# 研究ノート

# R の hclust 関数による 正しい ward 法分析の方法
## ～その 1：ward 法の特徴と検証実験～


**木村　敦** | Kimura Atsushi

㈱統計センター 理事・CIO

■ NTT にて ICT 関連開発に長年携わり、㈱ NTT ファシリティーズ総合研究所 取締役情報技術本部長を経て、2019 年 4 月から現職。1988 年 3 月名古屋大学大学院理学研究科博士課程（前期）修了、修士（理学）、専門統計調査士。


### 1. はじめに

本誌 2021 年 3 月号に、家計調査の階層的クラスター分析記事（木村･高部 [1]）を掲載した。記事を読んだ複数の方から「クラスター分析に興味があるが、R の hclust 関数での ward 法のオプションの違いと正しい使い方、分析結果の見方が良くわからない」との声を頂いた。

大隅 [2] は「階層的分類でよくある例だが、同じ手法名を掲げているものの、類似度・非類似度の選択や階層化手順の違い、独自に設けたオプションの差異などから、同一データに対して、用いたソフトウェアの出力結果がまるで異なることさえある」と指摘している。また Murtagh and Legendre[3] も世の中に存在する ward 法統計ソフト類に複数バリエーションが存在することを示し、利用にあたって注意喚起を行っている。R の hclust 関数はその後に機能追加が行われ ward.D オプションと ward.D2 オプションが実装されたのだが、R ドキュメントの記載は初学者向けに丁寧とは言い難い。インターネット上の情報等にも不正確な記述が散見される。ユーザーが迷うのも無理からぬ事である。

本稿では、バージョン 3.0.3 より新しい現在使われている hclust 関数の ward.D と ward.D2 オプションの違いと正しい使い方、加えて Ward[4] の基本的な考え方とその特徴を解説する。さらに、Ward[4] に則って筆者が独自に実装したプログラムと hclust 関数を実際のデータ分析事例 [1] に適用してその結果を比較し、hclust 関数の内部処理が説明したものと相違ないことを検証する。今回使用した独自実装プログラムの設計・実装については、次回詳細解説を行う。数式やアルゴリズムをプログラム実装する具体事例として参考になろう。なお本稿における見解は筆者個人のもので所属する組織を代表するものではない。また、本文章の誤記や誤りなどはすべて筆者の責に帰するものである。

### 2. R の hclust 関数の正しい使い方

先に結論を述べると、① Ward[4] に忠実なのは ward.D オプションであり、内部処理では Ward[4] で定義されている非類似度（以後「ward 非類似度」と記す）を使って計算が行われる。関数の第一引数には「平方ユークリッド距離の 1/2 倍（＝初回の「ward 非類似度」）」を渡すの

が正しい（1/2 倍の理由は第 3 節⑴で述べる）。

② ward.D2 オプションの場合、内部処理では「ward 非類似度」の平方根を元に計算を行っている。従って第一引数には『「平方ユークリッド距離の 1/2 倍」の平方根をとった値（＝『初回の「ward 非類似度」の平方根をとった値』）』を引き渡すのが正しい [5]。

上記①②の通りに関数に値を渡してやれば、両オプションによる処理は実は等価である。つまりクラスター凝集の順序や結果はどちらのオプションを指定しても同じになる。内部で扱う非類似度として「ward 非類似度」を用いるのか、『「ward 非類似度」の平方根』を用いるのかの違いだけである。結果の Height 値は凝集クラスター間の非類似度値を示している。このため、両オプションの分析結果からデンドログラムを描いた時の Height 値は、①の値の平方根をとったものが②の値と完全に一致する。

結論を簡潔にまとめると以下の通りである。hclust 関数に非類似度行列 d と method オプションを指定して分析結果を変数 hc に代入する場合、

hc <- hclust(d, method="xxxxxx")

である。ここで、xxxxxx には ward.D もしくは ward.D2 が入るものとする。クラスター分析したいデータ行列（標準化などの前処理が完了しているもの）を DATA とすると、

| xxxxxx | d に代入する値 |
|---|---|
| ward.D の時 | (1/2)*dist(DATA)^2 |
| ward.D2 の時 | sqrt(1/2)*dist(DATA) |

とするのが ward 基準的に正しい。

hclust 関数のドキュメントの Examples EX2 に ward 法の使用例が以下のように記載されている。

```
hcity.D <- hclust(UScitiesD, "ward.D") # "wrong"
hcity.D2 <- hclust(UScitiesD, "ward.D2")
```

