Nagoya J. Health, Physical Fitness, Sports Vol.42, No.1 (May 2019)
DOI: 10.18999/njhpfs.42.1.1

階層クラスタリングでクラスタ数を決定するための R プログラム

R programs for determining the number of clusters in hierarchical clustering

山　本　裕　二*　　高御堂　良　太**
千　鳥　司　浩***　　横　山　慶　子*

Yuji YAMAMOTO*　　Ryota TAKAMIDO**
Kazuhiro CHIDORI***　　Keiko YOKOYAMA*

We developed programming codes using R language to determine the optimal number of clusters in the hierarchical clustering method, which were “Upper tail method” and “Jain-Dubes method”. The developed R codes were applied to actual data sets, and we confirmed the programming codes and argued these methods to determine the number of clusters.

1　序

クラスター分析は、一組のデータをその間の類似性あるいは非類似性に基づいて、いくつかの部分集合、すなわちクラスターに分けようとするものである。統計学では多変量解析の一手法として位置づくが、データマイニングの手法として画像処理におけるパターン認識（Jain, Murty, & Flynn, 1999; Ngo, Pong, & Zhang, 2002）や神経細胞の発火パターンの検出（Kanakov, Osipov, Chan, & Kurths, 2007）、時系列データの解析（Alon, Sclaroff, Kollios, & Pavlovic, 2003）までさまざまな分野で用いられている。クラスタ分析には樹状図のような階層構造を求める階層的方法と、決められたクラスタ数ができるだけ等質になるようにする非階層的方法がある。前者の代表的な方法が Ward 法（Ward, 1963）で、後者の代表的な方法が *k*-mean 法（Forgy, 1965; MacQueen, 1967）である。

ここで非階層的な方法を用いる場合には、クラスタ数 k を事前に決定する必要がある。しかしながら探索的な研究では、クラスタ数は事前には知りえず、階層的方法を用いる場合が多い。そこではクラスタ数をどのように決定していくかが問題となり、クラスタ数の自動決定法に関して研究が行われている（志津・松田、2011；谷中・大澤、2016）。本稿では、階層的方法におけるクラスタ数を決定するために提案されている Upper Tail 法（Mojena, 1977）と Jain-Dubes 法（Davies & Bouldin, 1979; Jain & Dubes, 1988; Ngo et al., 2002）を、R のプログラムとして実装することを目的とする。

階層的方法では、クラスタを生成する方法として、最短距離法、最長距離法、群平均法、重心法、メディアン法、Ward 法がある。また、その基準となる距離には、ユークリッド距離や標準化ユークリッド距離、マハラノビス距離などがある。したがって、階層的方法ではクラスタの生成法と基準となる距離の組み合わせでさまざまな方法がある。本稿では、クラスタの生成法として Ward 法を、基準となる距離は標準化ユークリッド距離を用いる（Appendix A）。

2　クラスタ決定法

2. 1　Upper tail 法

Upper tail 法は Mojena（1977）によって提案され、統計的な停止規則を用いた階層的クラスタ分析におけるクラスタ数決定のための重要な方法である。これらの規則は、基準値 α となる分布が $N-1$ 個あるのを利用す

＊　名古屋大学総合保健体育科学センター
＊＊　名古屋大学大学院教育発達科学研究科
＊＊＊　中部学院大学
＊　Research Center of Health, Physical Fitness and Sports, Nagoya University
＊＊　Graduate School of Education and Human Development, Nagoya Univeristy
＊＊＊　Chubu Gakuin University

る。ここではクラスタ間の距離のみが基準値として用いられ、基準値の値はα_1、すなわちクラスタが２つの場合からα_{N-1}個まで取りうる。停止規則は、クラスタ数（j）が２個から始めて増加させていき、以下の条件を

$$\alpha_j \leq \bar{\alpha} + ks_\alpha \tag{1}$$

を満たすまで繰り返される。ここで、$\bar{\alpha}$とs_αは、それぞれの基準値αの分布の平均と不偏分散である。kは、αの分布の平均と不偏分散に基づき、上部棄却域を決める定数である。なお、Mojena（1977）においては、kの値は60から120までのデータ数に応じて２から４の値をとっている。

2. 2　Jain-Dubes 法

Davies and Bouldin（1979）と Jain and Dubes（1988）は、クラスタ間の距離を最大にしながら、クラスタ内の距離を最小にするクラスタ分割法を提案した。そのクラスタ分割指標$\rho(k)$は、以下のように定義されている。

$$\rho(k) = \frac{1}{k}\sum_{i=1}^{k}\max_{1\leq n\leq k}\left\{\frac{\eta_i + \eta_j}{\xi_{ij}}\right\} \tag{2}$$

