クラスタ分析

基本的な考え方と階層的方法

村田 昇

講義概要

- **第 1 日 : 基本的な考え方と階層的方法**
- 第 2 日 : 非階層的方法と分析の評価

クラスタ分析の考え方

クラスタ分析

- クラスタ分析 (**cluster analysis**) の目的
  個体の間に隠れている**集まり=クラスタ**を個体間の “距離” にもとづいて発見する方法
- 個体間の類似度・距離 (非類似度) を定義
  - 同じクラスタに属する個体どうしは似通った性質
  - 異なるクラスタに属する個体どうしは異なる性質
- さらなるデータ解析やデータの可視化に利用
- 教師なし学習の代表的な手法の一つ

クラスタ分析の考え方

- 階層的方法
  - データ点およびクラスタの間に **距離** を定義
  - 距離に基づいてグループ化
    - 近いものから順にクラスタを **凝集**
    - 近いものが同じクラスタに残るように **分割**
- 非階層的方法
  - クラスタの数を事前に指定
  - クラスタの **集まりの良さ** を評価する損失関数を定義
  - 損失関数を最小化するようにクラスタを形成

事例

- 総務省統計局より取得した都道府県別の社会生活統計指標の一部
  - 総務省 https://www.e-stat.go.jp/SG1/estat/List.do?bid=000001083999&cycode=0
  - データ https://noboru-murata.github.io/multivariate-analysis/data/japan_social.csv

`Pref` ： 都道府県名
`Forest` ： 森林面積割合（%）2014 年
`Agri` ： 就業者 1 人当たり農業産出額（販売農家）（万円）2014 年
`Ratio` ： 全国総人口に占める人口割合（%）2015 年
`Land` ： 土地生産性（耕地面積 1 ヘクタール当たり）（万円）2014 年
`Goods` ： 商業年間商品販売額［卸売業＋小売業］（事業所当たり）（百万円）2013 年

- データの内容

Name	Forest	Agri	Ratio	Land	Goods
Hokkaido	67.9	1150.6	4.23	96.8	283.3
Aomori	63.8	444.7	1.03	186	183
Iwate	74.9	334.3	1.01	155.2	179.4
Miyagi	55.9	299.9	1.84	125.3	365.9
Akita	70.5	268.7	0.81	98.5	153.3
Yamagata	68.7	396.3	0.88	174.1	157.5
Fukushima	67.9	236.4	1.51	127.1	184.5
Ibaraki	31	479	2.3	249.1	204.9
Tochigi	53.2	402.6	1.55	199.6	204.3
Gumma	63.8	530.6	1.55	321.6	270

