クラスタ分析

非階層的方法と分析の評価

村田 昇

講義の内容

- 第 1 回 : クラスタ分析の考え方と階層的方法
- **第 2 回 : 非階層的方法と分析の評価**

クラスタ分析の復習

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

階層的クラスタリング

- 凝集的手続き
  1. データ・クラスタ間の距離を定義
     - データ点間の距離
     - クラスタ間の距離
  2. データ点およびクラスタ間の距離を計算
  3. 最も近い 2 つを統合し新たなクラスタを形成
  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

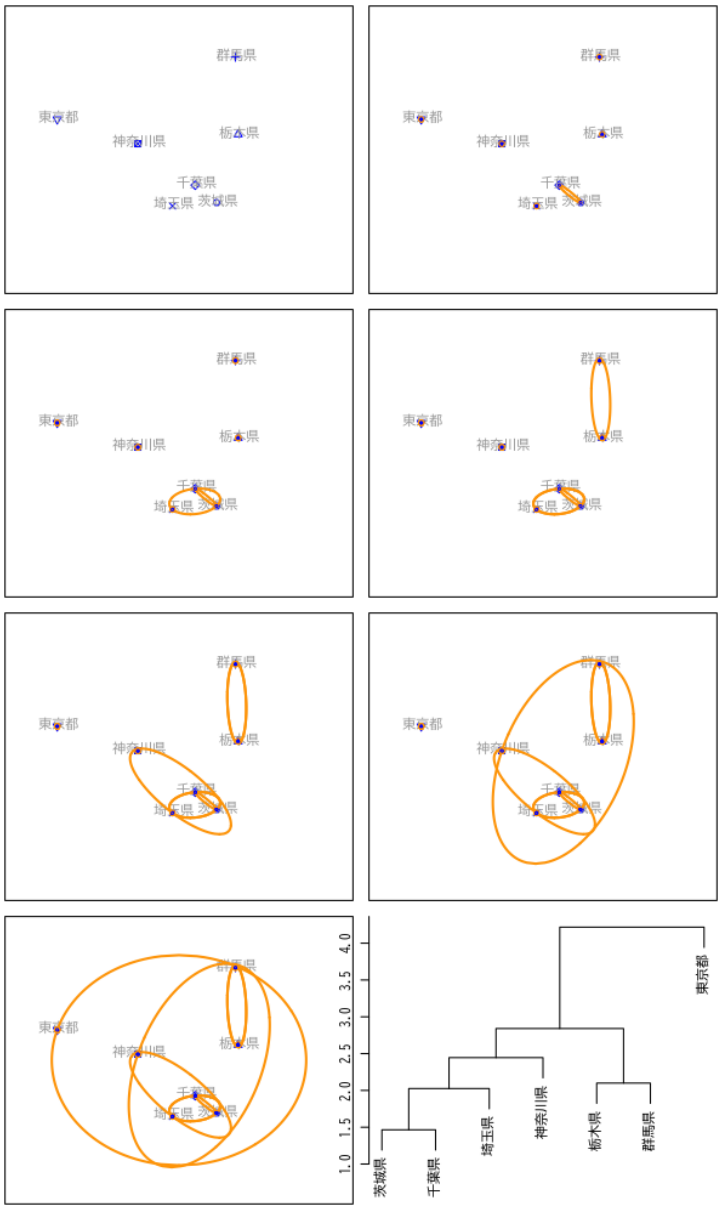
群馬県
東京都
神奈川県
栃木県
千葉県
埼玉県
茨城県
4.0
3.5
3.0
2.5
2.0
1.5
1.0
東京都
神奈川県
栃木県
群馬県
埼玉県
千葉県
茨城県

図 1: 凝集的手続きの例

非階層的方法

非階層的方法の手続き

- 対象の変数 : $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^\mathsf{T}$ (d 次元)
- 観測データ : n 個の個体の組

$$\{\boldsymbol{x}_i\}_{i=1}^n = \{(x_{i1}, x_{i2}, \ldots, x_{id})^\mathsf{T}\}_{i=1}^n$$

- 個体とクラスタの対応 C を推定

$$C(i) = (\text{個体 } i \text{ が属するクラスタ番号})$$

  - 対応 C の **全体の良さ** を評価する損失関数を設定
  - 観測データ $\{\boldsymbol{x}_i\}_{i=1}^n$ に最適な対応 $\{C(i)\}_{i=1}^n$ を決定

k-平均法の損失関数

- クラスタの個数 k を指定
- 2 つの個体 i, i' の **近さ=損失** を距離の二乗で評価

$$\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 = \sum_{j=1}^d (x_{ij} - x_{i'j})^2$$

- 損失関数 $W(C)$: クラスタ内の平均の近さを評価

$$W(C) = \sum_{l=1}^k \frac{1}{n_l} \sum_{i:C(i)=l} \sum_{i':C(i')=l} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2$$

k-平均法の性質

- クラスタ l に属する個体の平均

$$\bar{\boldsymbol{x}}_l = \frac{1}{n_l} \sum_{i:C(i)=l} \boldsymbol{x}_i$$

- 損失関数 $W(C)$ の等価な表現

$$W(C) = 2 \sum_{l=1}^k \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2$$

- 最適な対応 C : クラスタ内変動の総和が最小

演習

問題

- 以下の問に答えなさい
  - 損失関数 $W(C)$ の等価な表現を示しなさい

$$
\begin{aligned}
W(C) &= \sum_{l=1}^{k} \frac{1}{n_l} \sum_{i:C(i)=l} \sum_{i':C(i')=l} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 \\
&= 2 \sum_{l=1}^{k} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2
\end{aligned}
$$

