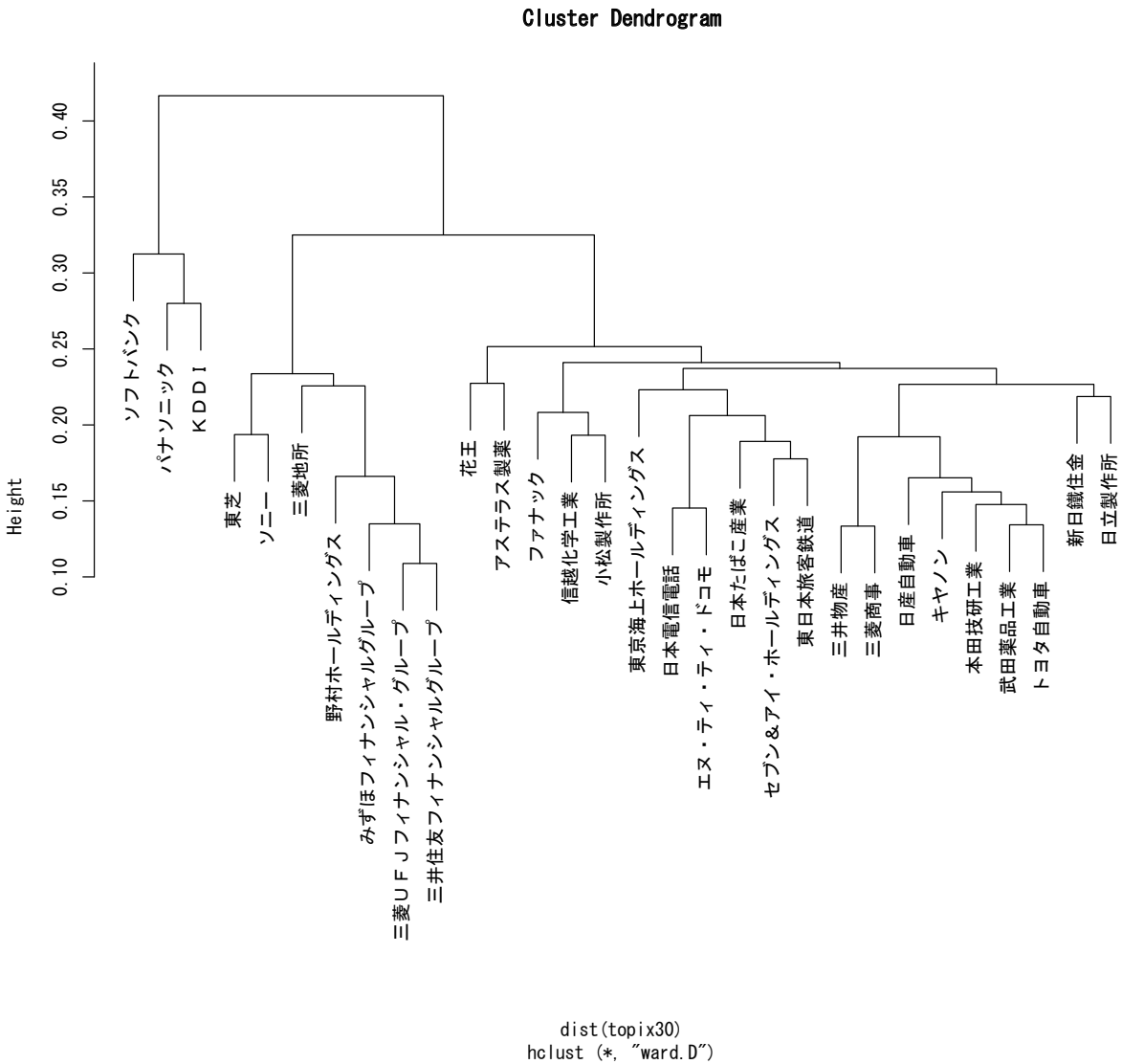
データ解析
第十四回「クラスタリング」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