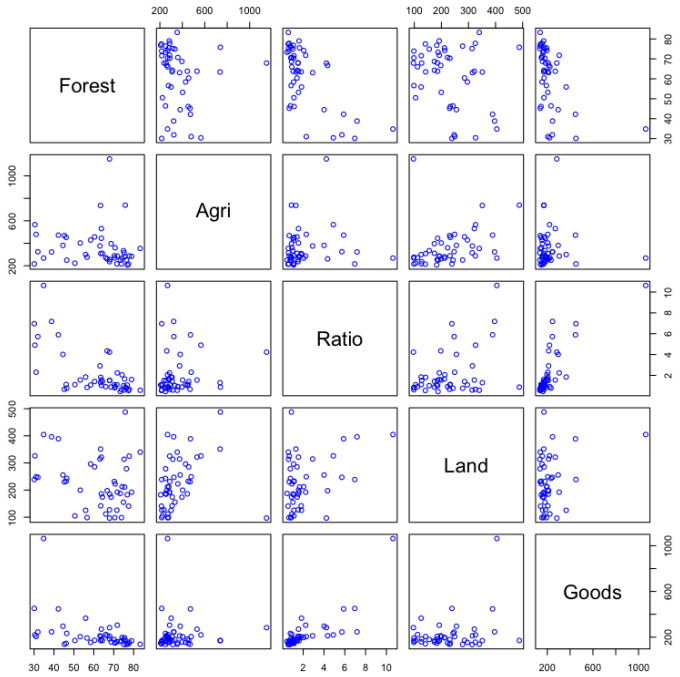

図 1: データの散布図

階層的方法

凝集的クラスタリング

1. データ・クラスタ間の距離を定義する
   - データ点とデータ点の距離
   - クラスタとクラスタの距離
2. データ点およびクラスタ間の距離を求める

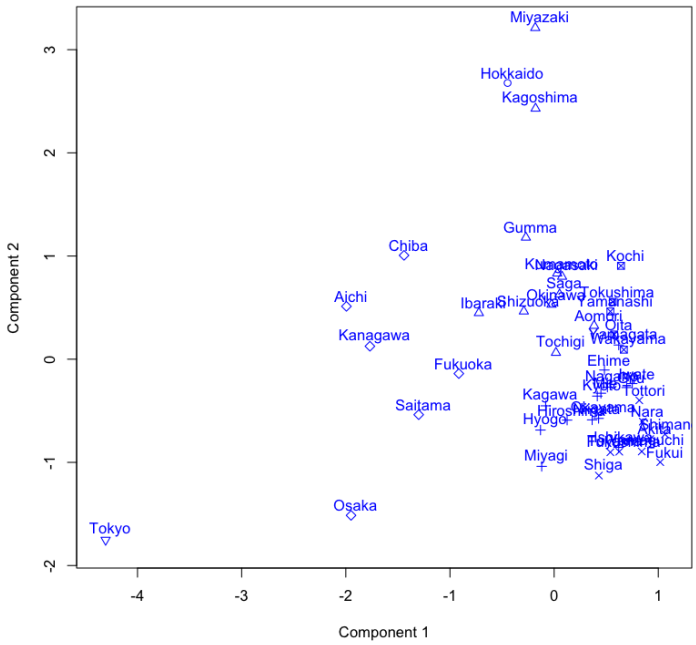

図 2: 主成分得点の散布図

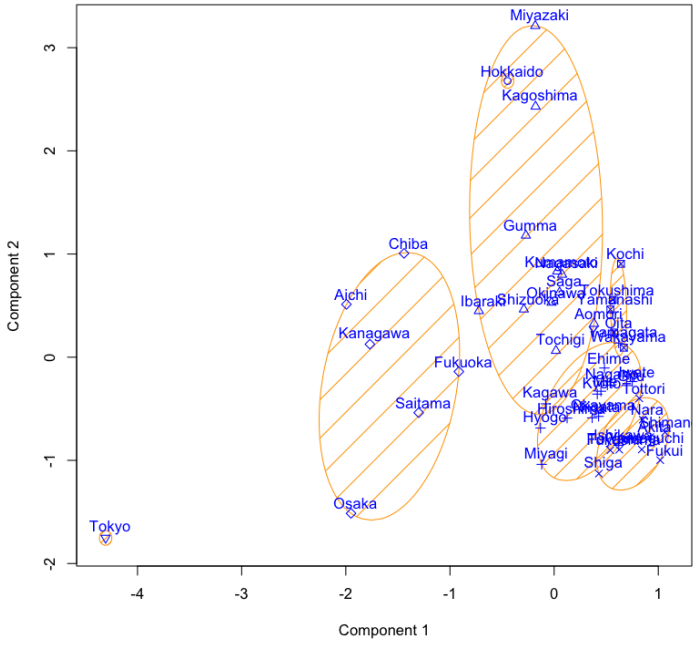

図 3: 散布図上のクラスタ構造 (クラスタ分析の概念図)

3. 最も近い 2 つを統合し新たなクラスタを形成する
    - データ点とデータ点
    - データ点とクラスタ
    - クラスタとクラスタ
4. クラスタ数が 1 つになるまで 2-3 の手続きを繰り返す

事例

- 社会生活統計指標の一部 (関東)

