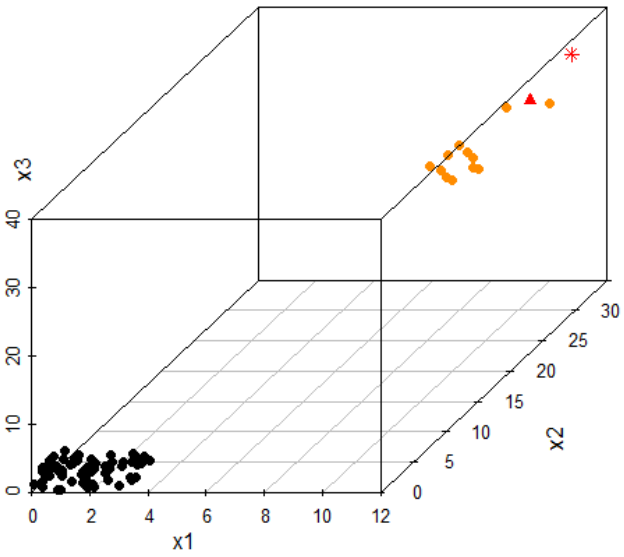


図 4.2.2　Hawkins-Bradu-Kaas データ
マハラノビス距離楕円

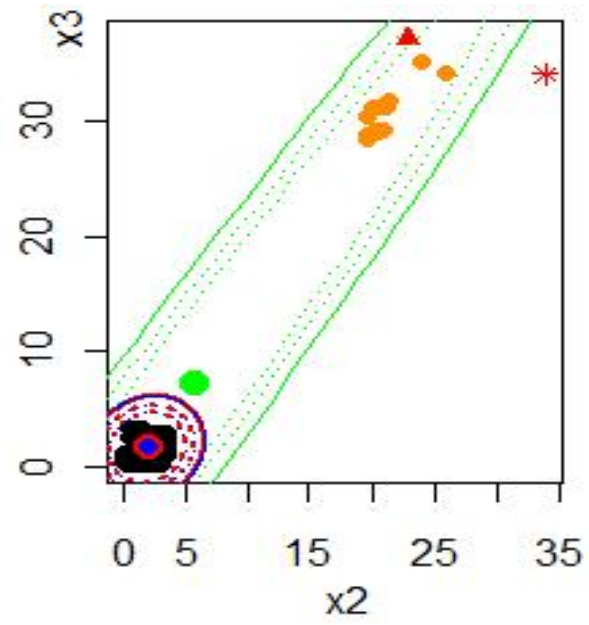
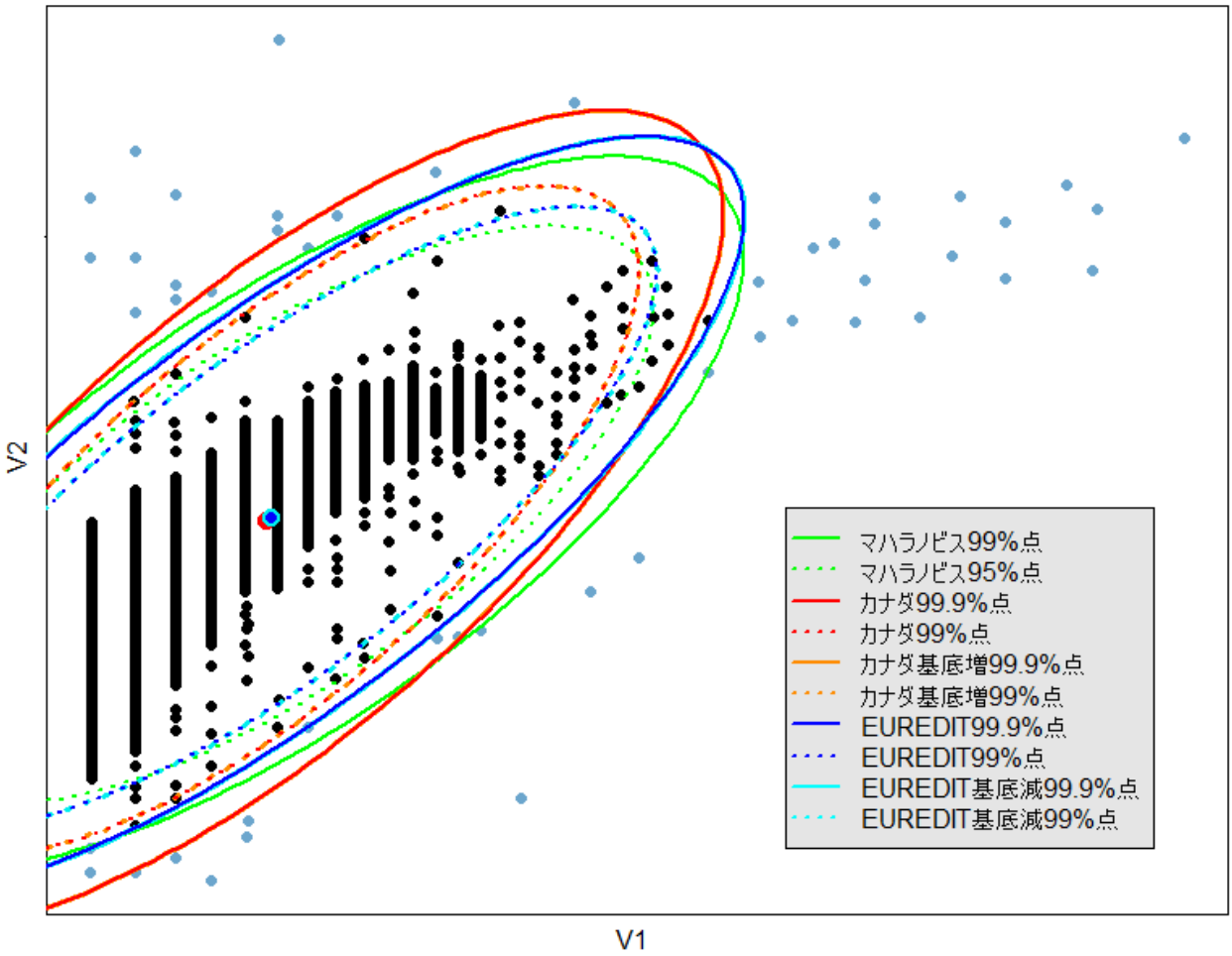


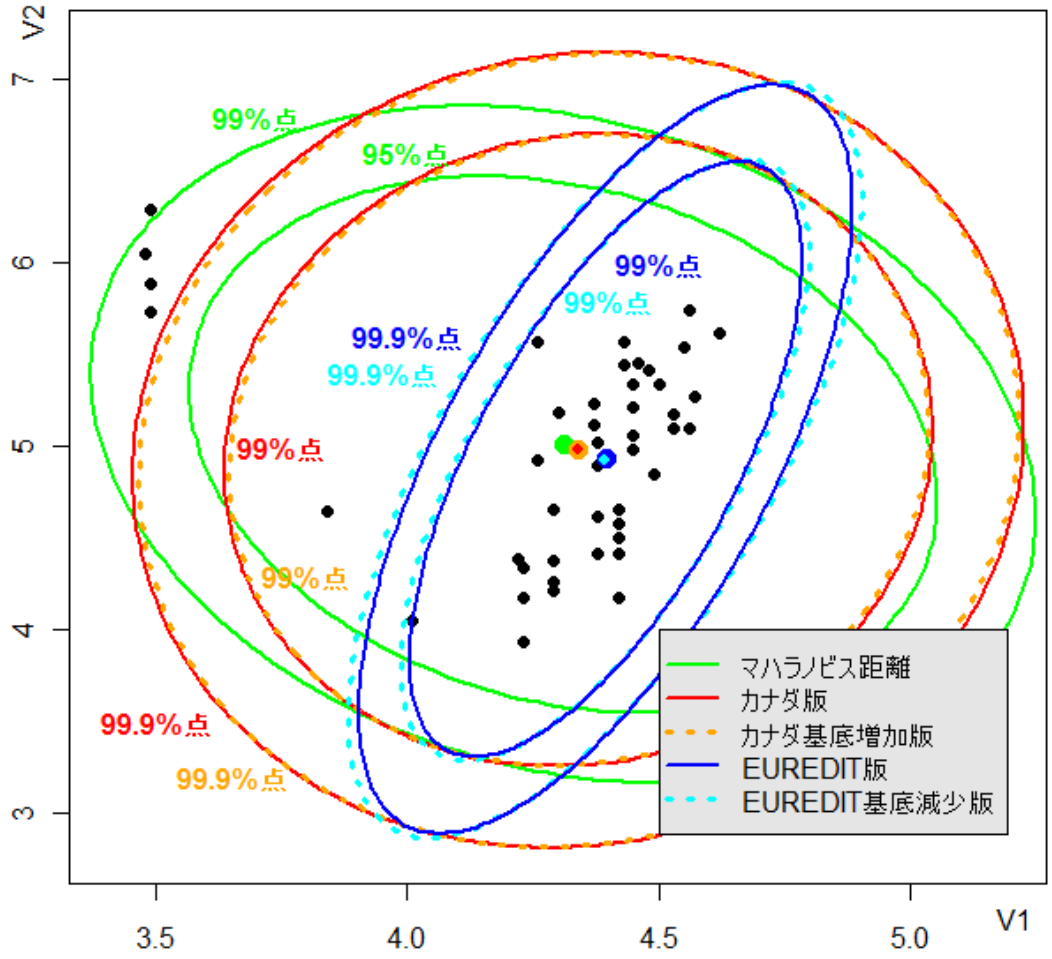
図 4.2.3　スイスのレストラン業データ散布図



(3) Hertzsprung-Russell データ

第Ⅰ章でも取り上げた２変数のデータセットである。散布図に各手法による確率楕円をプロットしたものを図 4.2.4 に示す。射影数の多寡による差異はほとんどないが、ウエイト付けの違いにより確率楕円の形が大きく変化しており、ロバストではない通常のマハラノビス距離やカナダ版よりも、EUREDIT 版の確率楕円の方が正常値の分布の形に沿うことが分かる。