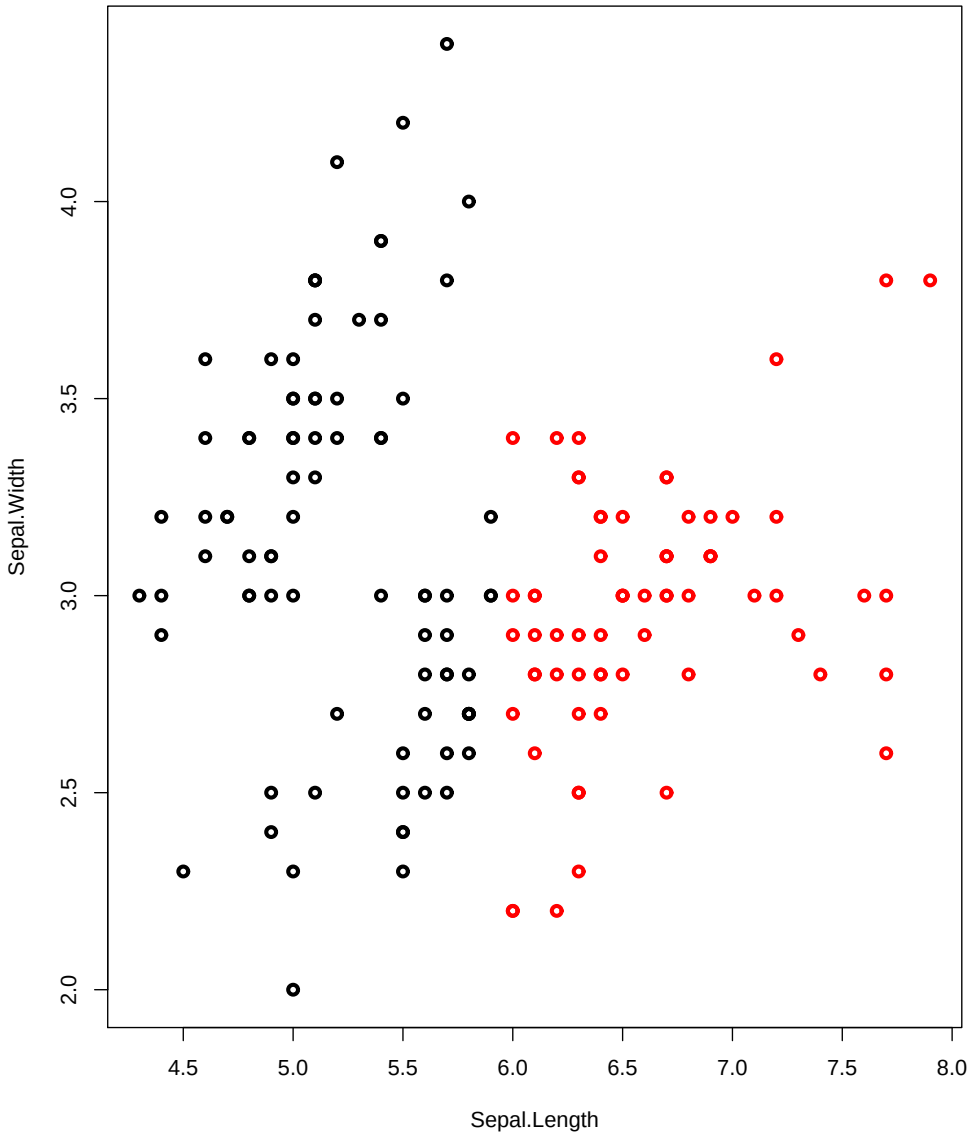
今日の講義内容

- 階層的クラスタリング
- k-means 法
- 混合ガウスモデルによるクラスタリング：EM-アルゴリズム
- TOPIX CORE 30 のクラスタリング

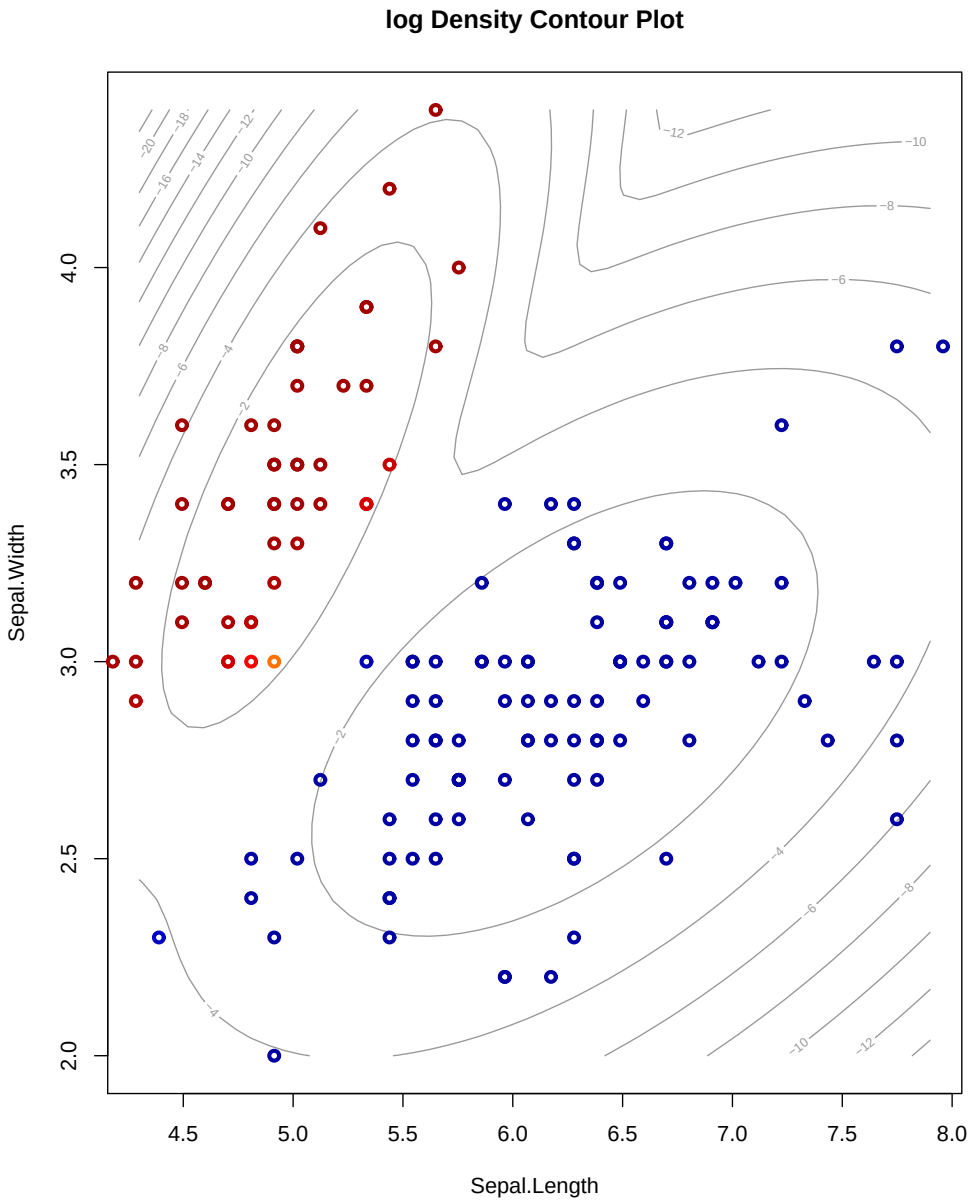
階層的クラスタリング


Cluster Dendrogram
Height
0.10
0.15
0.20
0.25
0.30
0.35
0.40
ソフトバンク
パナソニック
ＫＤＤＩ
東芝
ソニー
三菱地所
野村ホールディングス
みずほフィナンシャルグループ
三菱ＵＦＪフィナンシャル・グループ
三井住友フィナンシャルグループ
花王
アステラス製薬
ファナック
信越化学工業
小松製作所
東京海上ホールディングス
日本電信電話
エヌ・ティ・ティ・ドコモ
日本たばこ産業
セブン＆アイ・ホールディングス
東日本旅客鉄道
三井物産
三菱商事
日産自動車
キヤノン
本田技研工業
武田薬品工業
トヨタ自動車
新日鐵住金
日立製作所
dist(topix30)
hclust (*, "ward.D")

ハードクラスタリング



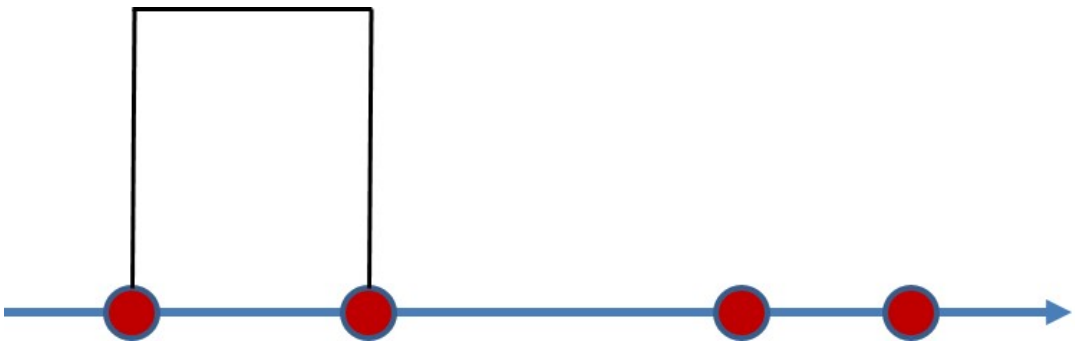
ソフトクラスタリング


log Density Contour Plot
Sepal.Width
Sepal.Length
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

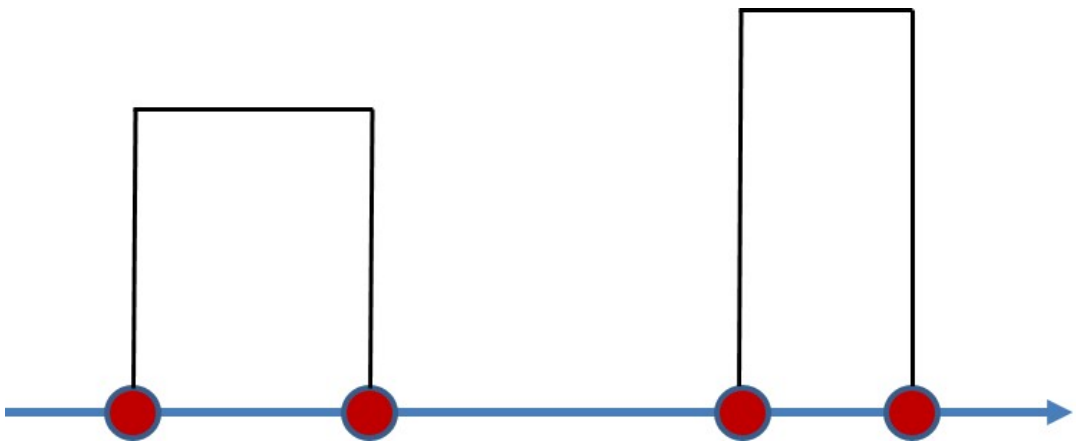
構成

階層的クラスタリングの手順

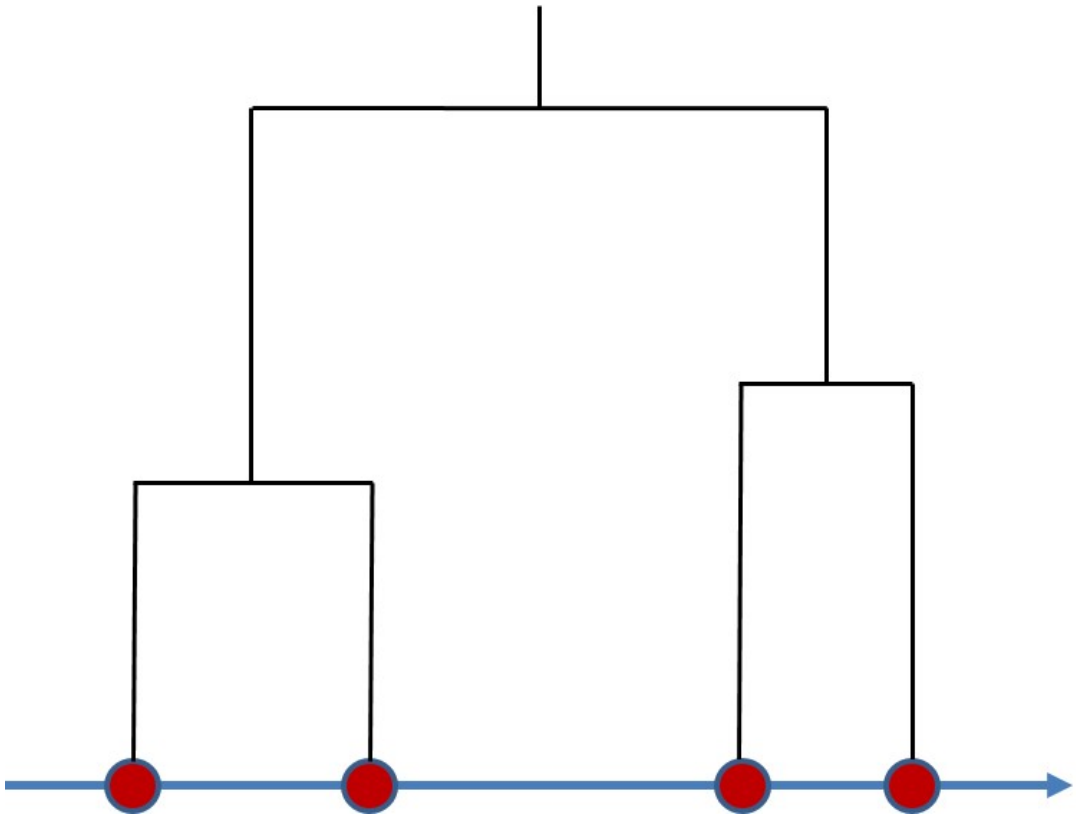
階層的クラスタリングの手順



階層的クラスタリングの手順



階層的クラスタリングの手順



サンプル間およびクラスタ間の距離が決まっていれば階層的クラスタリングは可能.