「ward.D2 例は ward 基準に従うが、ward.D 例は従っておらず分析結果も両者で異なる」と記載されている。しかし、“wrong” なのは ward.D 例で渡している「平方されていない歪んだ初回非類似度行列」であり、ward.D オプション指定自体が “wrong” なのでは無い。初回非類似度を次のように正しく設定すれば、EX2 の場合でも両例とも ward 基準に従い、分析結果も同等になる。

```
hcity.D <- hclust((1/2)*UScitiesD^2, "ward.D")
hcity.D2 <- hclust(sqrt(1/2)*UScitiesD, "ward.D2")
```

（なお、UScitiesD は米国主要都市間の距離行列であるため、dist 関数は不要である。）

## 3. ward 法の用語定義とアルゴリズムについて

### ⑴ 「情報の損失」と「非類似度」の定義

Ward[4] では、クラスター化に伴う「情報の損失（“loss” in information）」を *ESS*（error sum of squares）で表現し、「情報の損失の増加量」をクラスター間の「非類似度」として定義する。論文では 1 次元の例で式の展開とクラスター分析実施例の説明を行っている。本稿では多次元に拡張した形式で ward 法の意味を詳細に解説する。

凝集過程における $i$ 番目のクラスター $i$ について、要素数を $Ni$、クラスター $i$ の $j$ 番目の要素ベクトルを $\boldsymbol{x_{ij}}$、クラスター $i$ の重心ベクトルを $\boldsymbol{c_i}$ とする。各要素における変量の数を $m$ とすれば、$\boldsymbol{x_{ij}}$ も $\boldsymbol{c_i}$ もともに $m$ 次元ベクトルである。

$$\boldsymbol{x_{ij}} = (x_{ij1}, x_{ij2}, \cdots, x_{ijm}),\ \boldsymbol{c_i} = (c_{i1}, c_{i2}, \cdots, c_{im})$$

クラスター $i$ の *ESS* を *ESSi* と表記することにすると *ESSi* は以下の式で示される。

$$ESSi = \sum_{j=1}^{Ni} \| \boldsymbol{x}_{ij} - \boldsymbol{c}_i \|^2 \qquad (3\text{-}1)$$

ここで‖a‖はベクトルaの長さ（ユークリッド距離）である。次にクラスター$p$とクラスター$q$の「非類似度（$DSpq$）」は以下のようになる。クラスター$p$とクラスター$q$を凝集させてクラスター$r$を形成した場合を考えて、

$$DSpq = ESSr - (ESSp + ESSq) \qquad (3\text{-}2)$$

つまり、クラスター$p$とクラスター$q$が凝集することによる「情報の損失（“loss” in information）」の「増加量」を、クラスター$p$とクラスター$q$の「非類似度（$DSpq$）」と定義しているわけである。

さらに、式 (3-1) を用いて式 (3-2) を変形する。

$$DSpq = \sum_{j=1}^{Nr} \| \boldsymbol{x}_{rj} - \boldsymbol{c}_r \|^2 - \left( \sum_{j=1}^{Np} \| \boldsymbol{x}_{pj} - \boldsymbol{c}_p \|^2 + \sum_{j=1}^{Nq} \| \boldsymbol{x}_{qj} - \boldsymbol{c}_q \|^2 \right)$$

$$= \{Np \cdot Nq \,/\, (Np + Nq)\} \times \| \boldsymbol{c}_p - \boldsymbol{c}_q \|^2 \qquad (3\text{-}3)$$

$DSpq$はクラスター$p$とクラスター$q$の各重心ベクトルと各クラスターに含まれる要素数を用いて計算できる。初回の各クラスターは要素が全て1であり、$Np = Nq = 1$のため$Np \cdot Nq / (Np + Nq) = 1/2$となる。これが初回の非類似度において、平方ユークリッド距離を1/2倍している理由である。

### ⑵ クラスター凝集アルゴリズム

非類似度（$DSpq$）が最も小さい2つのクラスターを凝集させ、クラスターが最終的に1つになるまでそれを繰り返す。Ward[4] で提案されている ward 法の要点は以上である。