図 4.2.4　Hertzsprung-Russell データ散布図



(4) Bushfire（山火事）データ

第Ⅰ章でも取り上げた５変数データ。複数回検出を行って結果が安定しているのは EUREDIT 版のみであった。

外れ値を色分けした散布図行列を図 4.2.5 に示す。EUREDIT 版が検出する 12 個の外れ値のうち、箱ひげ図で検出される単変量の外れ値を赤、箱ひげ図では検出できないが MSD 法で検出できる３個の外れ値を緑で表示している。図 4.2.6 は、同じものを平行座標プロットで示しているが、図 1.4.1 と同じものである。

図 4. 2. 5　**Bushfire** データ散布図行列

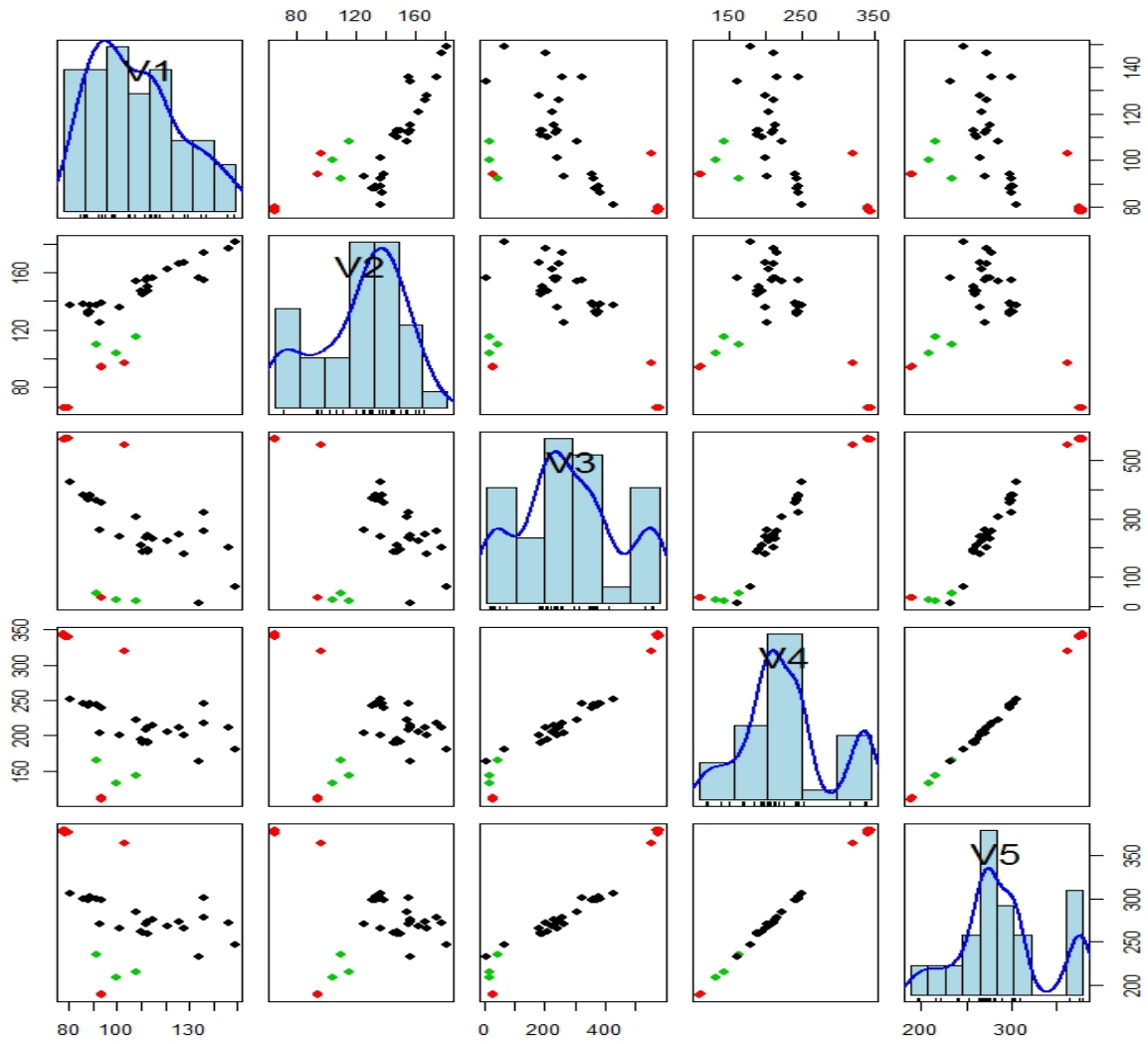
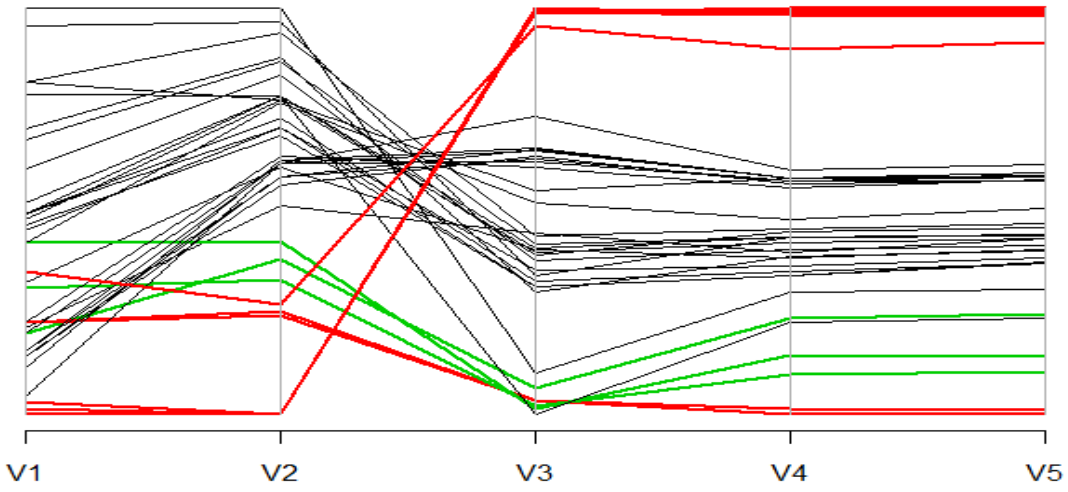


図 4. 2. 6　**Bushfire** データ平行座標プロット



(5) Stackloss データ

このデータセットの外れ値は、データ番号 No.1，2，3，4，21 の５つであることが分かっている。

乱数を変えて５回ずつ検出を行ったが、外れ値判定基準を検定統計量 F の 99.9%値とすると、どの条件でも結果が安定しない。

このため、基準を 99%値にしたところ、EUREDIT 版だけが５回すべてで正しい外れ値を検出できた。表 4.2.1 は、５つの外れ値についての検出結果を示しており、0 が正常値、1 が外れ値という判定である。この５つのデータポイント以外は、すべて正常値と判定される。

図 4.2.7 はこのデータの散布図行列で、表 4.2.1 内で青い四角で囲んだ部分の試行の際に、99.9%基準で検出された外れ値である No.1 と 21 を赤で、99%基準で検出された外れ値である No.2，3，4 を橙で表示している。図 4.2.8 と 4.2.9 は、同じ試行の際の D-D プロットと Q-Q プロットである。D-D プロットも Q-Q プロットも、データが同じ分布に従っていれば、緑で示す原点を通る傾き 1 の直線にすべてのデータポイントが乗る。