そして、

$$\eta_j = \frac{1}{n_j}\sum_{i=1}^{n_i} D\left(F_i^{(j)}, \mu_j\right) \tag{3}$$

$$\xi_{ij} = D(\mu_i, \mu_j) \tag{4}$$

である。ここでμ_iはクラスタiの平均ベクトルで、$F_i^{(j)}$はクラスタjにおけるi番目のベクトルである。そして、n_jはクラスタjにおけるベクトルの数である。したがって、η_{ij}は、クラスタjにおけるクラスタ内の距離を示し、ξ_{ij}は、クラスタiとjの間のクラスタ間の距離を示す。そして、$\rho(k)$がある範囲内で最小となるkが最適なクラスタ数となる。ここでは、kの取りうる範囲は志津・松田（2011）に従い、スタージェスの公式、すなわち、$2 \leq k \leq 1 + \log_2 n$を用いた。$n$はデータ数である。

3　適用例

3. 1　少数データの場合

ここでは、Chidori and Yamamoto（2017）のデータ*1を用いて書かれた R のプログラムによって Upper tail 法（Appendix B）と Jain-Dubes 法（Appendix C）によってクラスタ数を求めた結果を示す。

Figure 1 と Table 1 に２つの方法で分析した結果を示す。Upper tail 法では、Mojena（1977）に従い、$k = 2$とした結果、αは$j = 3$、すなわちクラスタ数が３個となった。しかしながら、Table 1 に示すとおり、$j = 4$におけるα_jは$j = 3$の場合と同値であった。

また Jain-Dubes 法では、$\rho(k)$の最小値は0.501であり、これはクラスタ数５の場合であった。つまり最適とされるクラスタ数が、このデータの場合には異なる値となった。

そこで、クラスタ数が３、４、５の場合について図示

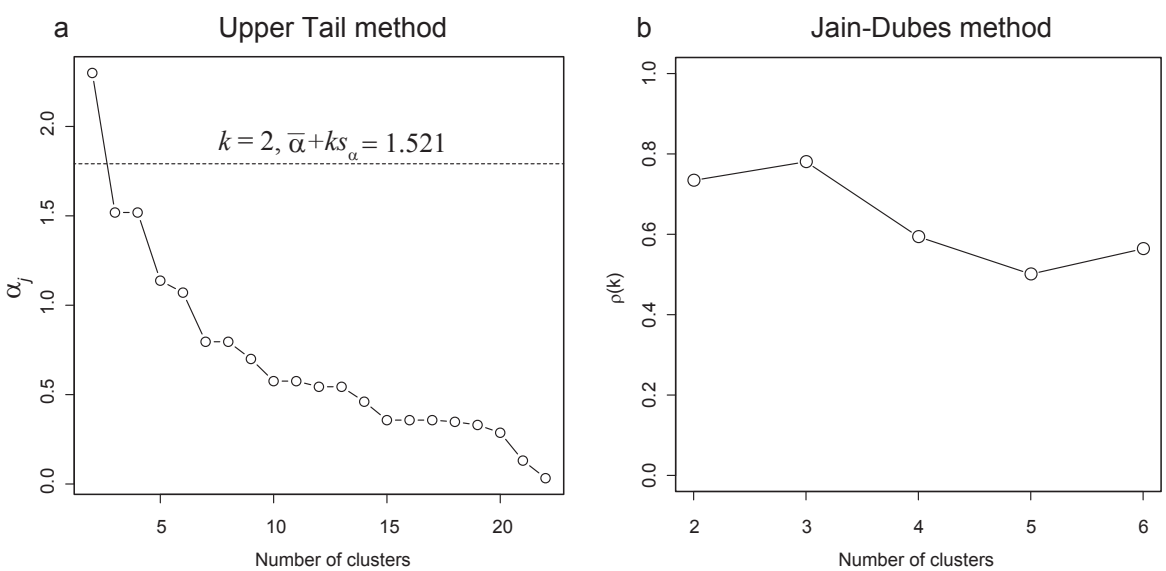

Figure 1:　**a**. Upper tail method. The dashed line shows $\bar{\alpha} + ks_\alpha$.　**b**. The Jain-Dubes method.