サンプル間の距離

- ユークリッド距離 ((dist(x,method="euclidean")):

$$d(x, x') = \sqrt{\sum_{j=1}^{d} (x_j - x'_j)^2}.$$

- ミンコフスキー距離 (ℓ_p-norm, dist(x,method="minkowski",p=3))

$$d(x, x') = \left(\sum_{j=1}^{d} |x_j - x'_j|^p \right)^{1/p}.$$

- 最大距離 (dist(x,method="maximum"))

$$d(x, x') = \max_{j=1,\ldots,d} |x_j - x'_j|.$$

サンプル間の距離 (続き)

- マンハッタン距離 (dist(x,method="manhattan"))

$$d(x, x') = \sum_{j=1}^{d} |x_j - x'_j|.$$

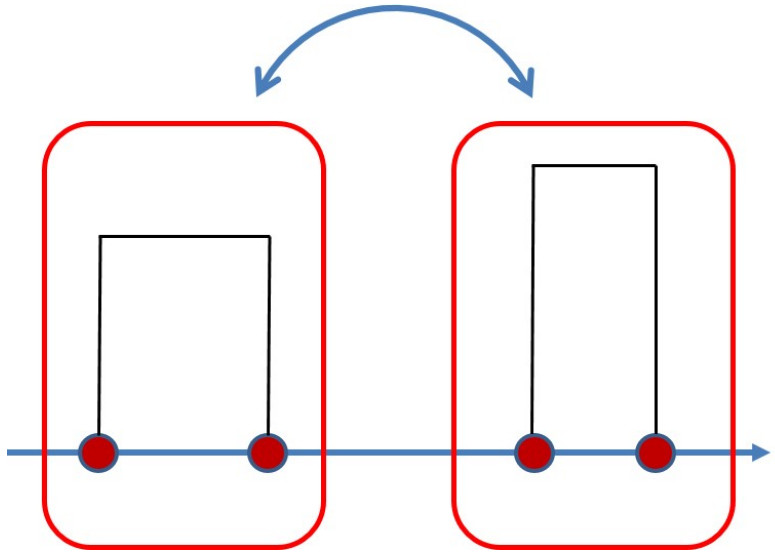
- キャンベラ距離 (dist(x,method="canberra"))

$$d(x, x') = \sum_{j=1}^{d} \frac{|x_j - x'_j|}{|x_j| + |x'_j|}.$$

- バイナリー距離 (x, x' が 0-1 値のバイナリベクトルのとき, dist(x,method="binary"))

$$d(x, x') = 1 - \frac{\sum_{j=1}^{d} \min(x_j, x'_j)}{\sum_{j=1}^{d} \max(x_j, x'_j)}.$$

クラスタ間の距離



- 最短距離法 (単連結法, single linkage method)

$$D(C, C') = \min_{x \in C, x' \in C} d(x, x')$$

  ノイズに弱い.
- 最長距離法 (完全連結法, complete linkage method)

$$D(C, C') = \max_{x \in C, x' \in C} d(x, x')$$

  保守的.

クラスタ間の距離 (続き)

- 群平均法 (average linkage method)

$$D(C, C') = \frac{1}{|C||C'|} \sum_{x \in C, x' \in C} d(x, x').$$

- ウォード法 (Ward's method)

(ユークリッド距離の場合) $\quad D(C, C') = \dfrac{\|\bar{x}_C - \bar{x}_{C'}\|^2}{1/|C| + 1/|C'|},$

ただし，$\bar{x}_C = \frac{1}{|C|} \sum_{x \in C} x$．ユークリッド距離以外では

$$D(C_1 \cup C_2, C_3) = \frac{(|C_1|+|C_2|)D(C_1,C_2)+(|C_2|+|C_3|)D(C_2,C_3)+(|C_1|+|C_3|)D(C_1,C_3)}{\sum_{j=1}^{3} |C_j|},$$

で再帰的に定義．もっとも広がりが小さく抑えられるクラスタを追加.

- McQuitty 法 (McQuitty's method)
- 重心法 (centroid method)
- メディアン法 (median method)

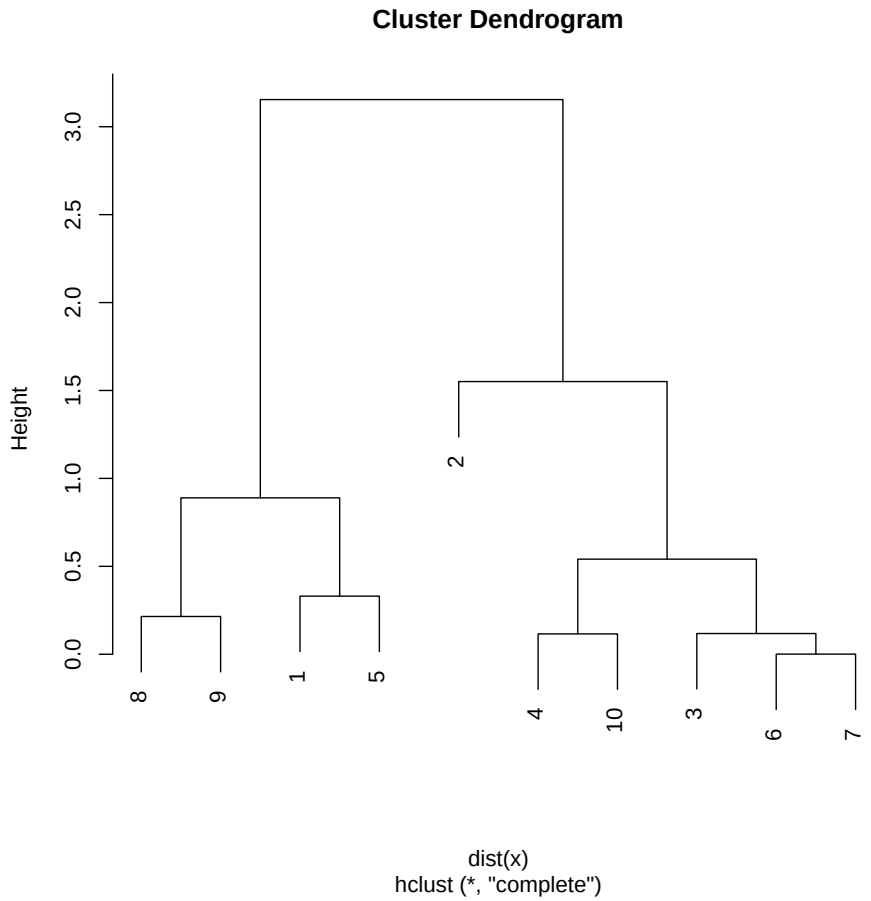
階層的クラスタリングの手順 (具体的に)

初期化: サンプル $\{x_i\}_{i=1}^n$ それぞれが 1 つのクラスタを形成するように n 個のクラスタを作成.
以下の手順を全体がひとつのクラスタになるまで続ける.

1. すべてのクラスタ間の距離を計算.
2. 最も距離の近いクラスタを統合.

階層的クラスタリングのRコード

```
hclust(dist(x))
hclust(dist(x,method="canberra"),method="ward")
```



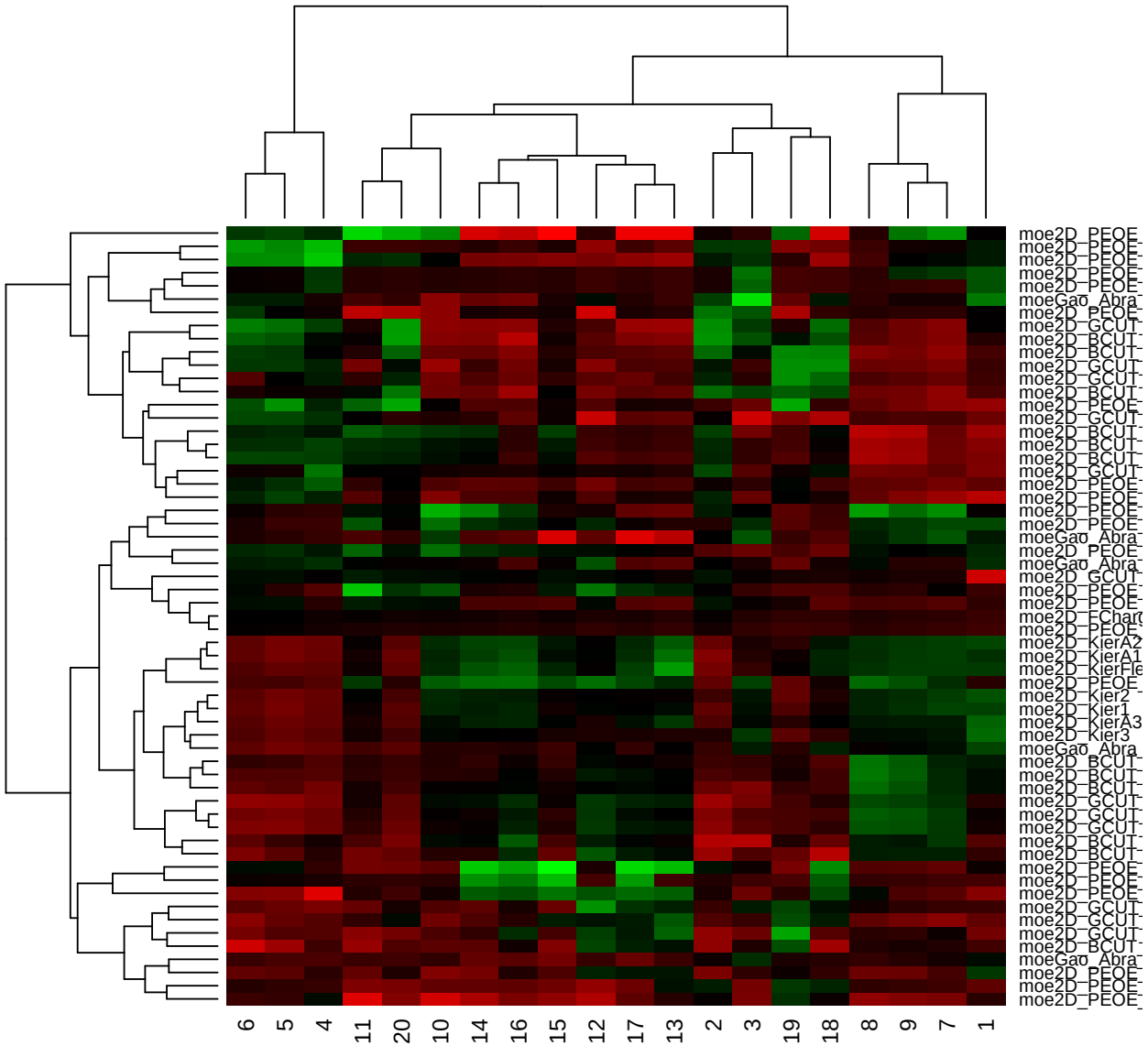
ヒートマップ

行列データがあった場合，行間の階層的クラスタリングと列間の階層的クラスタリングを同時に行い，プロットすることができる．

```
dsf <- function(x) dist(x,method="maximum")
hcf <- function(d) hclust(d, method = "complete")
heatmap(x,scale="column",col=hmcol,distfun=dsf,hclustfun=hcf)
```