図 4.2.7 Stackloss データ散布図行列

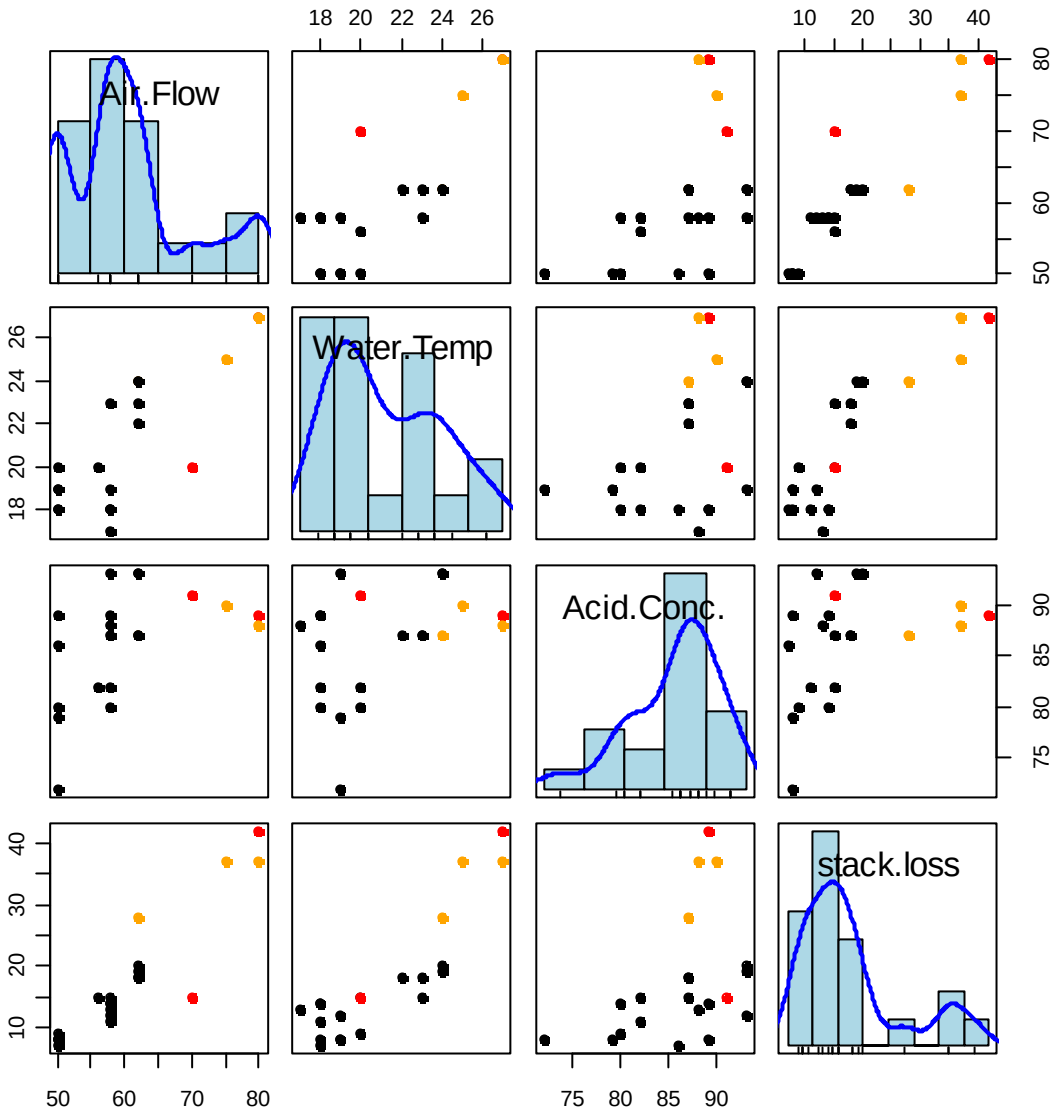


表 4. 2. 1 Stackloss データ検出結果

データ番号	99.9%基準					99%基準				
	No.1	No.2	No.3	No.4	No.21	No.1	No.2	No.3	No.4	No.21
カナダ版	1	0	1	1	0	1	1	1	1	1
	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
	1	0	1	1	1	1	0	1	1	1
	0	0	0	0	0	0	0	0	0	0
カナダ基底増加版	0	0	0	0	0	1	0	0	0	0
	1	0	1	1	1	1	1	1	1	1
	0	0	0	0	0	1	0	0	0	0
	0	0	0	0	0	1	0	0	0	0
	0	0	0	0	0	1	0	0	0	0
EUREDIT 基底減少版	1	0	1	0	0	1	1	1	1	0
	1	0	1	0	0	1	1	1	1	1
	0	0	0	0	0	1	0	0	0	0
	0	0	0	0	0	1	1	0	0	1
	0	0	0	0	0	0	0	0	0	0
EUREDIT 版	1	0	1	0	0	1	1	1	1	1
	1	0	1	0	0	1	1	1	1	1
	1	0	1	0	1	1	1	1	1	1
	1	0	0	0	1	1	1	1	1	1
	1	0	0	0	1	1	1	1	1	1

図 4. 2. 8 D-D プロット
［Stackloss データ］

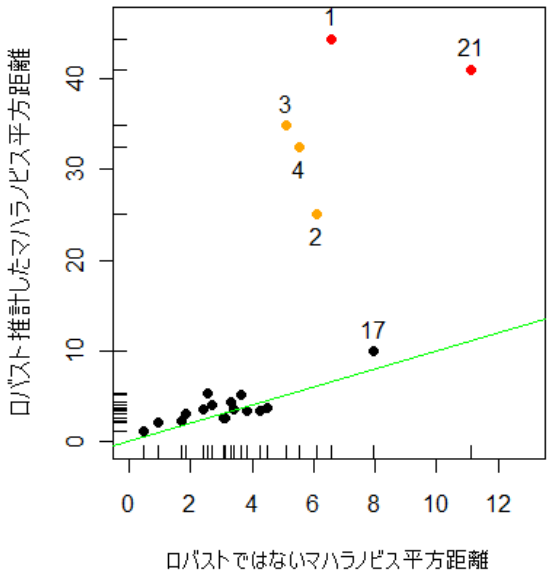
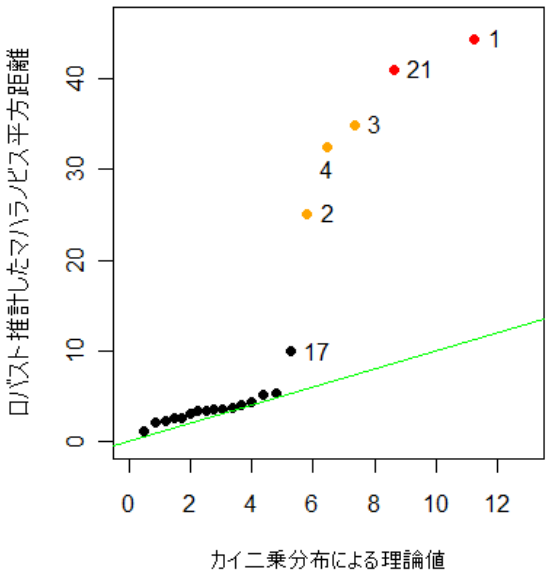


図 4. 2. 9 Q-Q プロット
［Stackloss データ］



（6）Modified Wood Specific Gravity データ

このデータセットの外れ値は､データ番号No.4，6，8，19の４つであることが分かっている。

外れ値判定基準を検定統計量 F の 99.9%値及び 99%値として、乱数を変えて５回ずつ検出を行った結果は表 4.2.2 のとおり。

射影数の多いカナダ基底増加版と EUREDIT 版の結果が明らかに良いが、どちらの条件でも５回のうち１回は検出漏れを起こした。この４つのデータポイント以外は、どの条件でもすべて正常値と判定される。

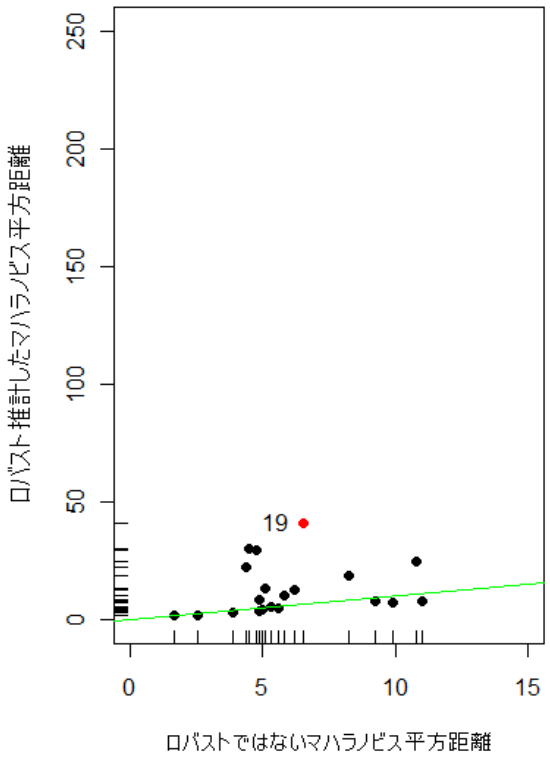
表 4.2.2 内で青と緑の四角で囲んだ部分をそれぞれ試行 A、試行 B とすると、試行 A では検出漏れが起きているが、試行 B ではすべての外れ値を正しく検出している。図 4.2.10 は、この２つの試行の D-D プロット及び Q-Q プロットを対比させ、それぞれの試行で検出される外れ値を赤で示したもの。図 4.2.11 の散布図行列では、試行 A でも B でも検出される No.19 を赤、試行 A で検出されないが試行 B で検出される No.4，6，8 を橙で示している。