  - 以下の $\hat{\boldsymbol{\mu}}$ を求めなさい

$$
\hat{\boldsymbol{\mu}} = \arg\min_{\mu} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2
$$

解答例

- 対称性に注意して標本平均のまわりで展開

$$
\begin{aligned}
&\sum_{l=1}^{k} \frac{1}{n_l} \sum_{i:C(i)=l} \sum_{i':C(i')=l} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 \\
&= \sum_{l=1}^{k} \frac{1}{n_l} \sum_{i:C(i)=l} \sum_{i':C(i')=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l + \bar{\boldsymbol{x}}_l - \boldsymbol{x}_{i'}\|^2 \\
&= \sum_{l=1}^{k} \frac{2}{n_l} \sum_{i:C(i)=l} \sum_{i':C(i')=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2 \\
&\quad - \sum_{l=1}^{k} \frac{2}{n_l} \sum_{i:C(i)=l} \sum_{i':C(i')=l} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l)^{\mathsf{T}} (\boldsymbol{x}_{i'} - \bar{\boldsymbol{x}}_l)
\end{aligned}
$$

- 中心化したデータの標本平均が 0 であることを利用

$$
\begin{aligned}
&= 2 \sum_{l=1}^{k} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2 \\
&\quad - \sum_{l=1}^{k} \frac{2}{n_l} \sum_{i:C(i)=l} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l)^{\mathsf{T}} \sum_{i':C(i')=l} (\boldsymbol{x}_{i'} - \bar{\boldsymbol{x}}_l) \\
&= 2 \sum_{l=1}^{k} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2
\end{aligned}
$$

- 以下の不等式が成立

$$
\begin{aligned}
\sum_{i:C(i)=l} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 &= \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l + \bar{\boldsymbol{x}}_l - \boldsymbol{\mu}\|^2 \\
&= \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2 + \sum_{i:C(i)=l} \|\bar{\boldsymbol{x}}_l - \boldsymbol{\mu}\|^2 \\
&\quad + 2 \sum_{i:C(i)=l} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l)^{\mathsf{T}} (\bar{\boldsymbol{x}}_l - \boldsymbol{\mu}) \\
&= \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2 + n_l \|\bar{\boldsymbol{x}}_l - \boldsymbol{\mu}\|^2 \\
&\geq \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2
\end{aligned}
$$

- 等号の成立の条件より

$$
\hat{\boldsymbol{\mu}} = \arg\min_{\mu} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 = \bar{\boldsymbol{x}}_l
$$

クラスタの標本平均を中心とすればよい

近似的な最適化

クラスタ対応の最適化

- 最適化 : 損失関数 $W(C)$ を最小とする C を決定
- 貪欲な C の探索
  - 原理的には全ての値を計算すればよい
  - 可能な C の数 : k^n 通り (有限個のパターン)
  - サンプル数 n が小さくない限り実時間での実行は不可能
- 近似的な C の探索
  - いくつかのアルゴリズムが提案されている
  - 基本的な考え方 : **Lloyd-Forgy のアルゴリズム**

    標本平均と変動の平方和の性質を利用

$$
\bar{\boldsymbol{x}}_l = \arg\min_{\mu} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2 \quad (\text{クラスタ } l \text{ の標本平均})
$$

Lloyd-Forgy のアルゴリズム

1. クラスタ中心の初期値 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ を与える
2. 各データの所属クラスタ番号 $C(i)$ を求める

$$
C(i) = \arg\min_{l} \|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|
$$

3. 各クラスタ中心 $\boldsymbol{\mu}_l$ $(l = 1, 2, \ldots, k)$ を更新する

$$
\boldsymbol{\mu}_l = \frac{1}{n_l} \sum_{i:C(i)=l} \boldsymbol{x}_i, \quad n_l = |\{\boldsymbol{x}_i | C(i) = l\}|
$$

4. 中心が変化しなくなるまで 2,3 を繰り返す

アルゴリズムの性質

- 結果は **確率的**
  - 初期値 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ に依存
  - アルゴリズムの成否は確率的なため，最適解が得られない場合もある
- 一般には複数の初期値をランダムに試して損失を最小とする解を採用する
- 平均の代わりにメドイド (medoid; 中心にある観測値) を用いる方法もある

$$\boldsymbol{x}_l^{\text{medoid}} = \arg\min_{\boldsymbol{x}_i} \sum_{i':C(i')=l} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2$$