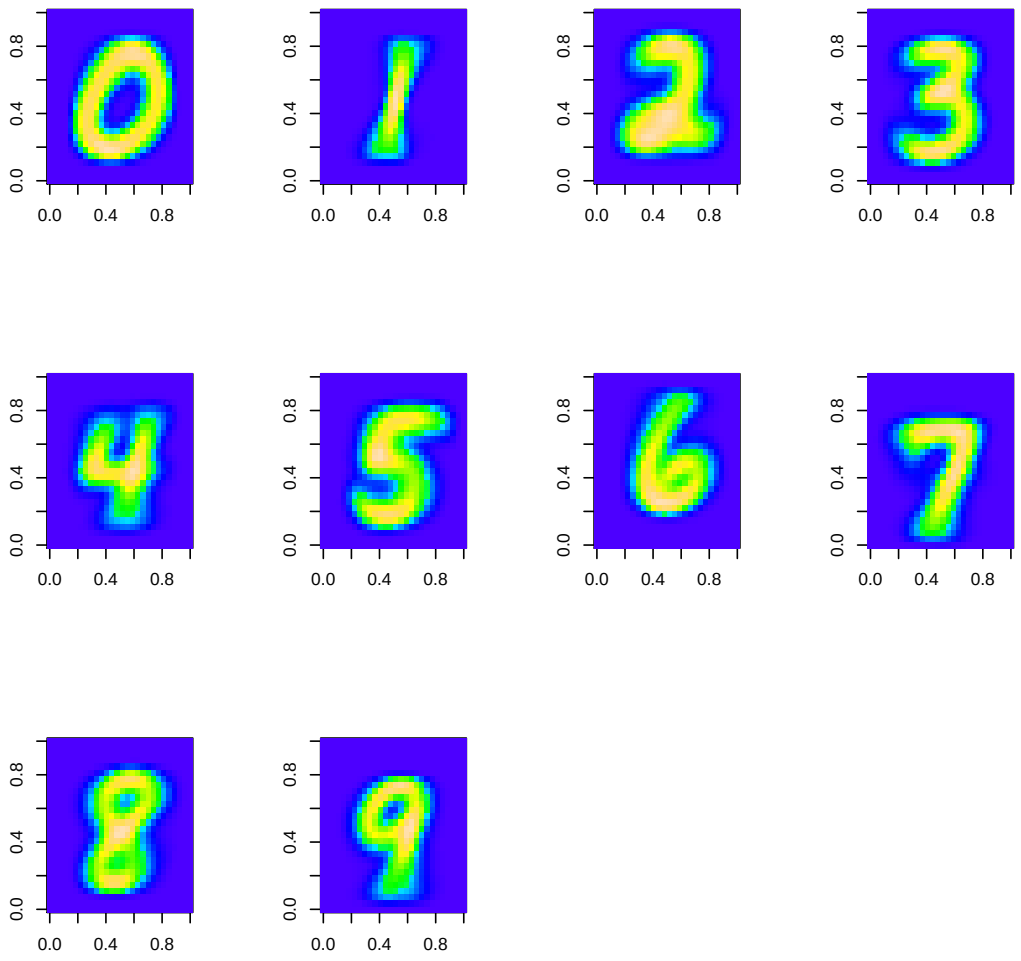
col 以降は省略可．

ジヒドロ葉酸還元酵素阻害剤データのヒートマップ



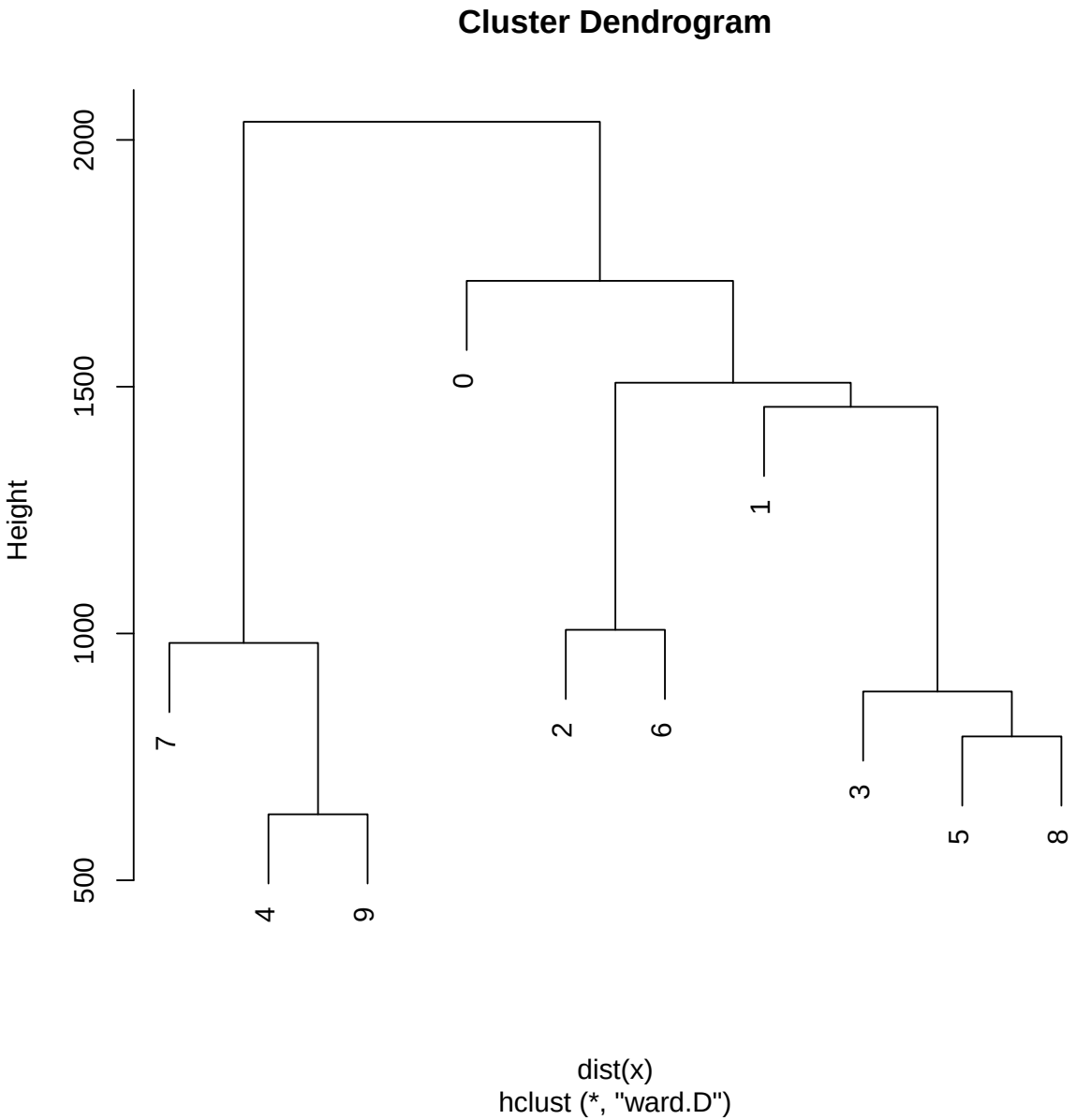
コードはスクリプトファイルを参照

手書き文字のクラスタリング



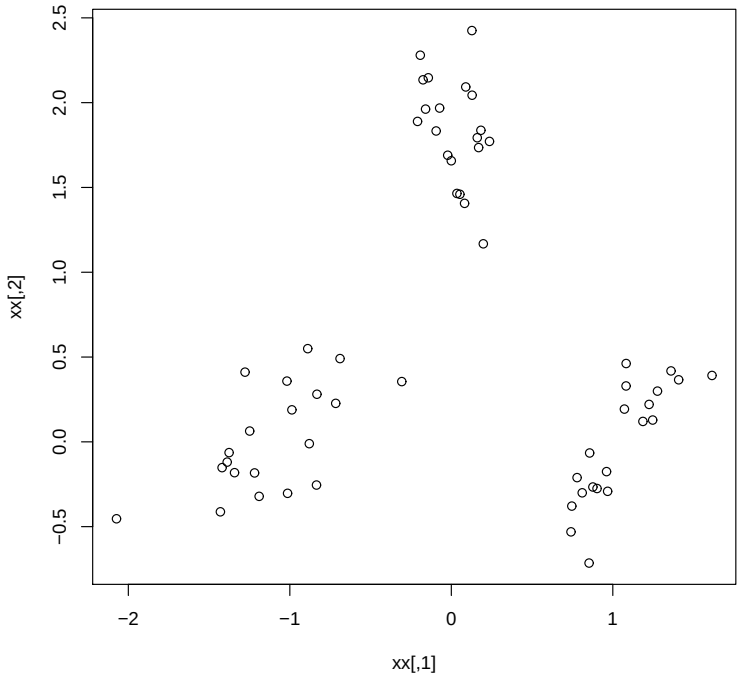
MNIST データセット．数字 0 から 9，各文字画像の平均値を計算．

手書き文字の階層的クラスタリング

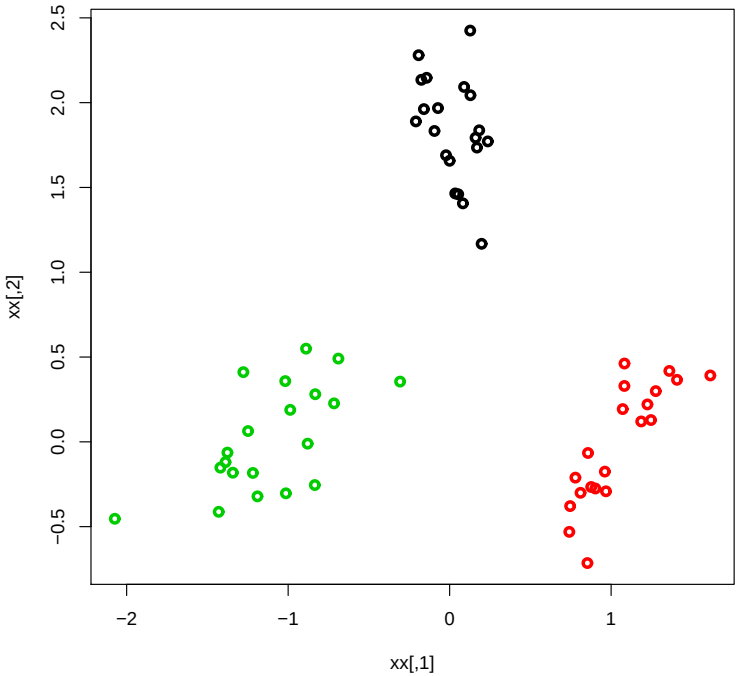


構成

k-means 法の結果



(a) 生データ



(b) クラスタリング結果

k-means 法の最適化問題としての定式化

K 個の代表点でサンプルを近似したい.

$$\min_{\mu_1,\ldots,\mu_K} \sum_{i=1}^{n} \|x_i - \mu_{k_i}\|^2$$

ただし，μ_{k_i} は x_i に一番近い $\{\mu_1, \ldots, \mu_K\}$ 内の点.