表 4.2.2 Modified Wood Specific Gravity データ検出結果

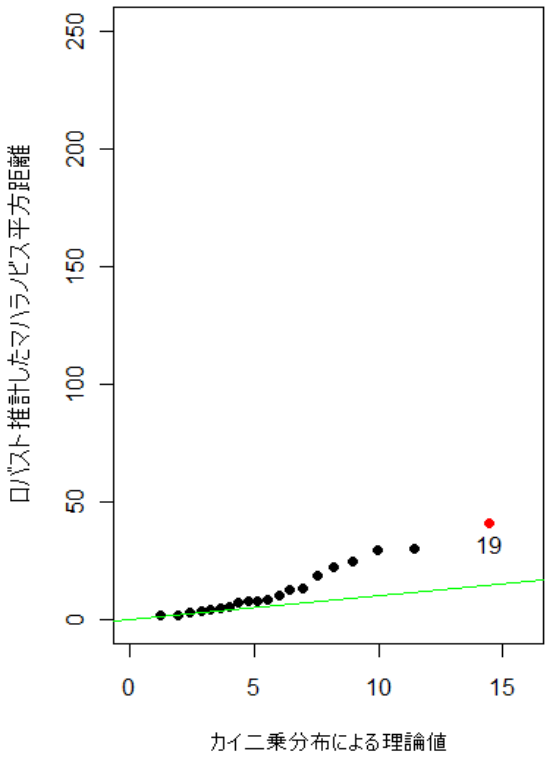
データ番号	99.9%基準				99%基準				
	No.4	No.6	No.8	No.19	No.4	No.6	No.8	No.19	
カナダ版	0	0	0	0	0	0	0	0	
	0	0	0	0	0	0	0	1	
	0	0	0	0	0	0	0	0	
	1	1	1	1	1	1	1	1	
	0	0	0	0	0	0	0	0	
カナダ基底増加版	0	1	1	1	1	1	1	1	
	1	1	1	1	1	1	1	1	
	0	1	1	1	1	1	1	1	
	1	1	1	1	1	1	1	1	
	0	0	0	0	0	0	0	0	
EUREDIT 基底減少版	0	0	0	0	0	0	0	0	
	0	1	1	1	1	1	1	1	
	0	0	0	0	0	0	0	1	
	0	0	0	0	0	0	0	0	
	0	0	0	0	0	0	0	0	
EUREDIT 版	0	0	0	0	0	0	0	1	A
	1	1	1	1	1	1	1	1	B
	1	1	1	1	1	1	1	1	
	0	1	1	1	1	1	1	1	
	0	1	1	1	1	1	1	1	

図 4. 2. 10 **D-D プロットと Q-Q プロット［Modified Wood Specific Gravity データ］**

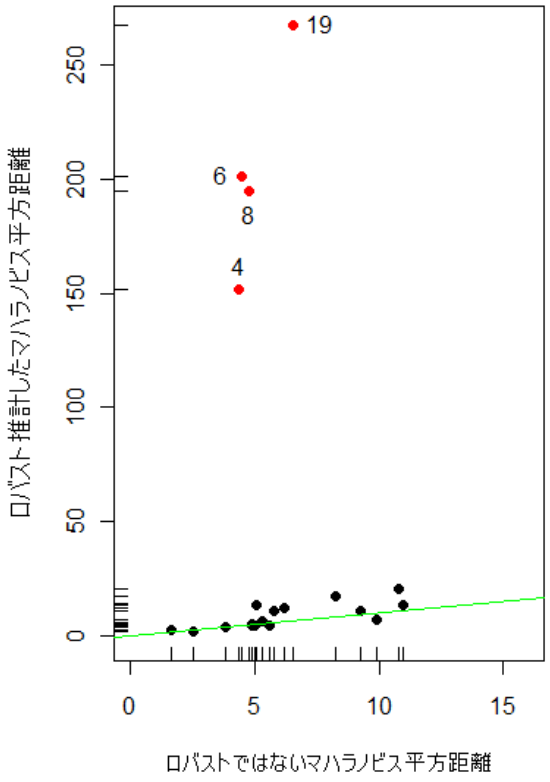
試行 A の D-D プロット



試行 A の Q-Q プロット



試行 B の D-D プロット



試行 B の Q-Q プロット

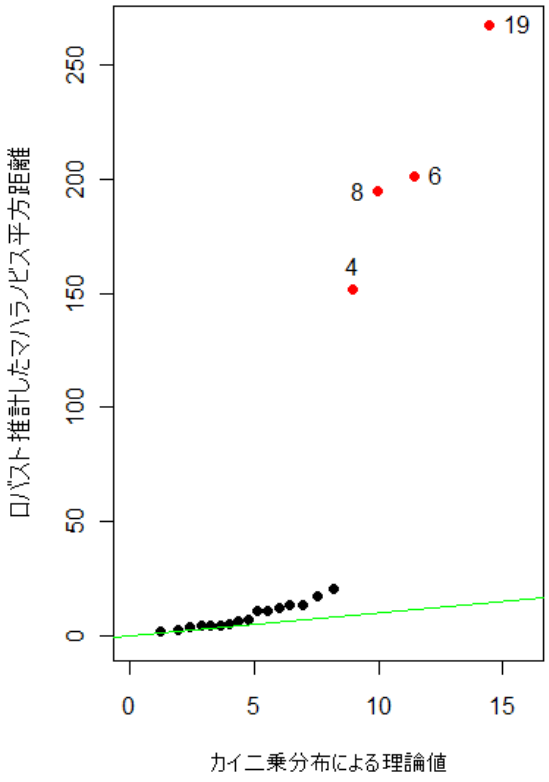
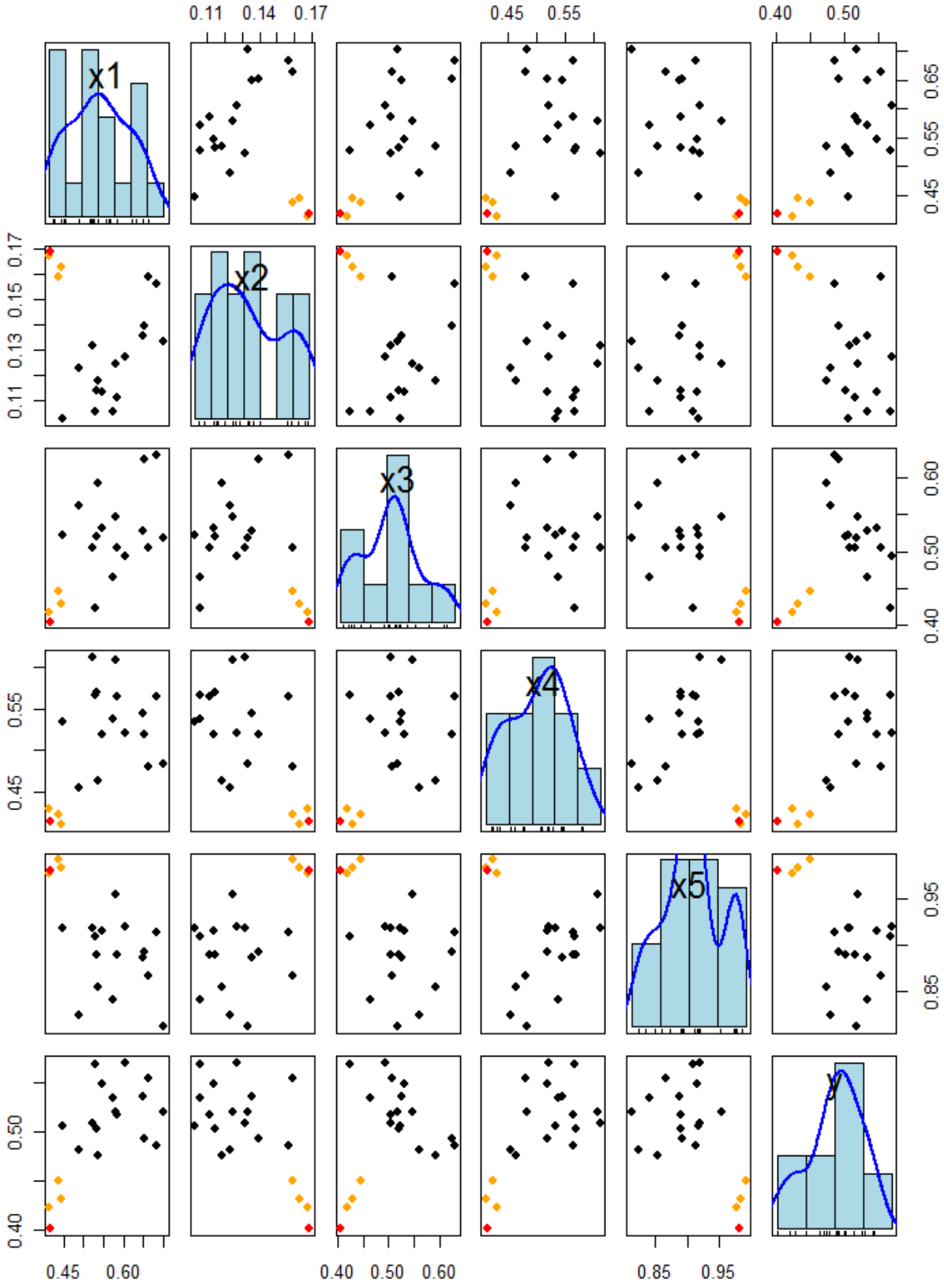