Table 1: Values of criterion α_j, j = 2 to 6 and cluster separation measure $\rho(k)$ at each stage

Number of clusters	2	3	4	5	6
α_j	2.299	**1.519**	1.519	1.138	1.071
$\rho(k)$	0.734	0.780	0.594	**0.501**	0.564

Bold digits show the optimal number of clusters in each method.

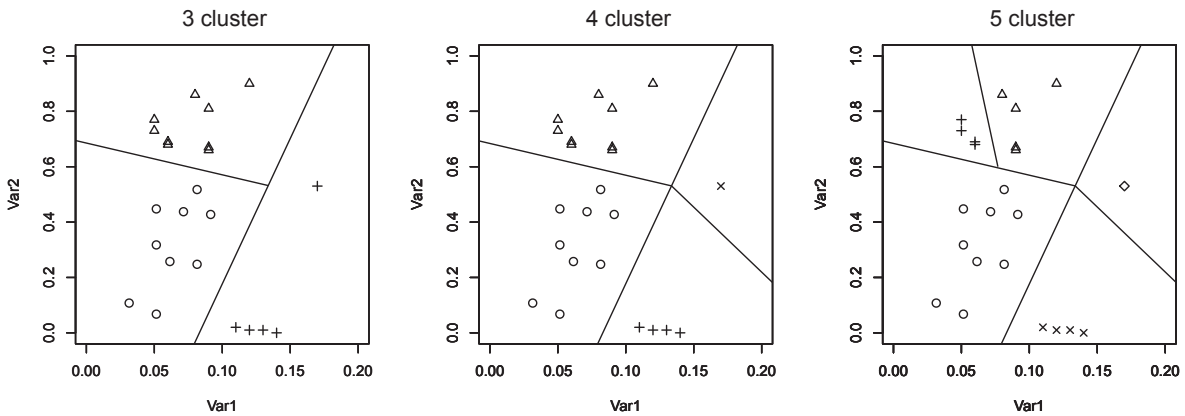

Figure 2: Separation into three (**a**), four (**b**), and five (**c**) clusters.

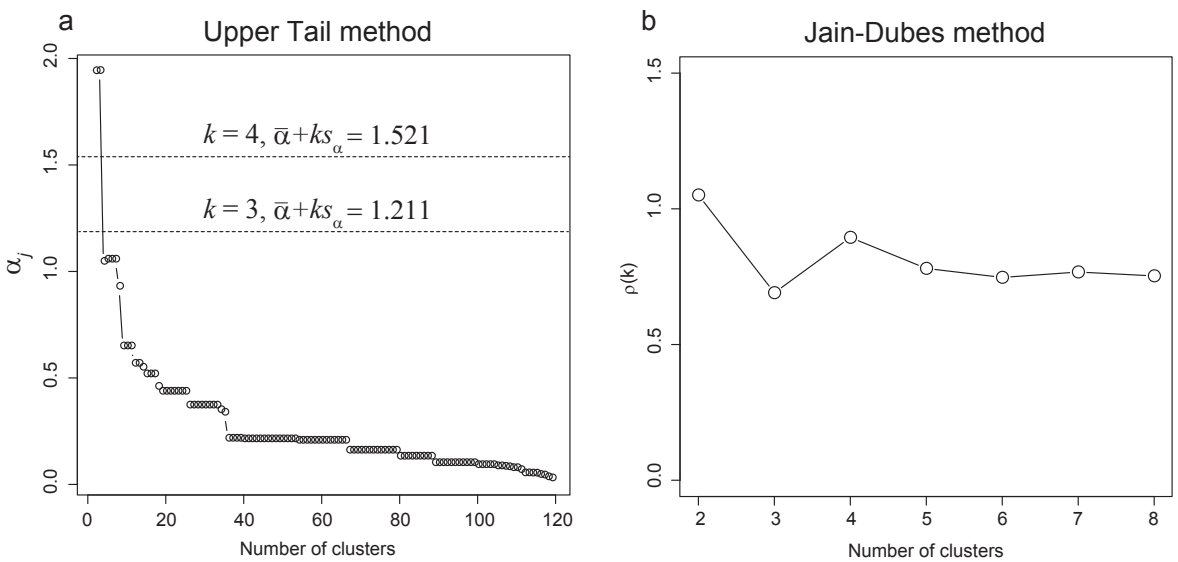

Figure 3: **a**. Upper tail method. The dashed line shows $\bar{\alpha} + \mathrm{ks}_{\alpha}$. **b**. The Jain-Dubes method.