この最適化問題は **NP-困難.**
局所的に $\{\mu_k\}_k$ を繰り返し更新して良い局所解を求める.

k-means 法の手順

1. クラスタリング:

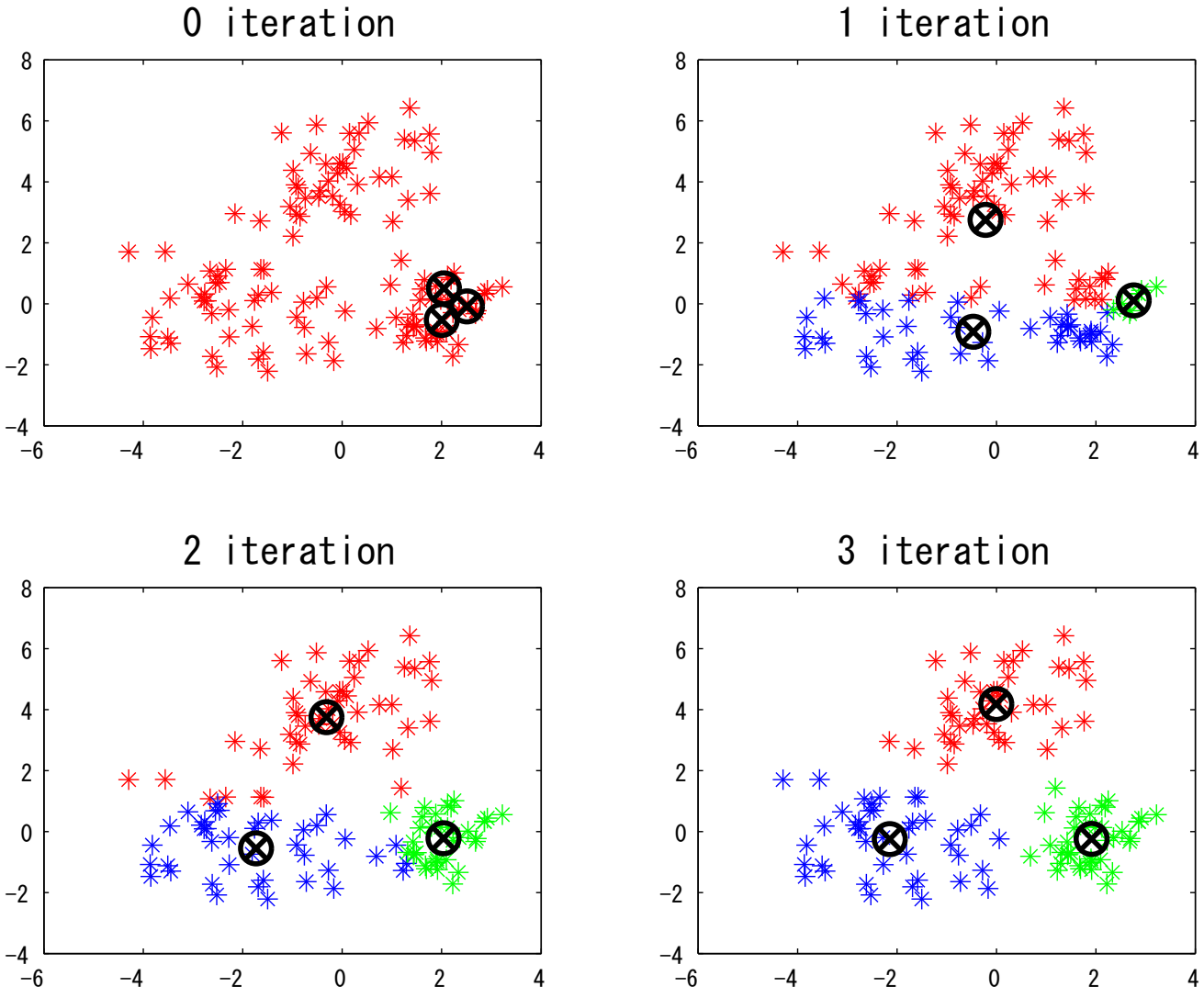
$$C_k = \{i \in \{1, \ldots, n\} \mid x_i \text{ にとって } \mu_k \text{ が最も近い } \}.$$

2. クラスタの平均を計算:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i.$$

上記の手順を収束するまで何度も繰り返す.

k-means 法の収束の様子

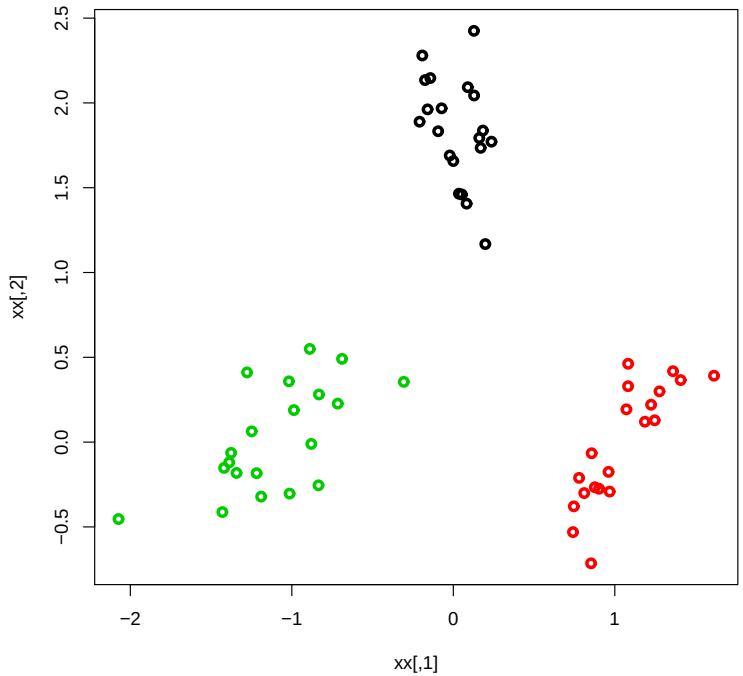


R による k-means 法

```
res <- kmeans(x,k)
res$centers  #cluster center
res$cluster  #cluster assignment
```