事例

- 都道府県別好きなおむすびの具 (一部) での例

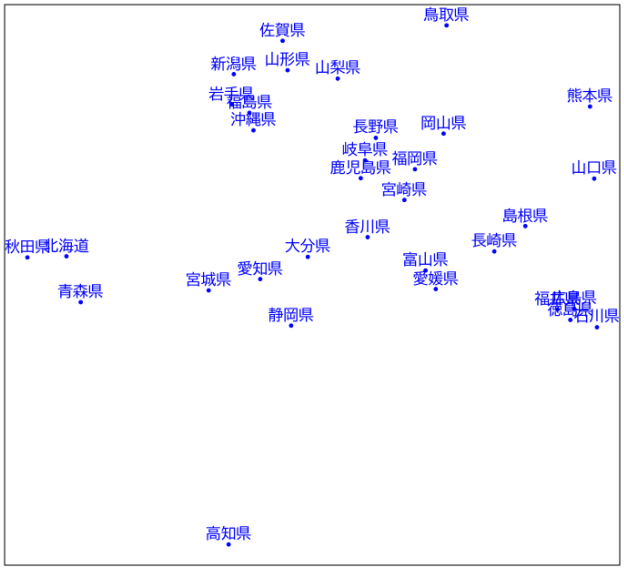

図 2: 非階層的クラスタリング

解析事例

都道府県別の社会生活統計指標

- データの属性

```
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
```

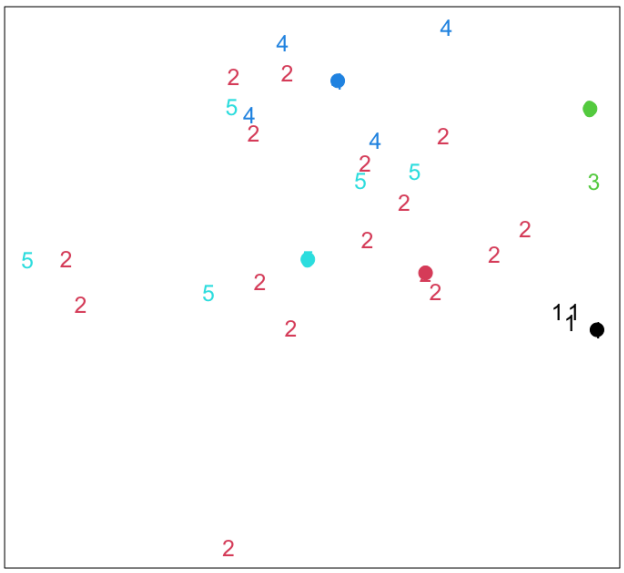

図 3: Lloyd-Forgy のアルゴリズム (その 1)

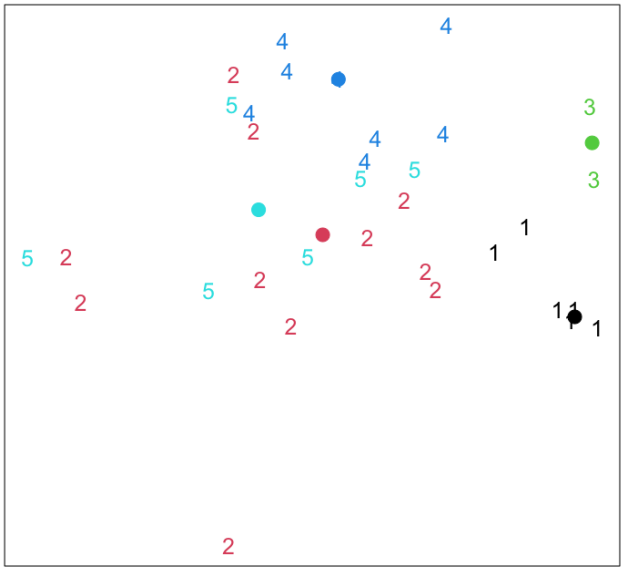

図 4: Lloyd-Forgy のアルゴリズム (その 2)

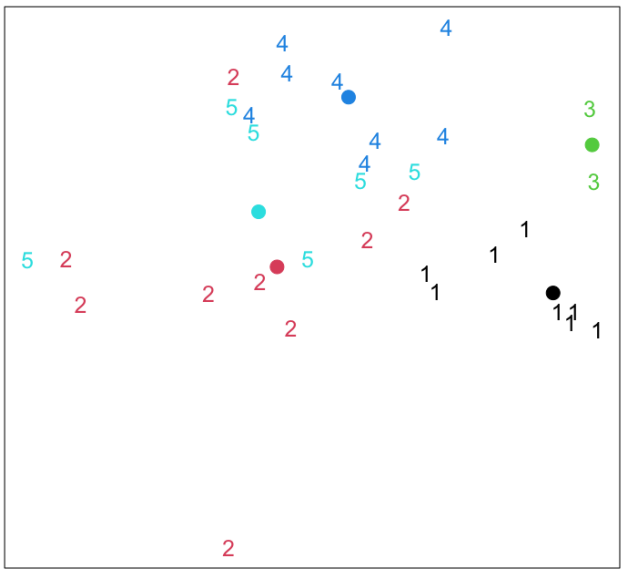

図 5: Lloyd-Forgy のアルゴリズム (その 3)

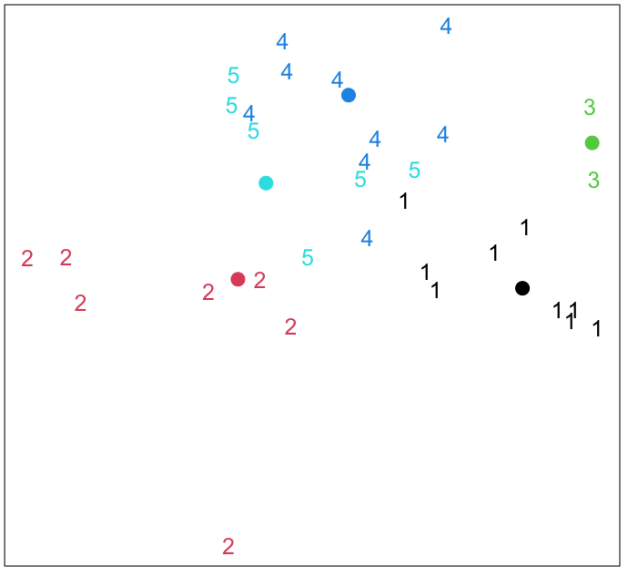

図 6: Lloyd-Forgy のアルゴリズム (その 4)