したのが、Figure 2 である。Chidori and Yamamoto（2017）においては、両方法の結果と解釈可能性から、クラスタ数を 4 に決定した。

3. 2　比較的多数データの場合

次に比較的多数データの例として、Takamido, Yokoyama, and Yamamoto（2019）で分析された N = 120のデータ*2に適用する。Figure 3 と Table 2 に 2 つの方法で分析した結果を示す。

Upper tail 法では、Mojena（1977）に従い、k = 3, 4 とした結果、いずれも α は j = 4、すなわちクラスタ数が 4 個となった。

また Jain-Dubes 法では、$\rho(k)$ の最小値は0.691であり、これはクラスタ数 3 の場合であった。つまり最適とされ

Table 2: Values of criterion α_j, $j = 2$ *to* 8 and cluster separation measure $\rho(k)$ at each stage for larger data set.

Number of clusters	2	3	4	5	6	7	8
α_j	1.939	1.940	**1.044**	1.054	1.054	1.054	0.927
$\rho(k)$	1.051	**0.691**	0.895	0.780	0.747	0.767	0.753

Bold digits show the optimal number of clusters in each method.

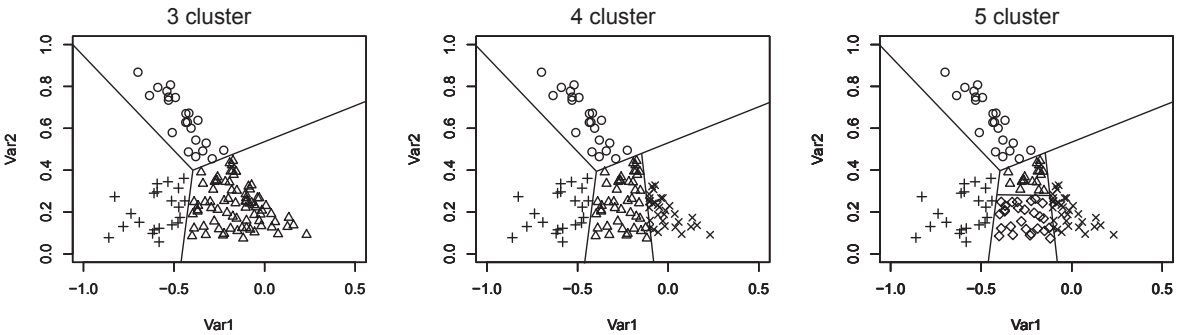

Figure 4: Separation into three (**a**), four (**b**), and five (**c**) clusters.

るクラスタ数が、このデータの場合にも異なる値となった。

そこで、クラスタ数が３、４、５の場合について図示したのが、Figure 4 である。Takamido et al.（2019）においては、両方法の結果と解釈可能性から、クラスタ数を３に決定した。

4　まとめ

本稿では、階層的クラスタ分析における、クラスタ数決定のための２つの方法、すなわち Upper tail 法と Jain-Dubes 法を、R のプログラムに実装することを目的とした。少数データと比較的多数データに適用した結果、最適なクラスタ数をこのプログラムによって決定することができた。

しかしながら、このように最適なクラスタ数を決定するために提案されている方法においても、必ずしも同じ結果が得られるとは限らず、対象としているデータを十分検討したうえでクラスタ数を決定しなければならない。つまり、クラスタ数決定のための絶対的な方法があるとは言えず、今後のアルゴリズムの開発を待ちたい。

注

*1　データは https://nuss.nagoya- u.ac.jp/s/047jhGUtgXBd5Se

*2　データは https://nuss.nagoya- u.ac.jp/s/PlZ6F6ftyaGO8jq

文　献

Alon, J., Sclaroff, S., Kollios, G., & Pavlovic, V. (2003). Discovering clusters in motion time-series data. *Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, 375–381.

Chidori, K. & Yamamoto, Y. (2017). Effects of the lateral amplitude and regularity of upper body fluctuation on step time variability evaluated using return map analysis. *PLoS ONE*, **12**, e0180898.

Davies, D. L. & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-1**, 224–227.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. *Biometrics*, **21**, 768–769.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice-Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, **31**, 264–323.

Kanakov, O. I., Osipov, G. V., Chan, C.-K., & Kurths, J. (2007). Cluster synchronization and spatiotemporal dynamics in networks of oscillatory and excitable Luo-Rudy cells. *Chaos*, **17** (015111), 1–8.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 281–297.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation. *The Computer Journal*, **20**, 359–363.

Ngo, C.-W., Pong, T.-C., & Zhang, H.-J. (2002). On clustering and retrieval of video shots through temporal slices analysis. *IEEE Transactions on Multimedia*, **4**, 446–458.