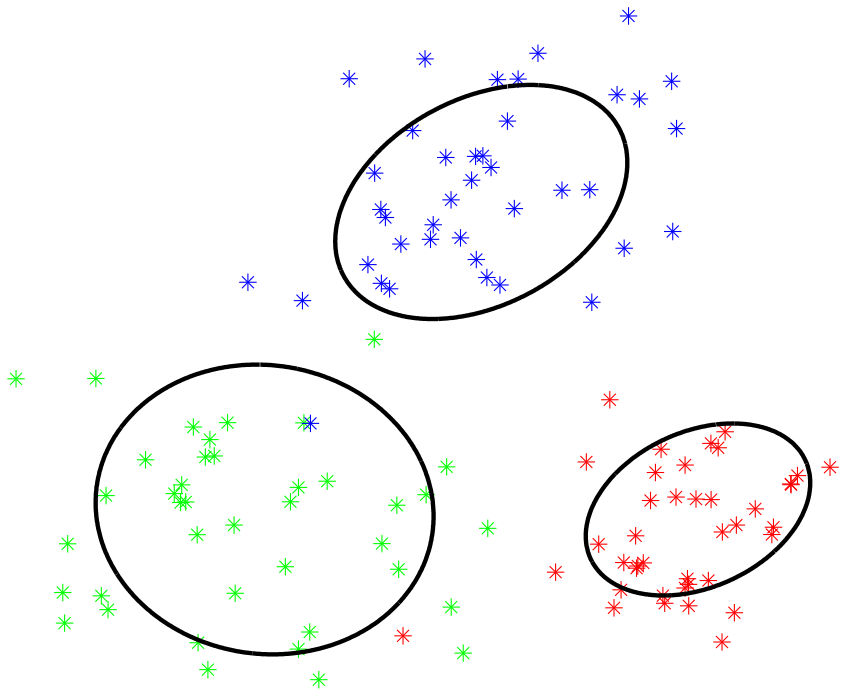
x は $n \times d$ (サンプル数 × 次元), k はクラスタ数.

人工データで実行した結果 (再掲)



構成

混合ガウス分布



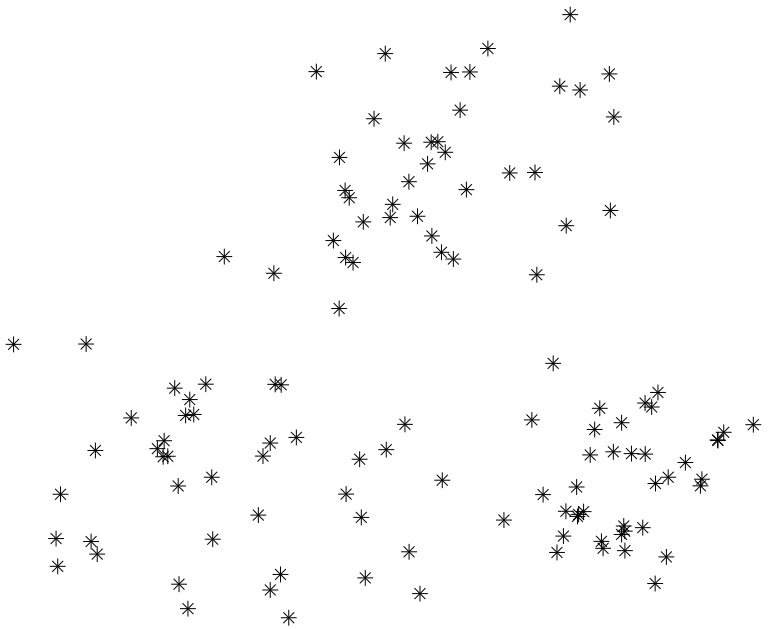
混合ガウス分布

$$P(x) = \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k)$$

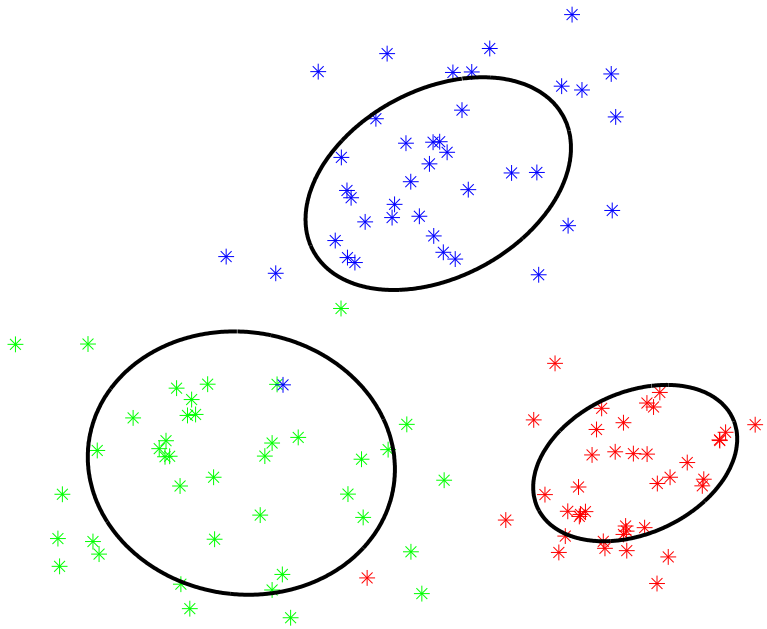
いくつかのガウス分布の足しあわせ $(\sum_{k=1}^{K} \pi_k = 1)$.

混合ガウス分布

実際は色分けはされていない.



(c) これから



(d) これを復元したい

混合ガウス分布の生成モデル

x の密度関数:

$$p(x) = \sum_{k=1}^{K} \pi_k g(x|\mu_k, \Sigma_k)$$

(g はガウス分布の密度関数とする)

クラスタラベル $Z = k$ は確率 π_k で得られる.

$$Z \sim \mathrm{Mult}(\pi_1, \ldots, \pi_K)$$

クラスタが $Z = k$ であるとき，x はガウス分布から得られる:

$$X|\{Z = k\} \sim N(\mu_k, \Sigma_k).$$

この時，X の周辺分布は $p(X)$ となる.

では，X が与えられたもとでの Z の分布はどうなるだろうか？

ガウス要素への寄与率

ベイズの定理

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

クラスタ k への寄与率:

$$P(Z = k|X) = \frac{g(X|\mu_k, \Sigma_k)\pi_k}{\sum_{k'=1}^{K} g(X|\mu_{k'}, \Sigma_{k'})\pi_{k'}}.$$

$P(Z = k|X)$ が大きいほど，X はクラスタ k に属している確率が高い．

パラメータをどう推定するか？
対数尤度関数が凹関数ではなく局所最適解がたくさんあるため，推定が難しい．
→ EM-アルゴリズム．

EM-アルゴリズム (※超重要)

EM-アルゴリズムのアイディア:

- 観測値 $\{x_i\}_{i=1}^n$ は不完全データ.
  本来は，どのクラスタに属しているかの**隠れ変数** $\{z_i\}_{i=1}^n$ が裏にある.
- 隠れ変数を固定してしまえば，パラメータの最尤推定量は簡単に求まる.

隠れ変数 $\{z_i\}_{i=1}^n$ を知っている時の対数尤度:

$$\log\left(\prod_{i=1}^{n} g(x_i|\mu_{z_i}, \Sigma_{z_i})\right).$$

これの $\{\mu_k, \Sigma_k\}_k$ に関する最大化は簡単.
EM-アルゴリズムでは z_i の分布を推定し，対数尤度の z_i に関する期待値を最大化.

EM-アルゴリズムの手続き

$\{z_i\}_{i=1}^n$ の分布と $\theta = \{\pi_k, \mu_k, \Sigma_k\}_k$ を交互に推定.（同時に推定は難しいが片方を固定した場合，もう片方の推定は簡単）

- **E-step:** 隠れ変数 $\{z_i\}_i$ の分布を推定.

$$p(z_i = k | x_i; \hat{\theta}_{[t-1]}) = \frac{\hat{\pi}_k g(x_i; \hat{\mu}_k^{[t-1]}, \hat{\Sigma}_k^{[t-1]})}{\sum_{k'=1}^K \hat{\pi}_{k'}^{[t-1]} g(x_i; \hat{\mu}_{k'}^{[t-1]}, \hat{\Sigma}_{k'}^{[t-1]})}.$$

- **M-step:** パラメータの推定．$w_i^{(k)} = p(z_i = k | x_i; \hat{\theta}_{[t-1]})$ として，

$$\hat{\pi}_k^{[t]} = \frac{\sum_{i=1}^n w_i^{(k)}}{n},$$

$$\hat{\mu}_k^{[t]} = \frac{\sum_{i=1}^n w_i^{(k)} x_i}{n},$$

$$\hat{\Sigma}_k^{[t]} = \frac{\sum_{i=1}^n w_i^{(k)} (x_i - \hat{\mu}_k^{[t]})(x_i - \hat{\mu}_k^{[t]})^\top}{n}.$$

EM-アルゴリズムの意味

M-step では，隠れ変数の分布を固定したもとでの重み付き尤度最大化を行っている．

$$\begin{aligned}\{\hat{\mu}_k, \hat{\Sigma}_k, \hat{\pi}_k\}_k &= \underset{\theta=\{\mu_k,\Sigma_k,\pi_k\}_k}{\arg\max} \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log(p(x = x_i, z_i = k|\theta)) \\ &= \underset{\theta=\{\mu_k,\Sigma_k,\pi_k\}_k}{\arg\max} \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log(g(x_i; \mu_k, \Sigma_k)\pi_k).\end{aligned}$$

右辺の max の中身を $L_0(\theta||\theta_{[t-1]})$ とおく．

L_0 は次のように展開できる.