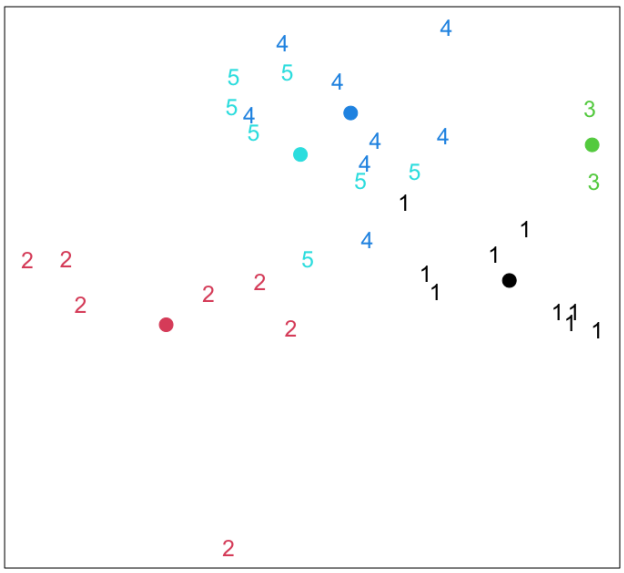

図 7: Lloyd-Forgy のアルゴリズム (その 5)

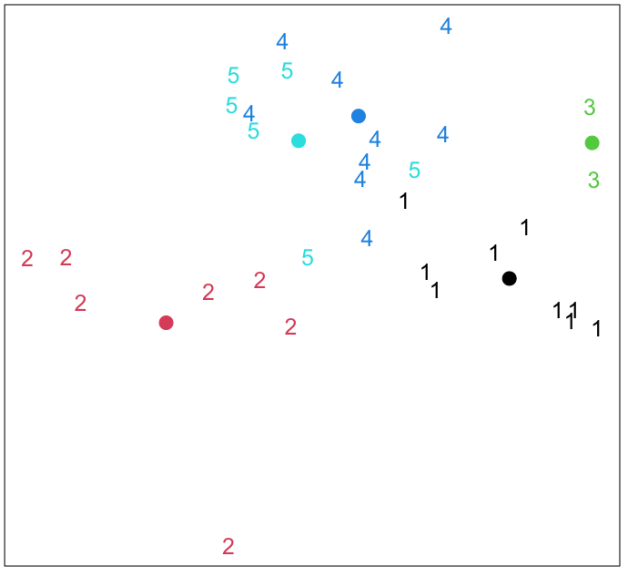

図 8: Lloyd-Forgy のアルゴリズム (その 6)

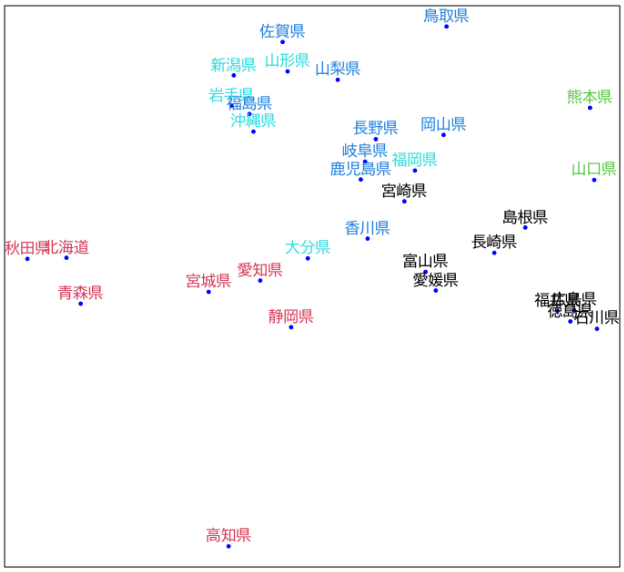

図 9: クラスタリングの結果

- 平均 0，分散 1 に正規化して解析
- ユークリッド距離 + k-平均法

```
=== cluster 1 ===
 [1] "岩手県" "宮城県" "秋田県" "山形県" "福島県" "新潟県" "富山県" "石川県" "福井県"
[10] "三重県" "滋賀県" "兵庫県" "島根県" "岡山県" "広島県" "山口県"
=== cluster 2 ===
[1] "東京都"
=== cluster 3 ===
[1] "宮崎県"   "鹿児島県"
=== cluster 4 ===
[1] "北海道"
=== cluster 5 ===
 [1] "青森県" "茨城県" "栃木県" "群馬県" "静岡県" "香川県" "佐賀県" "長崎県" "熊本県"
[10] "沖縄県"
=== cluster 6 ===
 [1] "山梨県"   "長野県"   "岐阜県"   "京都府"   "奈良県"   "和歌山県" "鳥取県"
 [8] "徳島県"   "愛媛県"   "高知県"   "大分県"
=== cluster 7 ===
[1] "埼玉県"   "千葉県"   "神奈川県" "愛知県"   "大阪府"   "福岡県"
```

- ユークリッド距離 + k-メドイド法

```
=== cluster 1 ===
[1] "北海道"
=== cluster 2 ===
 [1] "青森県"   "栃木県"   "群馬県"   "静岡県"   "佐賀県"   "長崎県"   "熊本県"
 [8] "宮崎県"   "鹿児島県" "沖縄県"
=== cluster 3 ===
 [1] "岩手県" "宮城県" "山形県" "新潟県" "長野県" "岐阜県" "三重県" "京都府" "兵庫県"
```