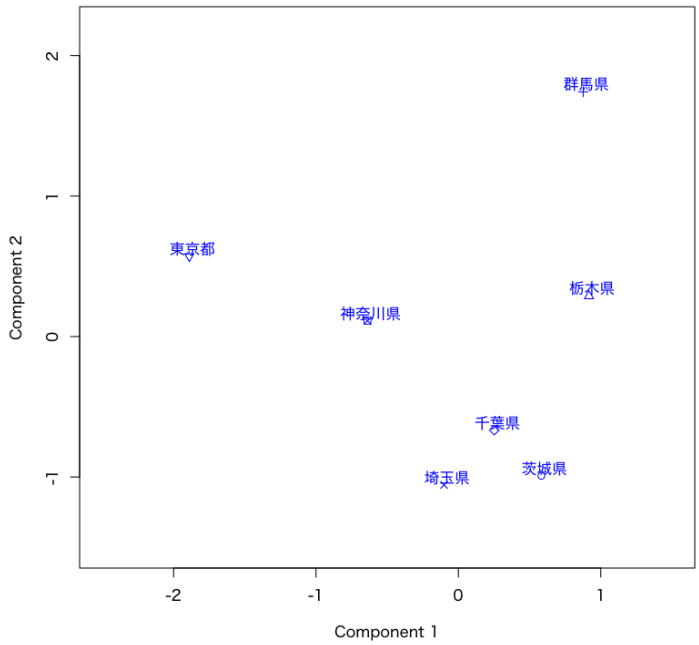

図 4: 凝集的クラスタリング

データ間の距離

データ間の距離

- データ : 変数の値を成分としてもつベクトル

$$\boldsymbol{x} = (x_1, \ldots, x_d)^\mathsf{T}, \boldsymbol{y} = (y_1, \ldots, y_d)^\mathsf{T} \in \mathbb{R}^d$$

- 距離 : $d(\boldsymbol{x}, \boldsymbol{y})$
- 代表的なデータ間の距離
    - Euclid 距離 (ユークリッド ; Euclidean distance)
    - Manhattan 距離 (マンハッタン ; Manhattan distance)
    - Minkowski 距離 (ミンコフスキー ; Minkowski distance)

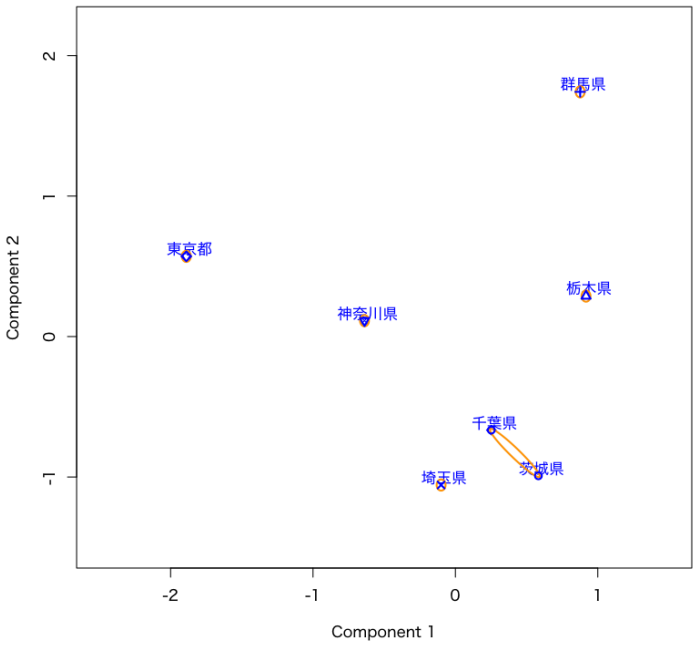

図 5: クラスタリングの手続き (その 1)

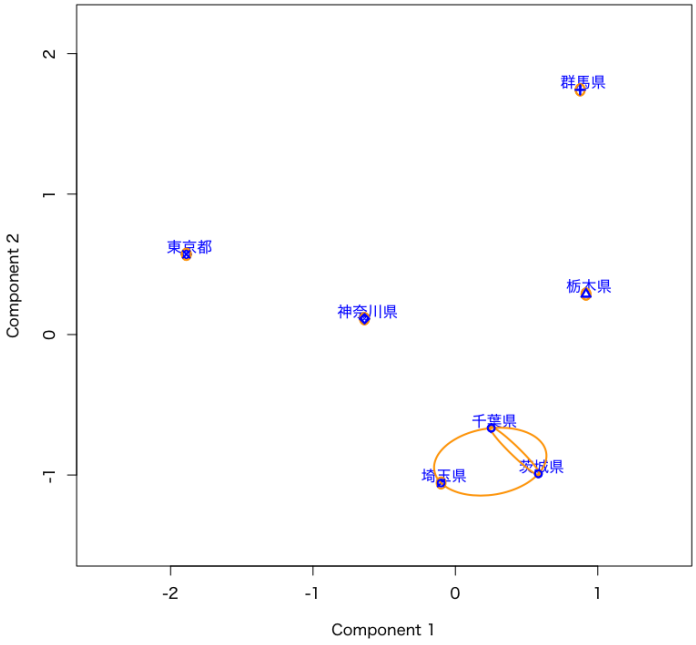

図 6: クラスタリングの手続き (その 2)