### ⑶ 非類似度再計算の効率化のための更新式

各工程において算定する必要が生じる非類似度は、前凝集工程で凝集した新しいクラスターとその他のクラスターとの非類似度だけである。凝集にかかわっていないクラスター間の非類似度は前工程までで算定済である。非類似度は、式 (3-3) を用いて計算し直せば良いが、新しいクラスターの重心ベクトルを計算して他クラスターの重心ベクトルとの距離計算を行う必要がある。

世の中のツール類における実装では、非類似度の再計算処理をできる限り減らすために前工程で計算した非類似度を用いて次工程の非類似度の算出をする。このために用いられるのが「Lance-Williams 更新式」である。ある凝集工程において、クラスター$p$とクラスター$q$が凝集対象として選定されてクラスター$r$に凝集された場合を考える。クラスター$r$とそれ以外のクラスター$s$の非類似度（$DSrs$）は、前工程で算定した$DSps$、$DSqs$、$DSpq$を用いて次のように算出することができる。ward 法の場合は、

$$DSrs = \{1/(Np + Nq + Ns)\} \times \{(Np + Ns) \times DSps + (Nq + Ns) \times DSqs - Ns \times DSpq\} \qquad (3\text{-}4)$$

ここで、$Np, Nq, Ns$は各々クラスター$p$, クラスター$q$, クラスター$s$の要素数である。式 (3-4) を利用することにより、クラスターの要素ベクトルに関する演算が不要となる [6]。

### ⑷ ward 法の特徴について

ward 法は階層的クラスター分析手法のなかでも比較的バランスが良く、鎖効果が発生しにくいと言われている。それは何故なのか？　世の中の書籍類をあたってみたが、筆者の知る限り明確に記載されているものは見当たらない。そこで、本稿ではその理由を簡単に説明しておく。

式 (3-3) を一部省略して再掲する。

$$DSpq = \{Np \cdot Nq \,/\, (Np + Nq)\} \times \| \boldsymbol{c}_p - \boldsymbol{c}_q \|^2$$

この式は ward 法におけるクラスター間の非類似度の定義式であった。$DSpq$ が最も小さくなる $p$ と $q$ をその凝集工程における凝集対象クラスターに選定する。式の右辺は 2 つの項の掛け算となっている。第一項はクラスターに含まれる要素数（$Np$ と $Nq$）だけから成る項、第二項は 2 つのクラスター重心間の平方ユークリッド距離である。第二項はクラスターに含まれる要素ベクトルの値を直接反映するものであり、クラスター間の特徴の違いの大きさを示すものである。第二項が同じ値となるクラスター $a, b$ とクラスター $c, d$ があった場合を考えてみる。この場合、非類似度の大きさを決定づけるのは第一項になる。第一項は要素数に対して単調増加の関数である。つまり、要素数が少ないクラスターペアほど非類似度が小さくなり、凝集が優先されることになる。

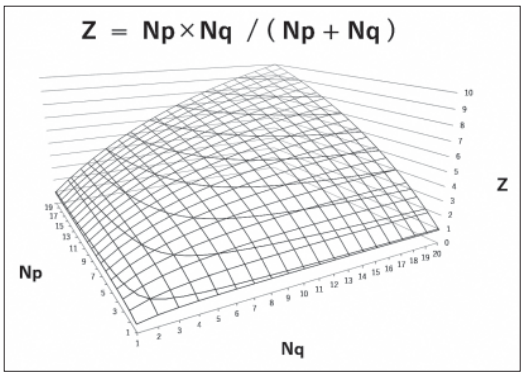


図 1　第一項のグラフ

図 1 は第一項のグラフ表示である。第一項の効果によって「要素数が小さいクラスターペアが優先的に凝集される」という特徴が生まれている。

## 4. hclust 関数の内部処理を確認してみよう

hclust 関数の ward 法関連オプションの正しい使い方について説明したが、「R の hclust 関数のドキュメントの記述と異なる説明だが本当？」とモヤモヤした感じが残っている方も多いだろう。

そこで、第 2 節で述べた結論が正しいことを検証するため､少々回りくどい方法ではあるが、筆者が独自実装した ward 法プログラムと R の hclust 関数とを比較し、内部処理を検証しておこう。独自実装プログラムは Ward[4] に完全に則り設計しており、初回の非類似度行列に第 2 節①で記載したものを用いる。この初回の非類似度行列は R 言語の dist 関数で作成したユークリッド距離行列の各要素を二乗して 1/2 倍したものと等しい。この初回の ward 非類似度行列をスタートとして､ward 法に従いクラスター凝集をすすめ、非類似度行列は式 (3-4) を用いて更新している。