図 4. 2. 11 **Modified Wood Specific Gravity** データ散布図行列



３．結果のまとめ

シミュレーション及びデータテスト結果からは、今回テストを行った表 3.1.1 に示す４つの条件のうち、EUREDIT 版が最も優れるといえる。

ただし、MSD 法は残差の刈り込みに正規分布の仮定を置くパラメトリックな手法であり、データが正規分布から遠くなれば検出力が落ちる。また、射影時に乱数を使用しているため、難易度の高いデータの場合は、外れ値検出の結果が安定しないことがあり、複数回の検出実行や他手法との併用が有用である。

Ⅴ　考察

１．カナダ統計局での実用化

Béguin and Hulliger（2003）によれば、Franklin and Brodeur（1997）で紹介されているカナダ統計局での年次卸売・小売業調査（AWRTS）への MSD 法の適用が、各国統計部局における多変量外れ値検出法実用化の唯一の例である。

AWRTS は、卸売・小売業の基礎統計収集を目的として、層別単純ランダム抽出で標本抽出が行われ、1995 年調査時点での標本規模は約 27,000 社と大きい。ミクロエディティングの負担を軽減するためにチェック前のデータに 1993 年から MSD 法が適用されている。検出された外れ値はホットデック法で補定を行うときのドナーや平均値補定を行うときの平均値算出など補定に関する処理からは除外されるため、この外れ値検出は省力化に加えてデータ品質向上にも貢献していると考えられる。

卸売・小売の別、チェーンか否か、会社規模（大・中・小）、商業グループにより区分された補定のためのドメイン（層）によりデータを細分化し、数百あるドメイン別に５つの数量項目に関して MSD 法による多変量外れ値検出を行っている。対象となる数量項目は、期首棚卸高、期末棚卸高、総支出、商品原価及び賃金・福利厚生費である。データ分布の対称性を確保するため、すべての数量項目について営業収入との比をとり、期首棚卸高及び期末棚卸高については対数変換も行われる。

外れ値検出プログラムは C 言語で開発され、人手による審査の作業量をコントロールするために検出する外れ値数を制限する機能を持ち、検出された外れ値はマハラノビス平方距離に最も寄与した変数を手掛かりに審査される。

Franklin and Brodeur（1997）は、1995 年の実績で、検出された外れ値は全体の 4%で、このうちデータ修正を行う誤データが全体の８割、データ修正を行わない特異値が残りの２割であったと述べている。

２．統計調査データへの適用に向けて

ロバストな多変量外れ値検出法は、どの手法を選び検出率がどのようなときにどの程度かということに加えて、どの調査項目を対象としてどのように適用するかということも重要になる。本節では、カナダ統計局の事例を踏まえて実用化に向けて考慮すべき幾つかの事項について考察を行う。

(1) 変数の選択

必要な変数を見落とすと外れ値が検出できないが、一方で、調査項目をすべてまとめて検出を行うことはコンピュータ処理の観点からも検出精度の観点からも現実的ではない。

一般に、ロバストな多変量外れ値検出法はデータ処理の負荷が高い。MSD 法の場合は特に射影数が多くなるほど処理時間が増えて使用メモリも増大するため、いかに対象とする変数を削るかが実用上とても重要である。加えて、対象とする変数が多い場合は、検出された外れ値の審査の負担が増える。このため、変数間に何らかの関係性のあるものだけを選び、対象変数として一緒に外れ値検出を行う必要がある。また、多変量外れ値検出法は単変量の外

れ値も検出するが、ある１つの変数で極端な値をとるが他の変数では全体の傾向から外れていないような単変量外れ値は、対象変数が多いほど検出されにくくなる傾向を持つ。

別紙４の調査票を見る限り、AWRTS で調査されている数量項目は対象とした５変数より多く、何らかの基準で対象とする変数を５つに絞っていると思われる。

(2) カテゴリ項目の処理

MSD 法などの多変量外れ値検出法は数量項目を対象としているが、通常統計調査データはカテゴリ項目と数量項目が混在する。このため、外れ値検出の対象とする数量項目に影響があるカテゴリ変数をすべて選んでクロス集計を行い、グループ内のデータができるだけ単一分布になるようデータを細分化し、グループごとに外れ値検出を行わなければならない。AWRTS の場合は、補定のためのドメイン（層）を使用している。

(3) 欠測値への対応

Franklin and Brodeur（1997）及び Béguin and Hulliger（2003）による MSD 法自体は欠測値に対応していないが、AWRTS では対象変数の数を変えて繰り返し外れ値検出を行うことにより欠測のあるデータ値に適用を行っている。

AWRTS は、対象となる５変数のうち欠測があるのが期首棚卸高と期末棚卸高であるため、対象変数に欠測のないデータについて、５変数すべて、期首棚卸高を除いた４変数、期首棚卸高・期末棚卸高を除いた３変数という形で３回外れ値検出を繰り返している。

(4) データ変換について

MSD 法は、ロバストな手法だがデータを対称な正規分布と仮定しており、このため AWRTS ではデータ分布の対称性を確保するために５変数すべてを営業収入との比率とし、期首棚卸高及び期末棚卸高は対数変換も行っている。

対称分布の仮定を置く手法については、データの対称性に問題がある場合に、対数変換を含む Box-Cox 変換など、データを正規分布化する前処理が行われることが多いが、こうした変換により、特に正常値に近接しており分散が正常値よりも大きな外れ値は検出しにくくなるため、注意が必要である。

図 5.2.1 は、シミュレーションテストで使用した対数正規分布データによる、対数変換の例である。正常値は、変数に相関があり標準偏差 1 で原点を中心とする多変量正規分布データを指数化したもの。外れ値は、A では標準偏差 5、B では標準偏差 0.1 の正規分布データを、第一軸だけ原点から距離 10 離れたところに加えている。外れ値の標準偏差が大きい A の場合、外れ値の標準偏差が小さい B と比較すると変換前もかなり正常値と外れ値が混ざってしまっているが、変換により更に正しい外れ値の検出が困難になることが分かる。

また、一般にゼロや負の数がある場合、データがすべて正の数になるよう平行移動してから Box-Cox 変換を行うが、Box-Cox 変換は平行移動不変ではない。つまり、どの程度動かすかにより外れ値検出の結果が変化してしまう。

(5) 人による審査の必要性

検定統計量 F の 99.9%点という MSD 法の外れ値検出基準は飽くまでも目安であり、第Ⅳ章で取り上げた Stackloss データや Modified Wood Specific Gravity データの検出例のように、必要に応じて調整が必要になる場合がある。また、外れ値検出の難易度が高いデータで複数

回の検出の実行や他の多変量外れ値検出法との併用を行うと、幾つもの異なる結果が得られる場合があり、最終的にどの結果が正しいかは、調査データに関する知見を持つ人間が総合的に判断する必要がある。

AWRTS でも検出した外れ値は人が審査している。検出プログラムはドメインごとに検出される外れ値の数を制御可能で、外れ値検出基準も調整できる仕様である。検出法はカナダ版の MSD 法のみを使用し、検出は１回のみだが、外れ値が少なく分布が比較的正規分布に近いとすれば、検出の難易度が高くないため問題は起きにくいと思われる。

おわりに

本稿で取り上げた、MSD 法のようなロバストな多変量外れ値検出法は、特定の変数で必ずしも極端な値をとらないが変数の関係性から見ると他の大部分のデータの傾向と異なるような外れ値を検出することができるのが特徴である。