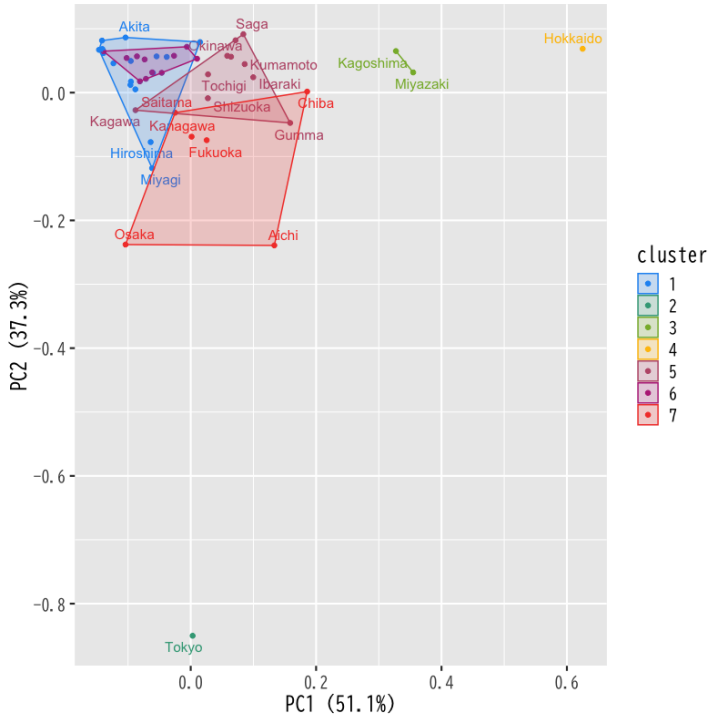

図 10: ユークリッド距離 + k-平均法

```
[10] "鳥取県" "岡山県" "広島県" "香川県" "愛媛県" "大分県"
=== cluster 4 ===
[1] "秋田県" "福島県" "富山県" "石川県" "福井県" "滋賀県" "奈良県" "島根県" "山口県"
=== cluster 5 ===
[1] "茨城県"   "埼玉県"   "千葉県"   "神奈川県" "愛知県"   "大阪府"   "福岡県"
=== cluster 6 ===
[1] "東京都"
=== cluster 7 ===
[1] "山梨県"   "和歌山県" "徳島県"   "高知県"
```

都道府県別好きなおむすびの具

- データの属性

  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
  【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%

  - 回答を県別に集計

- Hellinger 距離を利用

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

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

- Hellinger 距離 + k-メドイド法

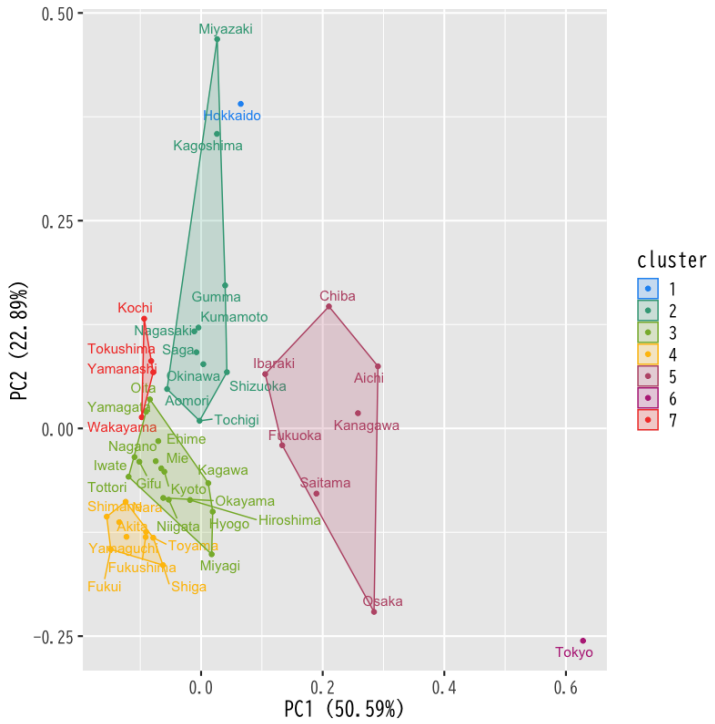

図 11: ユークリッド距離 + k-メドイド法

```
=== cluster 1 ===
[1] "北海道" "青森県" "秋田県"
=== cluster 2 ===
[1] "岩手県" "山形県" "新潟県" "沖縄県"
=== cluster 3 ===
 [1] "宮城県"   "茨城県"   "栃木県"   "群馬県"   "埼玉県"   "千葉県"   "東京都"
 [8] "神奈川県" "山梨県"   "長野県"   "宮崎県"
=== cluster 4 ===
[1] "福島県" "岐阜県" "愛知県" "鳥取県" "岡山県" "香川県" "佐賀県"
=== cluster 5 ===
 [1] "富山県"   "静岡県"   "三重県"   "滋賀県"   "京都府"   "大阪府"   "兵庫県"
 [8] "奈良県"   "和歌山県" "島根県"   "愛媛県"   "高知県"   "福岡県"   "長崎県"
[15] "熊本県"   "大分県"   "鹿児島県"
=== cluster 6 ===
[1] "石川県" "福井県" "広島県" "山口県" "徳島県"
```