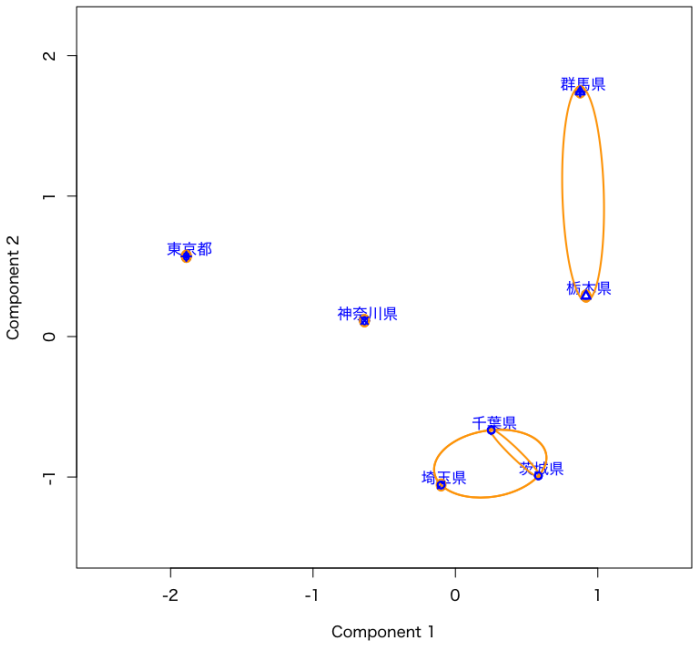

図 7: クラスタリングの手続き (その 3)

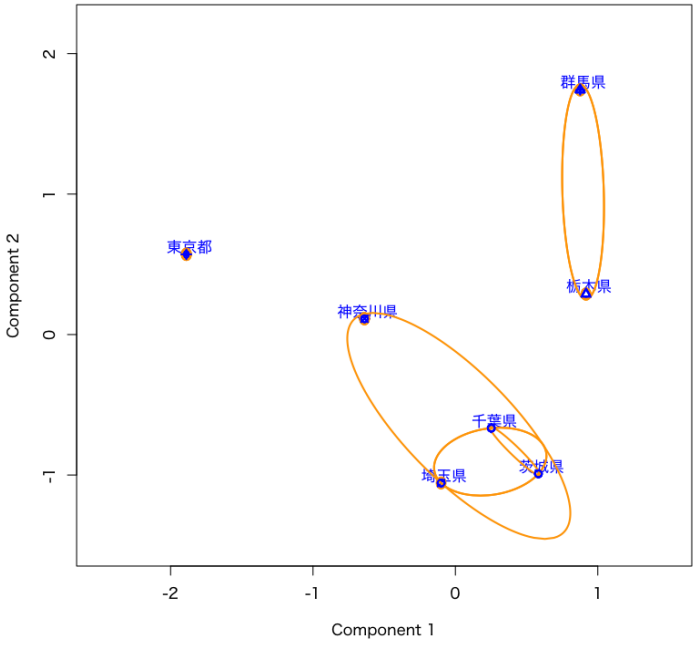

図 8: クラスタリングの手続き (その 4)

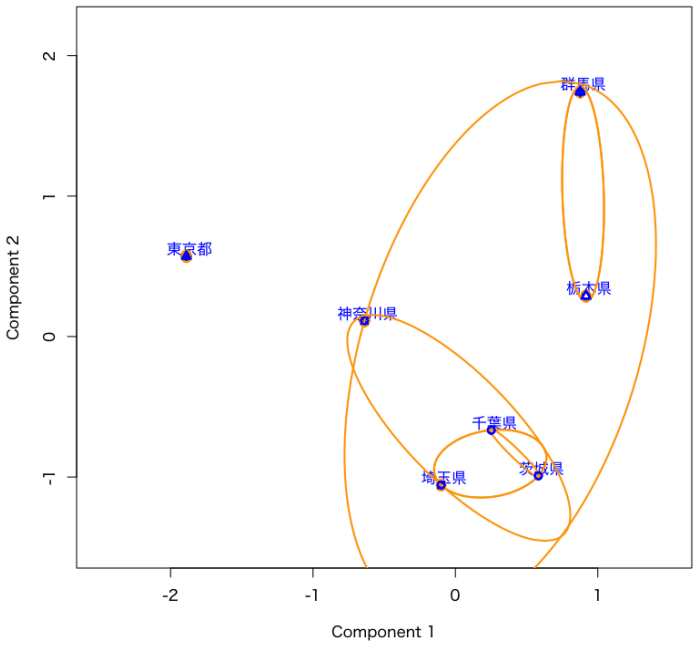

図 9: クラスタリングの手続き (その 5)