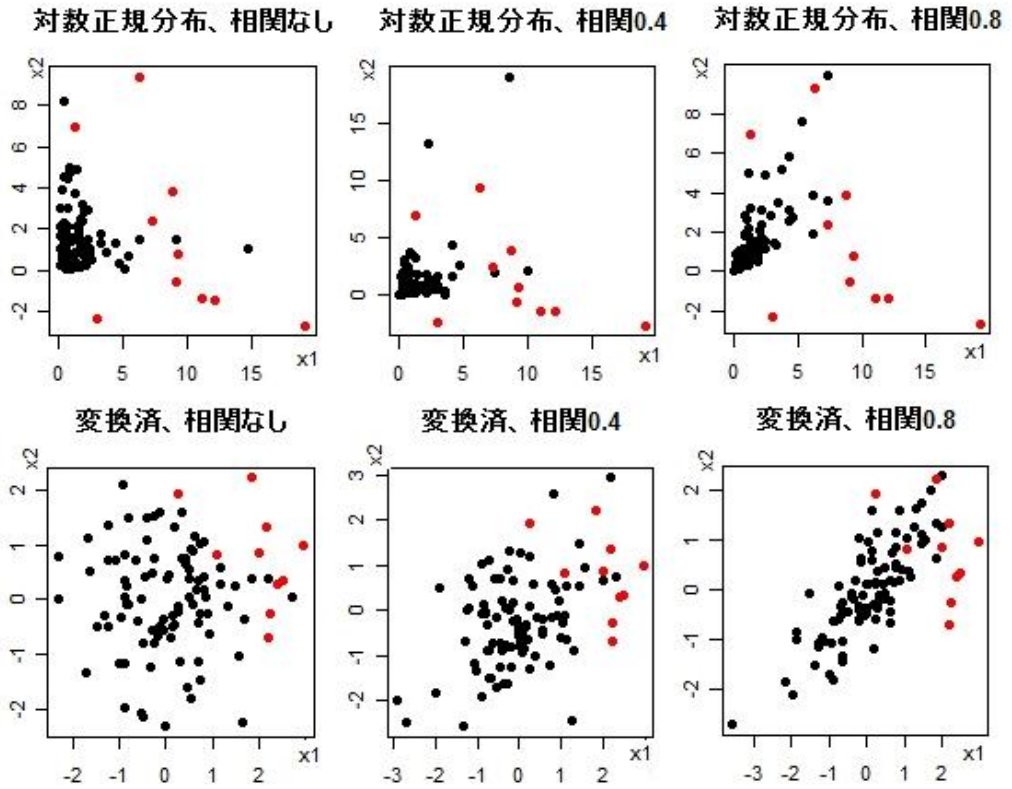
統計調査の製表業務において、集計表が最終成果物である場合、個々のデータの変数（調査項目）間の関係性の情報はほとんど残らないため、このような外れ値検出の必要性は高くないが、統計データの利用促進を図る新たな統計法が平成 21 年４月から全面施行され、個別データの提供可能性が拡大した。

集計表と異なり、個別データは変数間の関係性がそのまま保持されており、利用者もこのような関係性の分析を目的として個別データを利用することが多いため、MSD 法のような多変量外れ値検出法が有効になる可能性がある。

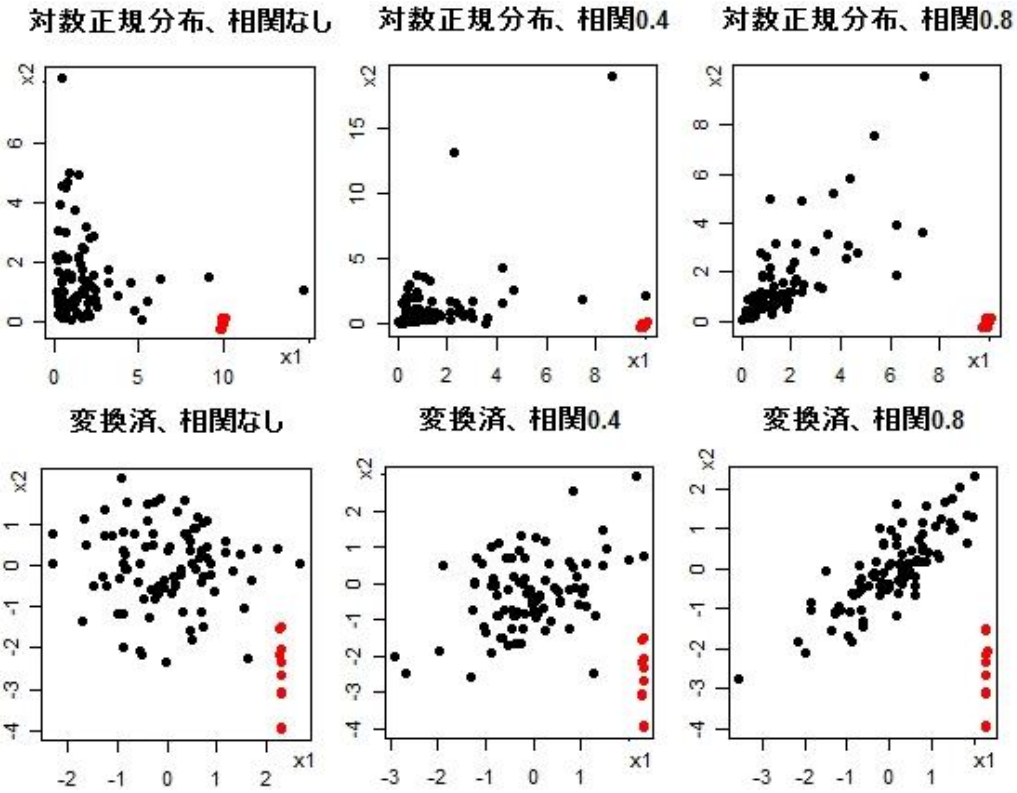
今後は、今回評価用に作成した MSD 法プログラムを大規模データにも対応できるよう改良を行うとともに、MSD 法以外の外れ値検出法も含めてベンチマーキングを行い、実用化につなげていきたい。