クラスタ構造の評価

階層的方法の評価

- 評価の対象
  - データ $\boldsymbol{x}_i$ と最初に統合されたクラスタ C の距離

$$d_i = D(\boldsymbol{x}_i, C)$$

  - 最後に統合された 2 つのクラスタ C', C'' の距離

$$D = D(C', C'')$$

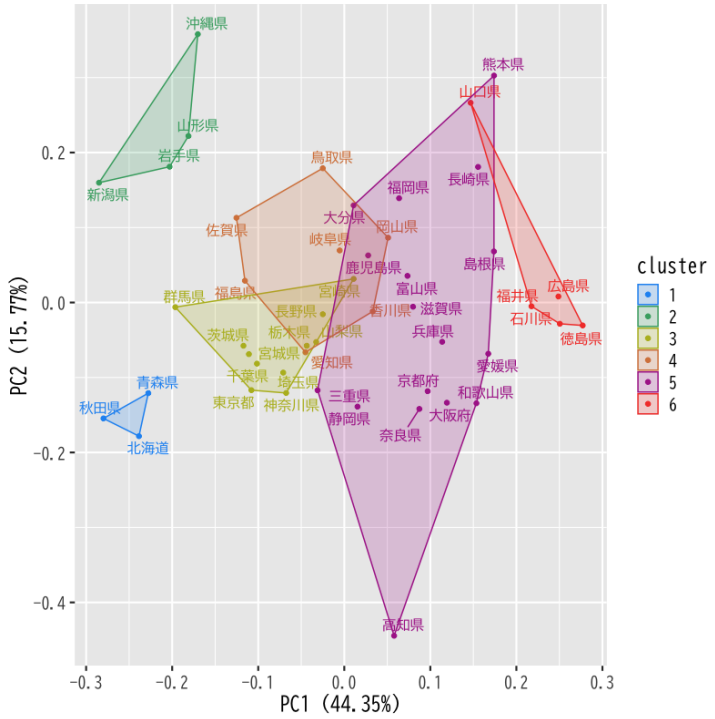

図 12: Hellinger 距離 + k-メドイド法

- **凝集係数** (agglomerative coefficient)

$$AC = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{d_i}{D}\right)$$

凝集係数の性質

- 定義より

$$0 \leq AC \leq 1$$

  - 1 に近いほどクラスタ構造が明瞭
- banner plot: 各 $(1 - d_i/D)$ を並べた棒グラフ
- banner plot の面積比として視覚化

非階層的方法の評価

- 評価の対象
  - $\boldsymbol{x}_i$ を含むクラスタ C^1 と $\boldsymbol{x}_i$ の距離

$$d_i^1 = D(\boldsymbol{x}_i, C^1 \setminus \boldsymbol{x}_i)$$

  - 一番近いクラスタ C^2 と $\boldsymbol{x}_i$ の距離

$$d_i^2 = D(\boldsymbol{x}_i, C^2)$$

- **シルエット係数** (silhouette coefficient)

$$S_i = \frac{d_i^2 - d_i^1}{\max(d_i^1, d_i^2)}$$

シルエット係数の性質

- 定義より

$$-1 \leq S_i \leq 1$$

  - 1 に近いほど適切なクラスタリング
- 全体の良さを評価するには S_i の平均を用いる
- 距離の計算を適切に行えば階層的方法でも利用可

演習

問題

- 以下の問に答えなさい
  - 群平均法において凝集係数が以下を満たすことを示しなさい

$$0 \leq AC \leq 1$$

  - シルエット係数が以下を満たすことを示しなさい

$$-1 \leq S_i \leq 1$$

解答例

- 2 つのクラスタ C_a, C_b が最も近いとする

$$D(C_c, C_d) \geq D(C_a, C_b), \quad \forall c, d$$

- 統合して計算される距離では下が成立

$$\begin{aligned} D(C_a + C_b, C_c) &= \frac{|C_a| D(C_a, C_c) + |C_b| D(C_b, C_c)}{|C_a| + |C_b|} \\ &\geq \frac{|C_a| D(C_a, C_b) + |C_b| D(C_a, C_b)}{|C_a| + |C_b|} \\ &= D(C_a, C_b) \end{aligned}$$

  統合した結果，それより短い距離が現れることはない
- 以上より

$$0 \leq d_i \leq D$$

$$0 \leq 1 - \frac{d_i}{D} \leq 1$$

よって

$$0 \leq AC \leq 1$$

- 非負値の大小関係に注意する

$$-\max(d_i^1, d_i^2) \leq d_i^2 - d_i^1 \leq \max(d_i^1, d_i^2)$$

より

$$-1 \leq S_i \leq 1$$

解析事例

都道府県別の社会生活統計指標

- 凝集係数を用いて階層的方法の距離を検討
- ユークリッド距離とマンハッタン距離を比較
  - 正規化は共通 (平均 0，絶対偏差 1)
  - クラスタ距離は群平均法

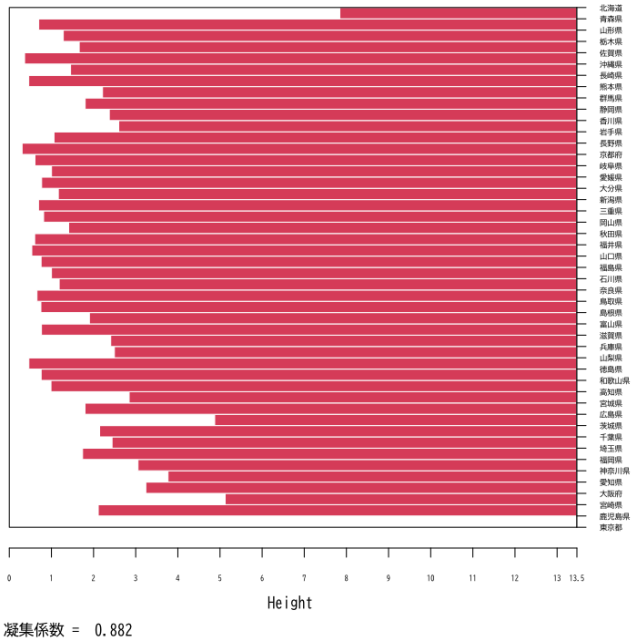

図 13: 凝集係数 (ユークリッド距離)