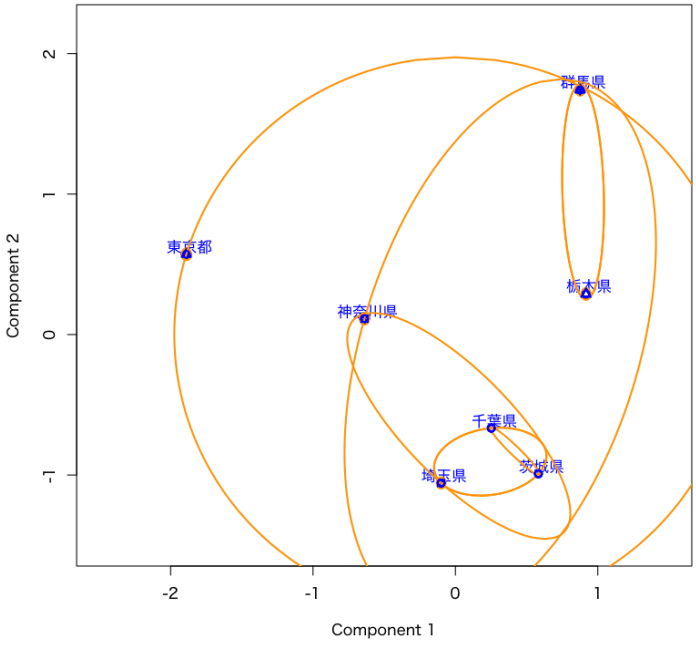

図 10: クラスタリングの手続き (その 6)

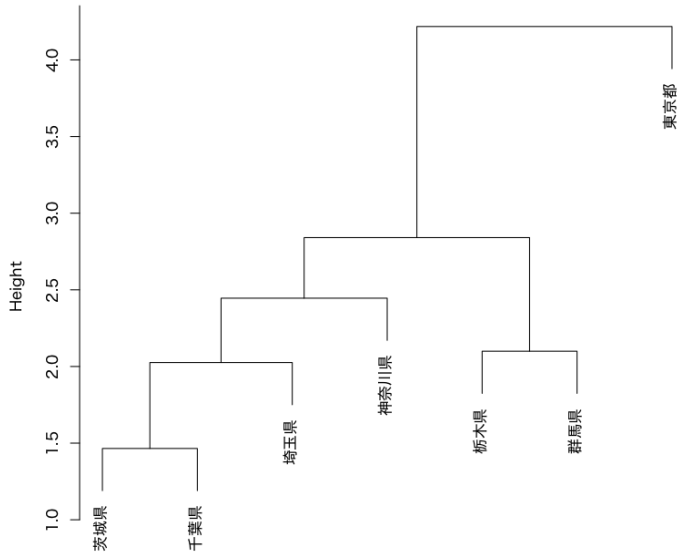

図 11: デンドログラムによるクラスタ構造の表示

Euclide 距離

- 最も一般的な距離
- 各成分の差の 2 乗和の平方根 (2 ノルム)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}$$

Manhattan 距離

- 後述する Minkowski 距離の $p = 1$ の場合
- 格子状に引かれた路に沿って移動するときの距離

$$d(\boldsymbol{x}, \boldsymbol{y}) = |x_1 - y_1| + \cdots + |x_d - y_d|$$

Minkowski 距離

- Euclid 距離を p 乗に一般化した距離
- 各成分の差の p 乗和の p 乗根 (p-ノルム)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left\{|x_1 - y_1|^p + \cdots + |x_d - y_d|^p\right\}^{1/p}$$