図 5.2.1 外れ値を含む対数正規分布データへの対数変換の影響

A. 外れ値の分散が大きい場合



B. 外れ値の分散が小さい場合



別紙１　外れ値検出プログラム

```
#########################################################################
#              MSD 法多変量外れ値検出関数  Ver. 1.6   09/07/14
#########################################################################
####### カナダ版と EUREDIT 版の比較評価用。
####### 対応できるデータの大きさ・次元数はメモリに依存する。
####### 外れ値検出データをインプットとして与えると、ロバスト推計した平均値ベクトルと共分散行列を計算する
#########################################################################
#######   関数 msd パラメーター一覧
#######
####### inp      必須。外れ値検出を行いたい n×p のデータ行列
#######   以下のパラメータはすべて省略可能。デフォルトは EUREDIT 版の設定。
####### sd       乱数のシード。再現性が必要な場合に指定(0 のときは毎回違う再現不能な乱数)。
####### nb       基底数コントロール用。1 次元当たりの基底数を設定できる。
####### tm       "CAN"でカナダ版、"EUR"で EUREDIT 版によるウエイトの刈り込みを行う。
#########################################################################
#######   関数 msd  戻り値一覧
#######
####### u1       一次ウエイトにより推計された平均値ベクトル
####### V1       一次ウエイトにより推計された共分散行列
####### bwt      一次ウエイト
####### u2       最終平均値ベクトル
####### V2       最終共分散行列
####### wts2     最終ウエイト
####### eg       一次ウエイトにより推計された共分散行列の固有値
####### ctb      一次ウエイトにより推計された共分散行列の固有ベクトル
#########################################################################

msd <- function(inp, nb=0, sd=0, tm="EUR") {

 inp_d <- ncol(inp)                                  # 次元数
 inp_n <- nrow(inp)                                  # データ数

###################
#     基底作成
###################

if (sd != 0) set.seed(sd)

## 必要基底数セット:
if (nb == 0) bb_n <- trunc(exp(2.1328+0.8023*inp_d) / inp_d)
   else bb_n <- nb
rn <- bb_n * inp_d ^2                               #  必要な一様乱数の数
basis <- array(runif(rn), c(inp_d, inp_d, bb_n))

# 直交化
basis <- apply(basis, 3, gso)
basis <- array(basis, c(inp_d, inp_d, bb_n))

###################
#     射影と残差計算
###################

prj     <- array(0, c(inp_n, inp_d, bb_n))          #  射影用
res     <- array(0, c(inp_n, inp_d, bb_n))          #  残差
wt      <- array(0, c(inp_n, inp_d, bb_n))          #  ウエイト
wts     <- array(0, c(inp_n, bb_n))                 #  次元別ウエイトの積和
```

```
bwt     <- rep(0, inp_n)                          #  データごとの最良基底によるウエイト
kijun   <- qchisq(0.95, inp_d)

Fprj    <- function(pj) t(pj %*% t(inp))          # 射影ベクトルの大きさ計算
prj <- apply(basis, 3, Fprj)
prj <- array(prj, c(inp_n, inp_d, bb_n))          # 整形

medi <- apply(prj, c(2, 3), median)               # 中位数
madx <- apply(prj, c(2, 3), mad)                  # 中央絶対偏差 / 0.674 (標準偏差化)

for (i in 1:bb_n) {                               #  基底数だけループ
  res[,,i] <- t(abs(t(prj[,,i]) - medi[,i]) / madx[,i])
}

###  ウエイト刈り込み
if (tm == "CAN") {
   k0     <- which(res <= 1.75)
   k1     <- which(res > 1.75 & res <= 3.5)
   k2     <- which(res > 3.5)
   wt[k0] <- 1
   wt[k1] <- 1.75 / res[k1]
   wt[k2] <- 0
}
else {                                            #  Huber-like な刈り込み
   k0      <- which(res <= sqrt(kijun))
   k1      <- which(res > sqrt(kijun))
   wt[k0]  <- 1
   wt[k1]  <- kijun / (res[k1]^2)
}
wts <- apply(wt, c(1,3), prod)                    # ウエイトの積和
bwt <- apply(wts, 1, min)                         # 最良基底を選択

###     最初のロバストな共分散行列
u1 <- apply(inp * bwt, 2, sum) / sum(bwt)
V1 <- t(t(t(inp) - u1) * bwt) %*% (t(t(inp) - u1) * bwt) / sum(bwt^2)

###     V1 のエラー処理
###    sum(bwt)がゼロだと u1 も V1 も NaN になるので、ゼロに置換して異常終了を回避する

u1 <- ifelse(is.nan(u1), 0, u1)
V1 <- ifelse(is.nan(V1), 0, V1)

###     ロバストな主成分算出
eg      <- eigen(V1, symmetric=TRUE)              # LAPACK 使用
ctb     <- eg$value / sum(eg$value)               # 寄与率

###################
#   二回目の射影と最終ウエイト決定
###################

res2    <- array(0, c(inp_n, inp_d))              # 残差
wt2     <- array(0, c(inp_n, inp_d))              # ウエイト 次元別
wts2    <- array(0, inp_n)                        # 最終ウエイト　積和

prj2  <- t(eg$vector %*% (t(inp) - u1))           # 射影ベクトルの大きさ
medi2 <- apply(prj2, 2, median)                   # 中位数
madx2 <- apply(prj2, 2, mad)                      # 中央絶対偏差
```

```
res2 <- t(abs(t(prj2) - medi2) / madx2)          # 残差計算

### 残差刈り込み
if (tm == "CAN") {
   k0      <- which(res2 <= 1.75)
   k1      <- which(res2 > 1.75 & res2 <= 3.5)
   k2     <- which(res2 > 3.5)
   wt2[k0] <- 1
   wt2[k1] <- 1.75 / res2[k1]
   wt2[k2] <- 0
}
else {                                           #  Huber-like な刈り込み
   k0      <- which(res2 <= sqrt(kijun))
   k1      <- which(res2 > sqrt(kijun))
   wt2[k0]  <- 1
   wt2[k1]  <- kijun / (res2[k1]^2)
}

wts2 <- apply(wt2, 1, prod)                      # 次元の積和

if (tm == "EUR") wts2 <- pmin(wts2, bwt)
# EUREDIT 版は一次ウエイトと比較して小さい方を採用
# カナダ版は二次ウエイトをそのまま使用

###################
#     最終位置ベクトルと共分散行列
###################

#  位置ベクトル
u2 <- apply(inp * wts2, 2, sum) / sum(wts2)
V2 <- t(t(t(inp) - u2) * wts2) %*% (t(t(inp) - u2) * wts2) / sum(wts2^2)

return(list(u1=u1, V1=V1, bwt=bwt, u2=u2, V2=V2, wts2=wts2, eg=eg, ctb=ctb))
}

#########################################################################
#     gso: 基底を直交化する関数
#########################################################################
#  Gram-Schmidt Orthonormalization  （関数 msd で使用）
#  正方行列を受け取り、横ベクトル同士を直交化して戻す
#########################################################################

gso <- function(basis)  {
        bd <- ncol(basis)                        # 横
        bn <- nrow(basis)                        # 縦
        basis[1,] <- basis[1,] / sqrt(t(basis[1,]) %*% basis[1,])
        for (i in 2 : bd ) {
                wk1 <- basis[i,]
                for (j in 1:(i-1)) {
                        wk2 <- basis[j,]
                        basis[i,] <- basis[i,] - (t(wk1) %*% wk2) * wk2
                }
                basis[i,] <- basis[i,] / sqrt(t(basis[i,]) %*% basis[i,])
        }
        return(basis)
}
```

```
#########################################################################
#               MSD 法多変量外れ値検出関数　使用例
#########################################################################

#source("MSD.r")          # 関数類をまとめてファイル MSD.r に収めた場合に使用

# 山火事データ呼び出し
  data(bushfire, package="robustbase")
  dat <- as.matrix(bushfire)              # データを行列化
  n <- nrow(dat)                          # データ数
  d <- ncol(dat)                          # 変数の数

# MSD 法
  msdout  <- msd(dat)                     # EUREDIT 版の設定になる

# 算出された最終平均値ベクトルと共分散行列から、各データの中心からのマハラノビス平方距離を算出
  mah  <- mahalanobis(dat, msdout$u2,  msdout$V2)

#  検定統計量を計算
  FF <- mah  * ( n - d )* n /(( n^2 - 1 )* d)

#  外れ値の基準はF分布
  cf99   <- qf(0.99,    d, n - d)
  cf999  <- qf(0.999,   d, n - d)         # 目安となる基準値

# 外れ値フラグ　　正常値は1
  ot <- rep(1, n)
  # 基準より大きいものは、フラグを2にセット
  ot[which(FF > cf999)] <- 2              # 外れ値

# 外れ値の数とデータ番号を表示
  length(which(ot==2))
  which(ot==2)

# 外れ値を色分けして散布図行列にプロット
  pairs(dat, pch=19, col=ot)

#  Q-Q プロット
  qqplot(qchisq(ppoints(n), df=d), mah,  pch=19, col=sort(ot), main = "Q-Q プロット",
      xlab="カイ二乗分布による理論値", ylab="ロバスト推計したマハラノビス平方距離")
  abline(0, 1, col = 'green')
```

別紙２　シミュレーションデータの設計

○ 正規分布

①	正常値の分布	平均 0、標準偏差 1 の正規分布
②	データ数	100
③	変数の数	5, 10, 20
④	変数間の相関	0, 0.4, 0.8
⑤	外れ値の分布	正規分布
⑥	外れ値割合	0%, 10%, 20%, 30%, 40%, 50%
⑦	外れ値の原点からの距離	5, 10, 100　第 1 軸方向のみ
⑧	外れ値の標準偏差	0.1, 1, 5

○ Skew-T 分布

①	分布の種類	平均 0、標準偏差 1 の正規分布
②	データ数	100
③	変数の数	5, 10, 20
④	変数間の相関	0, 0.4, 0.8
⑤	歪度（わいど, skewness）	0, 1, 5, 10　第 1 軸方向のみ
⑥	自由度	1, 10, Inf（無限大）
⑦	外れ値の分布	正規分布
⑧	外れ値割合	0%, 10%, 20%, 30%, 40%, 50%
⑨	外れ値の原点からの距離	10, 100　第 1 軸方向のみ
⑩	外れ値の標準偏差	0.1, 1, 5

○ 複合ポワソン分布

①	分布の種類	複合ポワソン分布 パラメータは p, mu, ph
	p: power, 2: Gamma, 3: Inverse-Gaussian	2.5
	mu: mean	1
	ph: dispersion	1

※　各変数ごとに単変量複合ポワソン分布に従う乱数を発生させているが、コレスキー分解によって相関導入しているため、最終的に作成する多変量データは厳密には複合ポワソン分布とはいえない。

②	データ数	100
③	変数の数	5, 10, 20
④	変数間の相関	0, 0.4, 0.8
⑤	外れ値の分布	正規分布
⑥	外れ値割合	0%, 10%, 20%, 30%, 40%, 50%
⑦	外れ値の原点からの距離	10, 100　第 1 軸方向のみ
⑧	外れ値の標準偏差	0.1, 1, 5

○ 対数正規分布

①	正常値の分布	平均 0、標準偏差 1 の相関付き正規分布を指数化
②	データ数	100
③	変数の数	5, 10, 20
④	変数間の相関	0, 0.4, 0.8
⑤	外れ値の分布	正規分布（すべて正の数になるよう絶対値化）
⑥	外れ値割合	0%, 10%, 20%, 30%, 40%, 50%
⑦	外れ値の原点からの距離	10, 100　第 1 軸方向のみ
⑧	外れ値の標準偏差	0.1, 1, 5

別紙３　使用したテストデータ一覧

１．Hawkins-Bradu-Kass データ

３変数（＋応答変数)、75 データ、外れ値は 14（No.1～14)。
1984 年に Hawkins らが作成した人工データセット。

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

２．スイスのレストラン業データ

２変数（従業者数［対数]、売上高［対数])、1273 データ。
1995 年スイス企業センサスでのレストラン業の擬似データ。データ自体は公開されていないため、プロット図の点の位置情報から擬似データを作成した。

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

３．Hertzsprung-Russell データ

２変数（星の表面温度の対数と光密度の対数）47 データ。

出典：Rousseeuw, P. J. and A. M. Leroy（1987）, Robust Regression and Outlier Detection, John Wiley & Sons

４．Bushfire（山火事）データ

５変数、38 データ。
山火事の痕跡を分析するため、衛星から測定したデータ

出典：Campbell, N. A.（1989）, Bushfire mapping using noaa avhrr data, Technical report, CSIRO

５．Stackloss データ

４変数 21 データ。外れ値は５つ（No.1，2，3，4，21)。
アンモニアを酸化して硝酸を作る工場の 22 日分の吸収塔損失データ。生産された硝酸は、向流吸収塔で吸収される。変数はそれぞれ操業率、冷却水の温度、酸の集中度、アンモニア損失量。

出典：Rousseeuw, P. J. and A. M. Leroy（1987）, Robust Regression and Outlier Detection, John Wiley & Sons

６．Modified Wood Specific Gravity データ

５変数（＋応答変数)、20 データ。人工データで、外れ値が４つ（No.4，6，8，19)。

出典：Rousseeuw, P. J. and A. M. Leroy（1987）, Robust Regression and Outlier Detection, John Wiley & Sons, P243

別紙４　カナダ卸売・小売業調査（AWRTS）調査票

Distributive Trades Division

Annual Wholesale and Retail Trade Survey 1997

Si vous préférez recevoir ce questionnaire en français, veuillez téléphoner au numéro approprié indiqué dans la section 2.