- 一部のデータの距離が大きいと凝集係数は大きくなりがち (理由を考えてみよう)

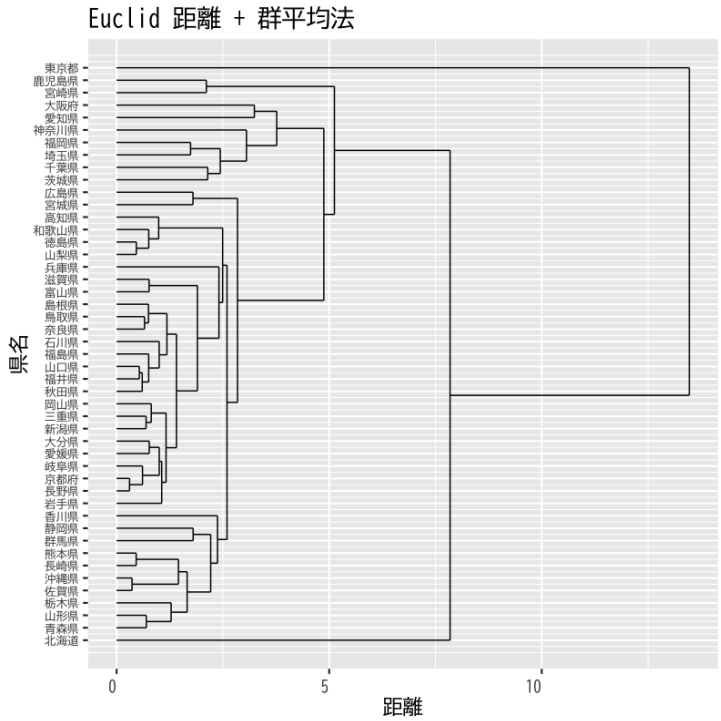

図 14: デンドログラム (ユークリッド距離)

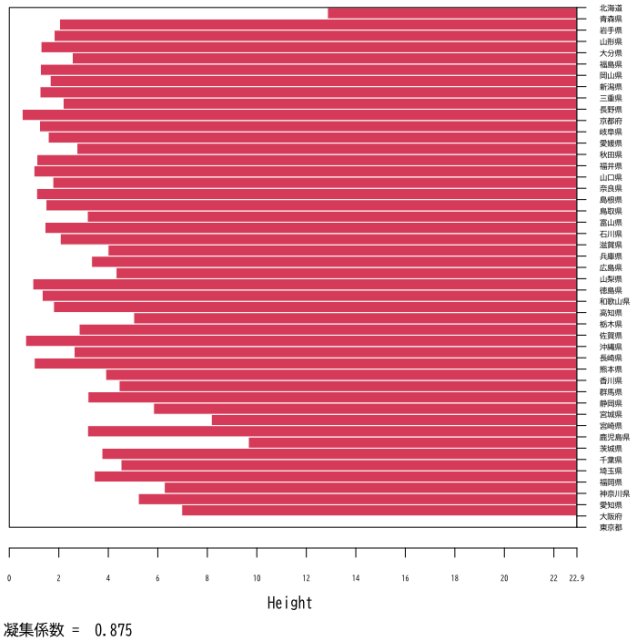

図 15: 凝集係数 (マンハッタン距離)

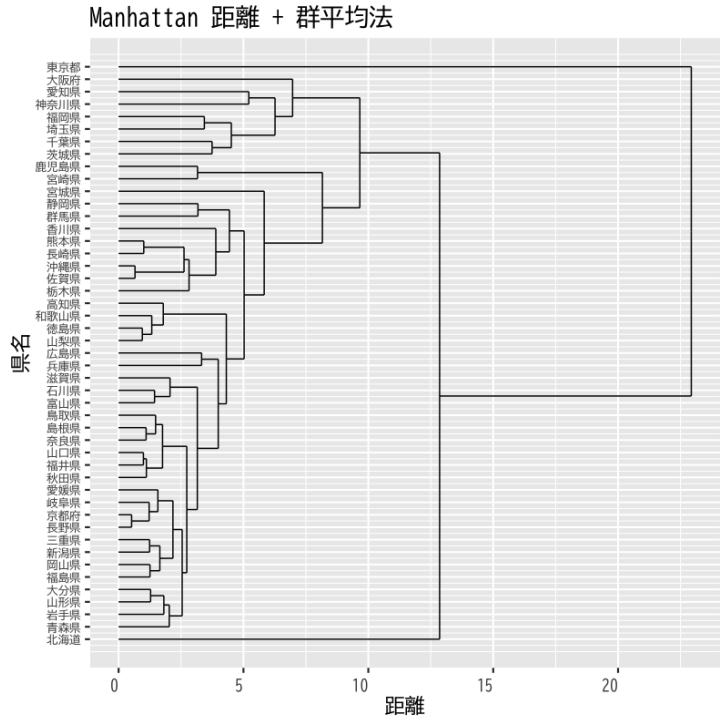

図 16: デンドログラム (マンハッタン距離)

- 北海道，東京，宮崎，鹿児島を除いて再計算する

  凝集係数（ユークリッド距離）
  [1] 0.807
  凝集係数（マンハッタン距離）
  [1] 0.782

都道府県別好きなおむすびの具

- シルエット係数を用いて非階層的方法のクラスタ数を検討
  - データ距離は Hellinger 距離
  - クラスタ距離は群平均法
- クラスタ数を 4-10 として比較