第 2 節での説明が正しければ、独自プログラムによる結果と hclust 関数の ward.D オプションでの結果は一致するはずである。また、hclust 関数で ward.D2 を指定し「初回の ward 非類似度」の平方根を初回非類似度として与えた分析結果の Height 値を二乗したものと一致するはずである。

## 5. 検証実験：分析結果を比較する

分析するデータには木村・高部 [1] 及び統計センター [7] の SSDSE-2020C（家計調査 食料品目）を使った。

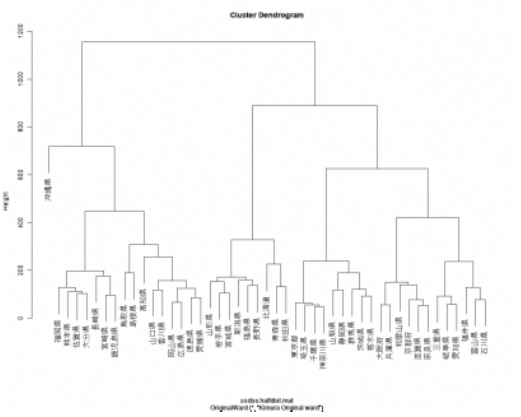


図２　独自プログラムの結果

図 2 が今回の独自実装による結果である。

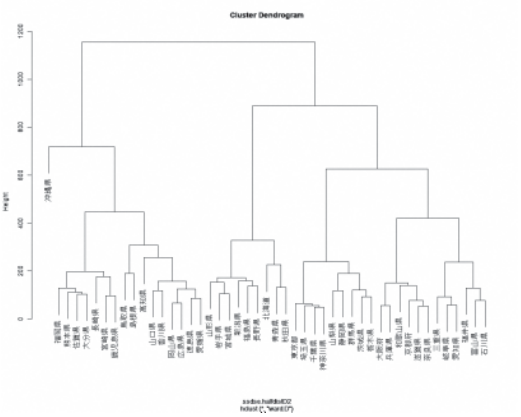


図３　ward.D 指定での hclust 関数結果

図 3 は、R の hclust 関数で ward.D オプション指定した結果である。図 2 と図 3 は、縦軸の Height 値も含めて完全に一致していることがわかる。

図 4 は ward.D2 オプションに『「初回 ward 非類似度行列」の平方根』を与えた結果である。クラスターの凝集の順番は図 2 や図 3 と一致しているが、縦軸の Height 値については前の図 2 と図 3 の Height 値の平方根をとった値になっている。

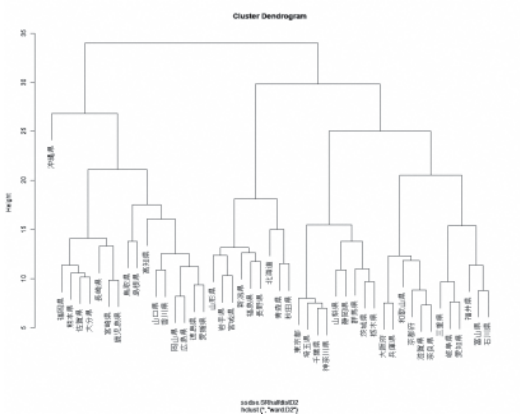


図４　ward.D2 指定の hclust 関数の結果

そこで、図 4 の Height 値を二乗したものでデンドログラムを描き直したものを図 5 に示す。

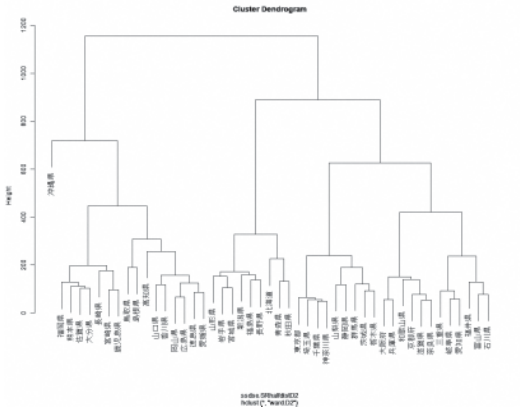


図５　図４の Height 値を二乗した分析結果

図2と図3に完全に一致していることがわかる。

なお、図 2 の横軸の都道府県名の並び順は hclust 関数の出力結果と合わせてある。

## 6. わざわざ独自実装プログラムを作った経緯

川端ほか [5] では、Murtagh and Legendre[3] を引用して説明されている。ところが R の hclust 関数のドキュメントにおいても Murtagh and Legendre[3] が引用されており、両者の主張が異なっているように読める。これには