志津綾香・松田眞一（2011）．クラスター分析におけるクラスター数自動決定法の比較．『アカデミア　情報理工学

編』, **11**, 17－34.
Takamido, R., Yokoyama, K., & Yamamoto, Y. (2019). Task constraints and stepping movement of fast-pitch softball hitting. *PLoS ONE*, **14**, e0212997.
Ward, Jr. J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, **58**, 236–244.
谷中瞳・大澤幸生（2016）. 特許文献を利用した技術課題の抽象化方法の検討. 『人工知能学会第30回大会』, 1J3–2.

Appendix A　階層的クラスタ分析

```
#　データ読み込み
d1<-read.table("demo1.txt",head=T,sep="\t")
#　データの標準化
d2<-scale(d1[,c("Var1","Var2")])
#　距離行列の作成
rownames(d1) <- d1$Sub
dist1 <- dist(d2, method="euclidean")
#　階層的クラスタ分析
demo.clust <- hclust(dist1^2, method="ward.D")
plot(demo.clust)
```

Appendix B　Upper tail 法

```
#　k は 2 とする
constK <- 2
alpha <- NULL
#　クラスタ数 2 個から（データ数－ 1）個まで検討する
maxk <- (nrow(d2)-1)
#　alpha の計算
for(nk in makx:2){
    memb <- cutree(demo.clust, k=nk)
    #　クラスタの重心（平均ベクトル）
    cent<-NULL
    for(i in 1:nk){
    cent<-rbind(cent, colMeans(d2[memb == i, , drop=FALSE]))
    }
    #　クラスタ間の距離（重心間の距離）
    cent_dist<-dist(cent, method="euclidean")
    alpha <- rbind(alpha, min(cent_dist))

    cat("nk = ",nk,"\t");cat("Alpha = ",alpha,"\n")
}
Upper <- mean(alpha)+constK*sqrt(var(alpha))
n.cluster<-matrix((nrow(d2)-1):2, ncol=1)
alpha <- cbind(n.cluster, alpha)
#　描画
plot(alpha, type="b")
par(new=T)
abline(h=Upper)
```

Appendix C　Jain-Dubes 法

```
pk <- NULL
#　スタージェスの公式（Sturges' rule）を用いる
maxk <- 1+round(log2(nrow(d2)))
#　rho(k) の計算
for(nk in 2:maxk){
    memb <- cutree(demo.clust, k=nk)
    # クラスタの重心（平均ベクトル）
    cent<-NULL
    for(i  in 1:nk){
    cent<-rbind(cent, colMeans(d2[memb == i, , drop=FALSE]))
    }
    #　クラスタ間の距離（重心間の距離）xi
    cent_dist<-as.matrix(dist(cent, method="euclidean"))
    #　変数名に変数を使って値を代入
```

```
    for(i  in 1:nk){
        assign(sprintf("d2.%d", i), d2[memb == i, , drop=FALSE])
    }
    #   クラスタ内の平均距離 eta
    for(i  in 1:nk){
        #   変数名に変数を
        row <- eval(parse(text=paste("nrow(d2.", sprintf("%d)",i),  sep="")))
        tmp <- 0
        for(j in 1:row){
            tmp <- tmp+eval(parse(text=paste(sprintf(
            "mean(sqrt(sum((d2.%d[%d,]-cent[%d,])^2)))/row",i,j,i))))
        }
        assign(sprintf("eta.%d",i),tmp)
    #   確認 eta
        cat("i = ",i,"\t");cat("eta = ",tmp,"\n")
    }
    tmp3=NULL
    for(i in 1:nk){
        tmp2=NULL
        for(j in 1:nk){
            tmp1=NULL
            tmp1 <- rbind(tmp1, eval(parse(text=paste(sprintf(
            "eta.%d + eta.%d",i,j,sep=""))))
            tmp2 <- rbind(tmp2, eval(parse(text=paste(sprintf(
            "if (i!=j){tmp1/cent_dist[%d,%d]}",i,j,sep=""))))
        }
        tmp3 <- rbind(tmp3, max(tmp2))
        cat("max = ",tmp3,"\n")
    }
    pk <- rbind(pk, mean(tmp3))
    cat("k = ",nk,"\t");cat("pk = ",pk,"\n")
}
n.cluster<-matrix(2:nk, ncol=1, nrow=nk-1)
pk<-cbind(n.cluster,pk)
#   描画
plot(pk, type="b", ylim=c(0,1), xlab="Numbers of clusters", ylab="p(k)")
```