  シルエット係数
  k = 4 [1] 0.172
  k = 5 [1] 0.152
  k = 6 [1] 0.176
  k = 7 [1] 0.2
  k = 8 [1] 0.204
  k = 9 [1] 0.206
  k = 10 [1] 0.195

次回の予定

- **第 1 日：時系列の基本モデル**
- 第 2 日：モデルの推定と予測

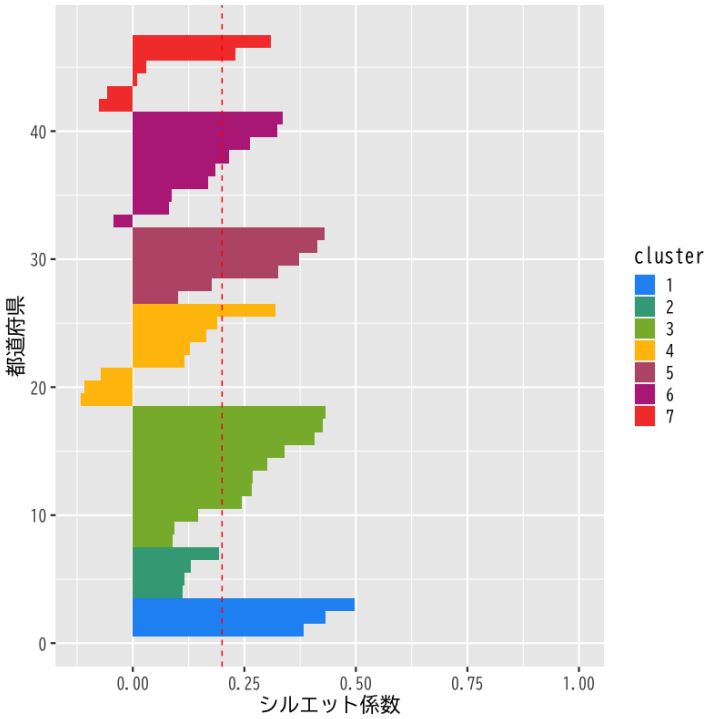

図 17: シルエット係数の分布 (k=7)

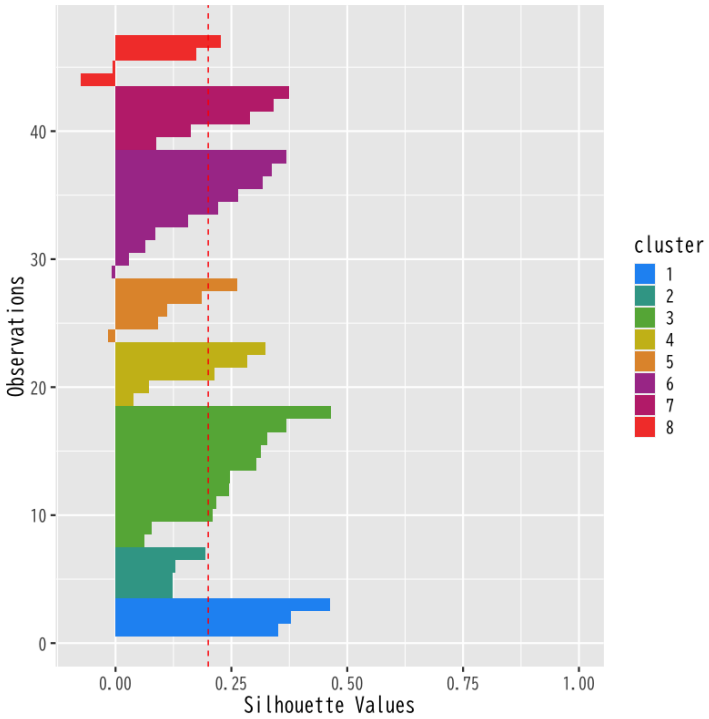

図 18: シルエット係数の分布 (k=8)

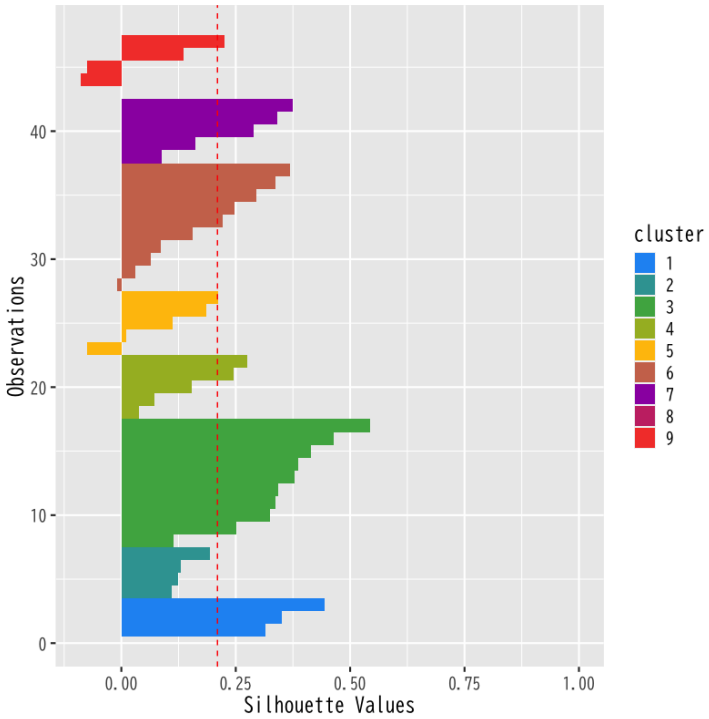

図 19: シルエット係数の分布 (k=9)

図 20: 非階層的クラスタリング (k=8)