$$
\begin{aligned}
L_0(\theta||\theta_{[t-1]}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log(p(x_i|z_i=k;\theta)\pi_k) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log(\sum_{k'=1}^{k} p(x_i|z_i=k';\theta)\pi_{k'}) \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log\left(\frac{p(x_i|z_i=k;\theta)\pi_k}{\sum_{k'=1}^{k} p(x_i|z_i=k';\theta)\pi_{k'}}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log(p(x_i|\theta)) + \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log\left(p(z_i=k|x_i,\theta)\right) \\
&= \sum_{i=1}^{n} \log(p(x_i|\theta)) + \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log\left(p(z_i=k|x_i,\theta)\right).
\end{aligned}
$$

右辺第二項を $-L_1(\theta||\theta_{[t-1]})$ とおくと，次が成り立つ:

$$
\sum_{i=1}^{n} \log(p(x_i|\theta)) \geq L_0(\theta||\theta_{[t-1]}) + L_1(\theta||\theta_{[t-1]}).
$$

ここで

$$L(\theta||\theta_{[t-1]}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log\left(p(z_i = k|x_i, \theta)\right)$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log\left(\frac{w_i^{(k)}}{p(z_i = k|x_i, \theta)}\right) - \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log(w_i^{(k)}).$$

今，$w_i^{(k)} = p(z_i = k|x_i, \theta_{[t-1]})$ であるので，

$$L_1(\theta||\theta_{[t-1]}) = \sum_{i=1}^{n} \mathrm{KL}\left(p(z_i|x_i, \theta_{[t-1]})||p(z_i|x_i, \theta)\right) - \sum_{i=1}^{n}\sum_{k=1}^{K} w_i^{(k)} \log(w_i^{(k)})$$

となる．ここで，$\mathrm{KL}(\cdot||\cdot)$ は KL-ダイバージェンスである．KL-ダイバージェンスは非負なので，

$$L_1(\theta||\theta_{[t-1]}) - L_1(\theta_{[t-1]}||\theta_{[t-1]}) = \sum_{i=1}^{n} \mathrm{KL}\left(p(z_i|x_i, \theta_{[t-1]})||p(z_i|x_i, \theta)\right) \geq 0 \quad (\forall\theta)$$

である．

以上より，対数尤度は

$$\sum_{i=1}^{n} \log(p(x_i|\theta)) \geq L_0(\theta||\theta_{[t-1]}) + L_1(\theta||\theta_{[t-1]})$$

で下から抑えられる．M-Step で第一項 $L_0(\theta||\theta_{[t-1]})$ を最大化することで，

$$\begin{aligned} &L_0(\theta_{[t]}||\theta_{[t-1]}) + L_1(\theta_{[t]}||\theta_{[t-1]}) - (L_0(\theta_{[t-1]}||\theta_{[t-1]}) + L_1(\theta_{[t-1]}||\theta_{[t-1]})) \\ \geq &L_0(\theta_{[t]}||\theta_{[t-1]}) - L_0(\theta_{[t-1]}||\theta_{[t-1]}) \geq 0 \end{aligned}$$

となるので，EM-アルゴリズムは 尤度の下界を単調に増大させること に対応する．

このような方法を下界最大化アルゴリズムと呼ぶ．
※去年のスライドはこの部分が間違っています．

EM-アルゴリズムの注意点

EM-アルゴリズムは本当の大域的最適解には到達しない.

むしろ，到達しないほうが良い.

なぜなら，ある一点に μ_k をおいて，$\Sigma_k \to O$ と極限を取れば尤度は無限大になるからである.

しかし，$\Sigma_k = O$ なる解は完全に過適合しており，推定量としては望ましくない.

EM-アルゴリズムは ちょうどよい局所最適解 を求める方法とみなすことができる.

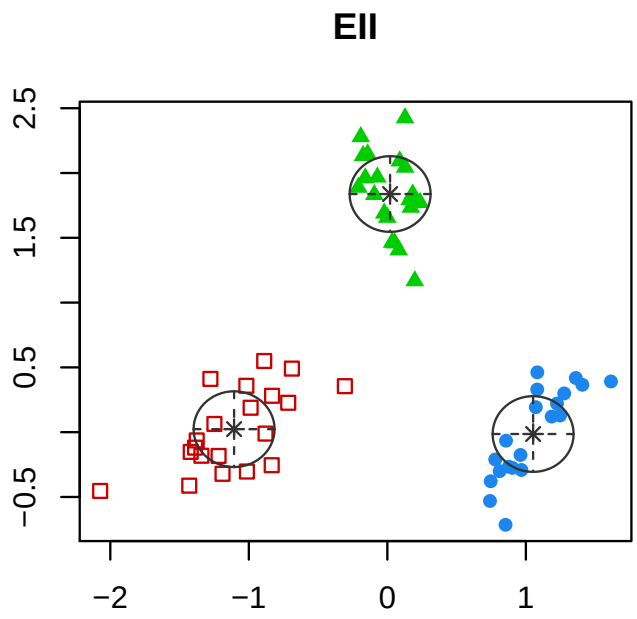
Rによる混合ガウス分布のクラスタリング

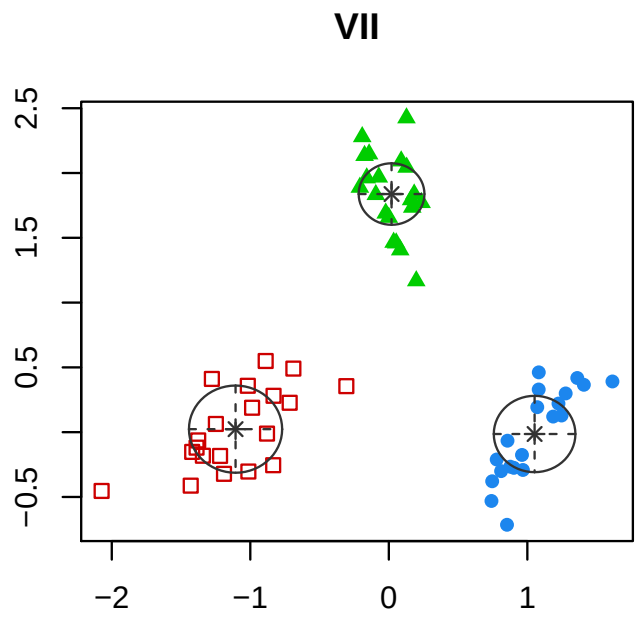
```
library(mclust)
result <- Mclust(data)
result <- Mclust(data,G=3,modelNames=''VVV'')
```

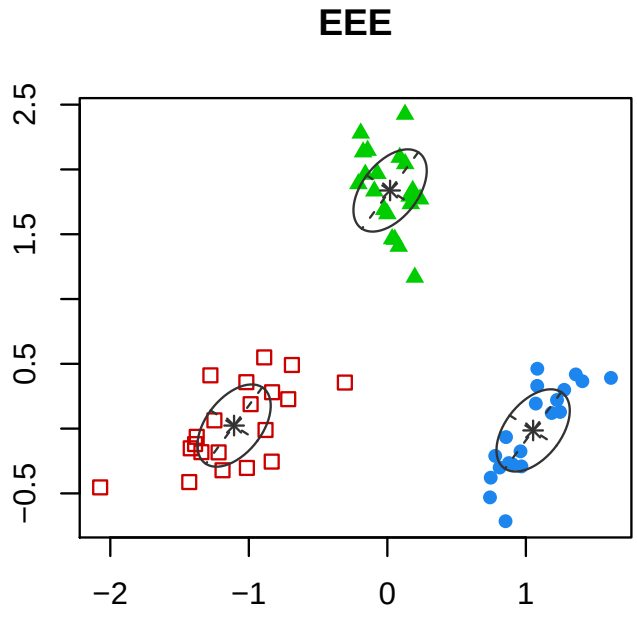
G=3 でクラスタ数を設定可能．modelNames で分散共分散行列のモデルを設定．

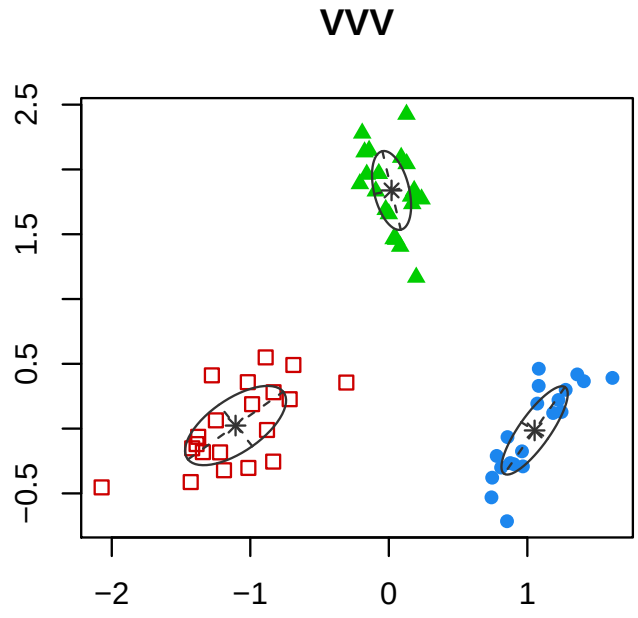
modelNames	分散のモデル	性質
EII	λI	等方，等分散, 等幅
VII	$\lambda_k I$	等方，等分散, 異幅
EEI	λA	対角，等分散，等幅
VEI	$\lambda_k A$	対角，等分散，異幅
EVI	λA_k	対角，異分散，等幅
VVI	$\lambda_k A_k$	対角，異分散，異幅
EEE	$\lambda D A D^T$	楕円，等分散，等幅
EEV	$\lambda D_k A D_k^T$	楕円，回転同値，等幅
VEV	$\lambda_k D_k A D_k^T$	楕円，回転同値，異幅
VVV	$\lambda_k D_k A_k D_k^T$	楕円，異分散，異幅