Confidential when completed

Correct pre-printed information if necessary using the corresponding boxes provided below.

Legal Name
Business Name
C / O
No. & Street
City
Province / Postal code
Contact
Telephone no. Area code / Extension
Facsimile no. Area code

1 - PLEASE READ CAREFULLY BEFORE COMPLETING

AUTHORITY
Collected under the authority of the Statistics Act, Revised Statutes of Canada, 1985, Chapter S19.

PURPOSE OF THE SURVEY
Data collected by this survey are used to provide aggregated industry information required by governments to develop national and regional economic programs and policies. This information is also used by the private sector to assist in decision making and to assess business conditions. The results of this survey are published in Statistics Canada catalogue number 63-236-XPB.

CONFIDENTIALITY
Statistics Canada is prohibited by law from publishing any statistics which would divulge information obtained from this survey that relate to any identifiable business without the previous written consent of that business. The data reported on this questionnaire will be treated in strict confidence. They will be used exclusively for statistical purposes and will be published in aggregate form only. The confidentiality provisions of the Statistics Act are not affected by either the Access to Information Act or any other legislation.

DATA SHARING AGREEMENTS
To avoid duplication of enquiry and to provide consistent statistics, agreements have been made under Section 11 of the Statistics Act, to exchange information collected by this survey with the Bureaus of Statistics of Alberta (for retail trade data) and Manitoba (for retail and wholesale trade data). Agreement has also been made under Section 12 of the Statistics Act, with the Northwest Territories Bureau of Statistics for the sharing of information from this survey. Under Section 12, you may refuse to share your information with the Northwest Territories Bureau of Statistics by writing to the Chief Statistician and returning your letter of objection along with the completed questionnaire.

2 - INQUIRIES

The questionnaire should be completed and returned in the postage paid envelope **within 30 days** of receipt. If you require assistance in the completion of the questionnaire or have any questions regarding the survey, please call the nearest Statistics Canada Regional Office.

	LOCAL	TOLL FREE	FAX
Halifax	426 - 5662	1 - 800 - 565 - 1685	902 - 426 - 8292
Montréal	283 - 5724	1 - 800 - 363 - 6720	514 - 283 - 7969
Toronto	954 - 9069	1 - 800 - 263 - 3072	416 - 973 - 6524
Edmonton	495 - 4627	1 - 800 - 661 - 9884	403 - 495 - 4788
Vancouver	666 - 2100	1 - 800 - 663 - 0172	604 - 666 - 6495

3 - REPORTING PERIOD

Please report for your 1997 fiscal year *(normal business year)* ending any time between April 1, 1997 and March 31, 1998

FROM D M Y　TO D M Y

4 - KIND OF BUSINESS

List the main lines of merchandise and services sold and indicate the estimated percentage of total operating revenue

4.1	1		%
4.2	2		%
4.3	3		%

5 - REVENUE BY CLASS OF CUSTOMER

Indicate the estimated percentage of total operating revenue:

5.1	Made to households or individuals for personal use		%
5.2	Made to retail businesses		%
5.3	Made to wholesale businesses		%
5.4	Made to industrial, commercial and other business users		%
5.5	Made to farmers, for farm operations		%
5.6	Made to clients outside Canada (exports)		%
5.7	**Total**	100	%

5-3200 (1997-12-12)　STC/IND-375-75008

Statistics Canada　Statistique Canada

Canada

営業収入

期首棚卸高

期末棚卸高

商品原価

賃金・
福利厚生費

総支出

6 - REVENUE (Canadian dollars)

6.1	Sales of goods purchased for resale *(Report gross sales of new and/or used goods **less** returns, discounts, provincial sales taxes and GST or HST. Do not deduct the value of trade-ins)*	$	.00
6.2	Revenue from gross commissions earned from buying and/or selling merchandise on account of others	$	.00
6.3	Sales of products manufactured at your company facilities and included in your total revenue	$	.00
6.4	Labour receipts from installation and repair of goods *(Report parts in (6.1) above)*	$	.00
6.5	Revenue from rental and leasing of goods and equipment	$	.00
6.6	Other operating revenue *(such as rental from real estate, food serving and other activities; **If a grain wholesaler**, include revenue from handling, storage, drying and other charges)*	$	.00
6.7	**Total Operating Revenue** *(Sum of 6.1 to 6.6 above)*	$	.00
6.8	**Non-operating revenue** *(such as subsidies, interest, dividends and gains on disposal of investments and capital assets)*	$	.00
6.9	**Total Revenue** *(Sum of 6.7 and 6.8)*	$	.00

7 - COST OF GOODS SOLD (Canadian dollars)

7.1	Opening inventory	$	.00
7.2	Purchases of new and used goods for resale *(Include freight-in and the value of goods taken in trade, less returns and discounts.)*	$	.00
7.3	Closing inventory	$	.00
7.4	**Cost of goods sold** *(Sum of 7.1 and 7.2) less 7.3*	$	.00

8 - EXPENSES (Canadian dollars)

8.1	Salaries, wages and benefits: *(Include bonuses, commissions and any other payments to employees. Report gross payments before deductions for such items as income tax, pension plans, U.I. premiums)*	$	.00
8.2	Depreciation on buildings and other fixed assets	$	.00
8.3	Interest expenses	$	.00
8.4	Other expenses *(**Exclude** income taxes, cost of goods sold and purchases of new and used goods for resale)*	$	.00
8.5	**Total expenses** *(Sum of 8.1 to 8.4)*	$	.00

9 - LOCATIONS AND PROVINCIAL DISTRIBUTION

9.1 Indicate the number of trading locations (if retail) or distribution centres/sales offices (if wholesale) operated in Canada during your reporting period, irrespective of the length of time they were open. ➔

9.2 Do you operate your trading locations (if retail) or distribution centres/sales offices (if wholesale) in more than one province? ☐ No ▶ *Skip to 10* ☐ Yes ▶ *If yes, report the following items.*

Province / Territory	Trading Locations (if retail) or distribution centres/ sales offices (if wholesale) (9.1)	Operating Revenue (6.7)	Salaries, Wages and Benefits (8.1)	Cost of Goods Sold (7.4)
	Number	$ or %	$ or %	$ or %
Newfoundland				
Prince Edward Island				
Nova Scotia				
New Brunswick				
Québec				
Ontario				
Manitoba				
Saskatchewan				
Alberta				
British Columbia				
Yukon Territory				
Northwest Territories				
Canada				

10 - CERTIFICATION

I certify that the information contained herein is complete and correct to the best of my knowledge and belief.

Signature of authorized person	Title	Date
Name of contact for further information (please print)	Telephone no. Area code / Extension	Facsimile no. Area code

Thank you for completing this survey. Please maintain a copy for your records.
You may be called at a later date to verify the data.

参考文献

[1] Béguin, C. and B. Hulliger（2003）, Robust Multivariate Outlier Detection and Imputation with Incomplete Survey Data, EUREDIT Deliverable D4/5.2.1/2 Part C

[2] Campbell, N. A.（1989）, Bushfire mapping using noaa avhrr data, Technical report, CSIRO

[3] Donoho, D. L.（1982）, Breakdown properties of multivariate location estimators, Ph.D. Qualifying paper, Harvard University

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

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

[6] Istat, CBS, SFSO, Eurostat（2007）, Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys, EDIMBUS Project

[7] Maronna, R. A., and V. J. Yohai（1995）, The behavior of the Stahel-Donoho robust multivariate estimator, Journal of the American Statistical Association, Vol.90, No.429, pp.330-341

[8] Patak, Z.（1990）, Robust principal component analysis via projection pursuit, M. Sc. Thesis, University of British Columbia, Canada

[9] Peña, D. and F. J. Prieto（2001）, Multivariate outlier detection and robust covariance matrix estimation, Technometrics, Vol.43, pp.286-300

[10] Rousseeuw, P. J. and A. M. Leroy（1987）, Robust Regression and Outlier Detection, John Wiley & Sons

[11] Stahel, W. A.（1981）, Breakdown of covariance estimators, Research Report 31, Fachgruppe für Statistik, E.T.H. Zürich

[12] 岡本政人（2004），多変量外れ値検出法の研究動向，製表技術研究レポート 1，（独）統計センター研究センター，pp.1-34

[13] 小林良行（2009），ヨーロッパにおけるデータエディティング及び補定に関する調査報告～EDIMBUS プロジェクトを中心に～，統計研究彙報第 66 号，総務省統計研修所，pp.101-129

[14] 吉澤　正（1992），「統計処理」，岩波書店，pp.11

[15] 和田かず美（2004），多変量外れ値検出法の比較，2004 年度統計関連学会連合大会講演報告集，pp.95-96