筆者も最初は大いに混乱した。ここで改めてMurtagh and Legendre[3]の結論を確認しておこう。次の箇条書きは筆者によるMurtagh and Legendre[3]の結論の概要である。

・世の中のward法の実装アルゴリズムには、F.Murtagh[8]のコードに基づいた実装（以後Ward1）とKaufman and Rousseeuw[9]のコードに基づいた実装（以後Ward2）の2種類がある。
・Ward1とWard2では、内部で用いている非類似度が異なり、非類似度更新式も異なる。Ward1では非類似度に平方ユークリッド距離、Ward2ではユークリッド距離を用いている。
・しかしながら、Ward1とWard2アルゴリズムには同一性があり、両者の非類似度を適切に関係づけることにより同等性が確保できる。

つまり、非類似度として平方ユークリッド距離を用いているWard1こそがWard法に則った実装ということになる。Rのhclust関数は、Murtagh and Legendre[3]発表時点ではWard1実装（オプションmethod = ”ward”）のみであった。Murtagh and Legendre[3]には「今後、Rのhclust関数にもWard1とWard2の両実装の盛り込みが準備中とのことである」との記述があるが、hclust関数に後に追加されたward.Dとward.D2オプションが最終的にどう実装されたのかは当然記載されていない。もしward.D指定時にRのドキュメント通りdist関数の結果をそのまま引数dとして渡すのが正解なら、hclust内部でdを二乗し1/2倍していなければならない。

内部処理確認のためhclust関数のソースプログラムを読んでみたが、肝心の処理がFortranサブルーチンになっており直接の確認ができなかった。そこでWard[4]準拠独自プログラムを作り、hclustと比較することにしたわけである。

## 7. おわりに

Ward[4]の基準に厳密に準拠したRのhclust関数の正しい使い方を具体的に解説した。ward.Dオプション指定の際には、平方ユークリッド距離の非類似度を使用せねばならない。dist関数の出力を平方せずにそのまま渡すと、強く歪ませてしまったデータを分析していることになってしまうので注意が必要である。

なお、市販の書籍でhclust関数の使用例として初回非類似度の1/2や1/2の平方根などの係数を掛けていないものも見受けられる。この場合、結果のHeight値はward基準とは異なるためHeight値自体を評価する際には問題があるが、クラスター凝集順序や凝集結果だけを評価する場合であれば分析結果に違いは無いのでご安心を。

＊参考文献


[1] 木村敦・高部勲（2021）「家計消費データから見る日本の食料嗜好地域性 ～人文社会系知見との連携も見据えて～」『ESTRELA』No.324, pp.36-42.
[2] 大隅昇（2000）「多次元データ解析における分類手法の役割 —分けて知ることの効用と難しさ—」『ESTRELA』No.79, pp.10-20.
[3] F.Murtagh and P.Legendre (2014) “Ward’s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward’s Criterion?”, Journal of Classification, 31, pp.274-295.
[4] Joe H. Ward, Jr. (1963) “Hierarchical Grouping to Optimize an Objective Function”, Journal of the American Statistical Association, Vol. 58, No. 301(Mar., 1963), pp.236-244.
[5] 川端一光・岩間徳兼・鈴木雅之（2018）「コラム23：2種類のウォード法」『Rによる多変量解析入門（データ分析の実践と理論）』オーム社, p.313.
[6] イリチュ美佳・高木英明（2017）「分かるために分けるクラスター分析」『サービスサイエンスの事訳』（高木英明編著）, 筑波大学出版会, pp.65-116.
[7] ㈱統計センター「SSDSE（教育用標準データセット）」https://www.nstac.go.jp/SSDSE/index.html
[8] MURTAGH, F. (1985), Multidimensional Clustering Algorithms, Vienna: Physica-Verlag.
[9] KAUFMAN, L., and ROUSSEEUW, P.J. (1990), Finding Groups in Data: An Introduction to Cluster Analysis, New York: Wiley.