その他の距離

- 類似度や乖離度などデータ間に自然に定義されるものを用いることは可能
  - 語句の共起 (同一文書に現れる頻度・確率)
  - 会社間の取引量 (売上高などで正規化が必要)
- 擬似的な距離でもアルゴリズムは動く

実習

R：関数 dist()

- データフレームを用いた基本的な計算方法

```
### 距離の計算，返値は dist class (特殊なベクトル)
dst <- dist(x, method = "euclidean", diag = FALSE, upper = FALSE)
## x: データフレーム
## method: 距離 (標準はユークリッド距離，他は"manhattan","minkowski"など)
## diag: 対角成分を持たせるか
## upper: 上三角成分を持たせるか (標準は下三角成分のみ)

### 距離行列全体の表示
dst # または print(dst)

### 特定の成分の取得
as.matrix(dst)[i, j]
## i,j: 行・列の指定 (数値ベクトル，データフレームの行名)
```

R：関数 cluster::daisy()

- cluster: クラスタ分析用のパッケージ
- 関数 dist() とほぼ同様

```
### パッケージの読み込み (標準で含まれているので install は不要)
library(cluster) # require(cluster)
### 距離の計算，返値は dissimilarity class (dist とほぼ互換)
dsy <- daisy(x, metric = "euclidean", stand = FALSE)
## x: データフレーム
## metric: 距離 (標準はユークリッド距離，他は"manhattan"など)
## stand: 正規化 (平均と絶対偏差の平均による) の有無

### 距離行列全体の表示
dsy # または print(dsy)
### 特定の成分の取得
as.matrix(dsy)[i, j]
## i,j: 行・列の指定 (数値ベクトル，データフレームの行名)
```

練習問題

- 都道府県別の社会生活統計指標を用いて以下を確認しなさい

```
### データの読み込み
js_data <- read.csv(file="data/japan_social.csv", row.names=1)
```

  - 正規化せずにユークリッド距離とマンハッタン距離の計算を行いなさい
  - 正規化して上記と同様の計算を行いなさい
  - 関東の都県同士の距離を表示しなさい (daisy による正規化を用いなさい)
  - 大阪と四国の間の距離を表示しなさい
  - ユークリッド距離とマンハッタン距離の散布図を描き比較しなさい

クラスタ間の距離

クラスタ間の距離

- クラスタ：いくつかのデータ点からなる集合

$$C_a = \{\boldsymbol{x}_i | i \in \Lambda_a\},\ C_b = \left\{\boldsymbol{x}_j | j \in \Lambda_b\right\},\quad C_a \cap C_b = \emptyset$$

- 2 つのクラスタ間の距離 : $D(C_a, C_b)$
  - データ点の距離から陽に定義する方法
  - クラスタの統合にもとづき再帰的に定義する方法
- 代表的なクラスタ間の距離
  - 最短距離法 (単連結法 ; single linkage method)
  - 最長距離法 (完全連結法 ; complete linkage method)
  - 群平均法 (average linkage method)

最短距離法

- 最も近い対象間の距離を用いる方法

$$D(C_a, C_b) = \min_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_b} d(\boldsymbol{x}, \boldsymbol{y})$$

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \min\left\{D(C_a, C_c), D(C_b, C_c)\right\}$$

最長距離法

- 最も遠い対象間の距離を用いる方法

$$D(C_a, C_b) = \max_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_b} d(\boldsymbol{x}, \boldsymbol{y})$$

- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \max\left\{D(C_a, C_c), D(C_b, C_c)\right\}$$

群平均法

- 全ての対象間の平均距離を用いる方法

$$D(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{\boldsymbol{x} \in C_a,\, \boldsymbol{y} \in C_b} d(\boldsymbol{x}, \boldsymbol{y})$$

  - ただし $|C_a|, |C_b|$ はクラスタ内の要素の数を表す
- 統合前後のクラスタ間の関係

$$D(C_a + C_b, C_c) = \frac{|C_a| D(C_a, C_c) + |C_b| D(C_b, C_c)}{|C_a| + |C_b|}$$

距離計算に関する注意

- データの性質に応じて距離は適宜使い分ける
  - データ間の距離の選択
  - クラスタ間の距離の選択
- 変数の正規化は必要に応じて行う
  - 物理的な意味合いを積極的に利用する場合はそのまま
  - 単位の取り方などによる分析の不確定性を避ける場合は平均 0，分散 1 に正規化
- データの性質を鑑みて適切に前処理

実習

R：関数 `hclust()`

- 距離行列を用いた階層的クラスタリング

```
hclst <- hclust(d, method = "complete")
## d: 距離行列
## method: 分析法 (標準は最長距離法，他は"single","average"など)

### 系統樹の表示 (一般的な plot のオプションが利用可能)
plot(hclst)

### クラスタの分割
cutree(tree = hclst, k = NULL, h = NULL)
## tree: hclust の結果を指定
## k: クラスタ数を指定して分割
## h: クラスタ距離を指定して分割

### クラスタの分割表示 (cutree とほぼ同様のオプション)
rect.hclust(tree = hclst, k = NULL, h = NULL)
```

練習問題

- 都道府県別の社会生活統計指標を用いて以下の分析を行いなさい
  - 平均 0，分散 1 に正規化したデータのユークリッド距離を用いて，群平均法による階層的クラスタリングを行いなさい
  - クラスタ数を 5 つとして分割を行いなさい

R：関数 `cluster::agnes()`

- `cluster` パッケージによる階層的クラスタリング

```
agns <- agnes(x, metric = "euclidean", stand = FALSE,
              method = "average")
## x: データフレーム，または距離行列
## metric: 距離 (標準はユークリッド距離，他は"manhattan"など)
## stand: 正規化 (平均と絶対偏差の平均による) の有無
## method: 分析法 (標準は群平均法，他は"single","complete"など)

### 系統樹の表示 (一般的な plot のオプションが利用可能)
plot(agns, which.plots=2)
## which.plots=1 は評価の際に利用
```

- 関数 `cutree()`, `rect.hclust()` も利用可能

R：関数 cluster::clusplot()

- 2 次元でのクラスタ表示

```
clusplot(x, clus, stand =FALSE,
         lines = 2, shade = FALSE, labels= 0,
         col.p = "dark green", col.txt = col.p, col.clus = 5)
## x: データフレーム
## clus: クラスタ分割
## stand: 正規化の有無
## lines: クラスタ間の繋がりの表示 (0: 無, 1: 外, 2: 中心)
## shade: 網掛けの有無
## labels: ラベルの表示 (0: 無, 2: データとクラスタ, 3: データ, 4: クラスタ, など)
## col.p/txt/clue: データ点・文字・クラスタの色指定
```

データセットの準備

- Web アンケート (都道府県別好きなおむすびの具)
  - 「ごはんを食べよう国民運動推進協議会」(平成 30 年解散)
    (閉鎖) http://www.gohan.gr.jp/result/09/anketo09.html
  - データ https://noboru-murata.github.io/multivariate-analysis/data/omusubi.csv

- アンケート概要 (Q2 の結果を利用)

【応募期間】2009 年 1 月 4 日〜2009 年 2 月 28 日
【応募方法】インターネット、携帯ウェブ

【内　　容】
Q1. おむすびを最近 1 週間に、何個食べましたか？
　　そのうち市販のおむすびは何個でしたか？
Q2. おむすびの具では何が一番好きですか？
　　A. 梅 B. 鮭 C. 昆布 D. かつお E. 明太子 F. たらこ G. ツナ H. その他
Q3. おむすびのことをあなたはなんと呼んでいますか？
　　A. おにぎり B. おむすび C. その他
Q4. おむすびといえば、どういう形ですか？
　　A. 三角形 B. 丸形 C. 俵形 D. その他

【回答者数】

男性	9,702 人	32.0%
女性	20,616 人	68.0%
総数	30,318 人	100.0%

練習問題

- 上記のデータを用いて以下の分析を行いなさい

```
### データの読み込み
om_data <- read.csv(file="data/omusubi.csv", row.names=1)
```

  - Hellinger 距離を用いて距離行列を作成しなさい

    $\boldsymbol{p}, \boldsymbol{q}$ を確率ベクトルとして定義される確率分布の間の距離

$$d_{hel}(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{\sqrt{2}} d_{euc}(\sqrt{\boldsymbol{p}}, \sqrt{\boldsymbol{q}})$$

  - 群平均法による階層的クラスタリングを行いなさい
  - クラスタ数を定めて 2 次元でのクラスタ表示を作成しなさい

次週の予定

- 第 1 日 : 基本的な考え方と階層的方法
- **第 2 日 : 非階層的方法と分析の評価**