どのモデルがデフォルトで用いられるかは reference を参照．


EII
2.5
1.5
0.5
−0.5
−2
−1
0
1


VII
2.5
1.5
0.5
−0.5
−2
−1
0
1


EEE
2.5
1.5
0.5
−0.5
−2
−1
0
1


VVV
2.5
1.5
0.5
−0.5
−2
−1
0
1

クラスタ数設定

クラスタ数 K はいくつ？

クラスタ数が事前にわからない場合，データから推定する必要がある．

R の関数 Mclust では「BIC」が最良のモデル (クラスタ数) を用いる．

BIC：これまで観測したデータと同じ分布にしたがうデータを独立に発生させた時に，それにどれだけよく当てはまるかを推定した量．AIC と同様，過適合を防ぐためのモデル選択規準．

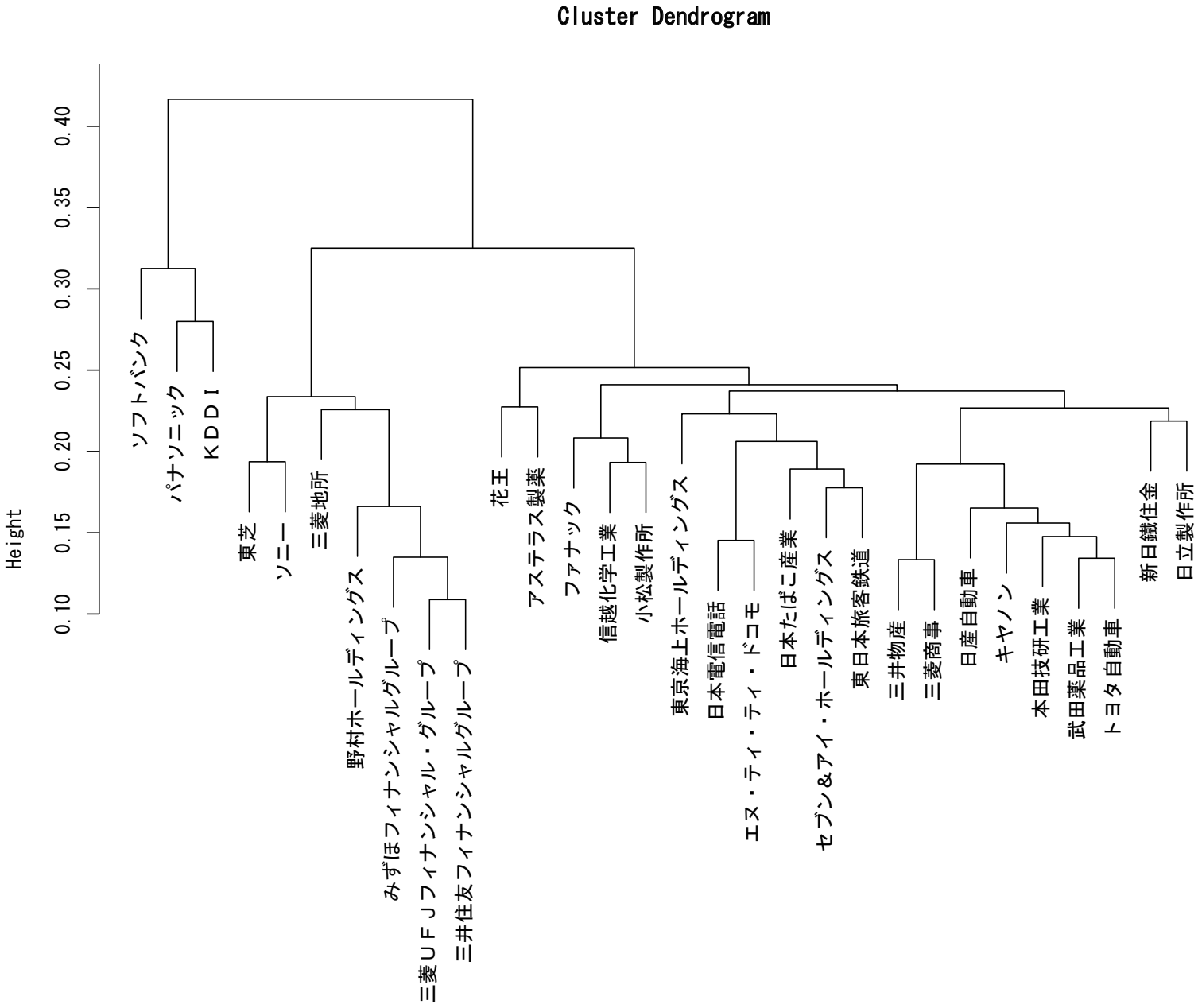
本来は混合モデルに安易に BIC を当てはめるのは間違いだが，
Mclust ではこれがデフォルトになっている．

構成

TOPIX CORE 30 銘柄時系列

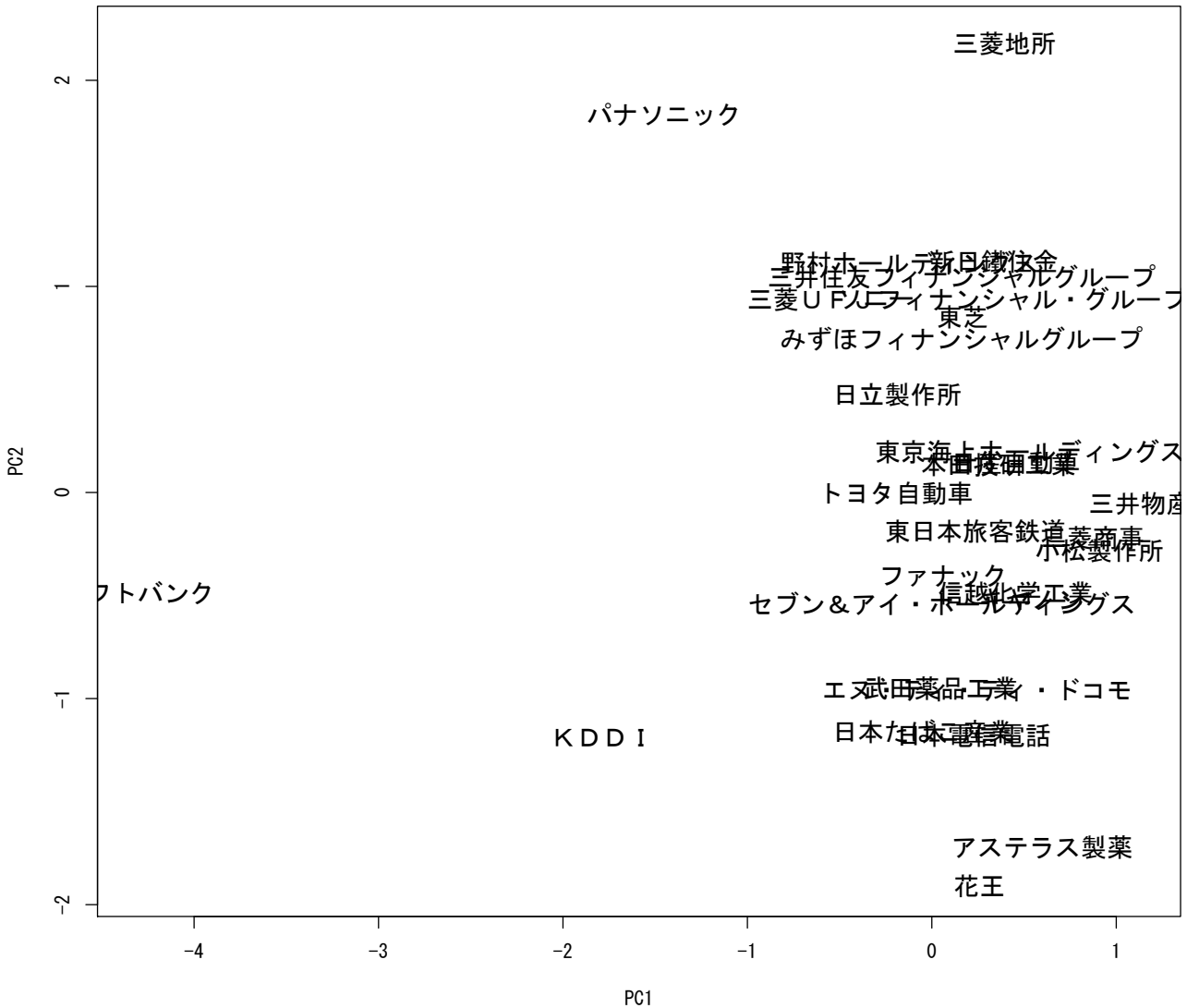
- TOPIX CORE 30 指標に含まれる 30 銘柄.
- 過去 250 日分 (2013/6/28–2014/7/4) の始値，安値，高値，終値 (日足) が格納されている.
- 今回は始値/終値を日にちごとに計算し，時系列を構成.

TOPIX CORE 30: 階層的クラスタリング



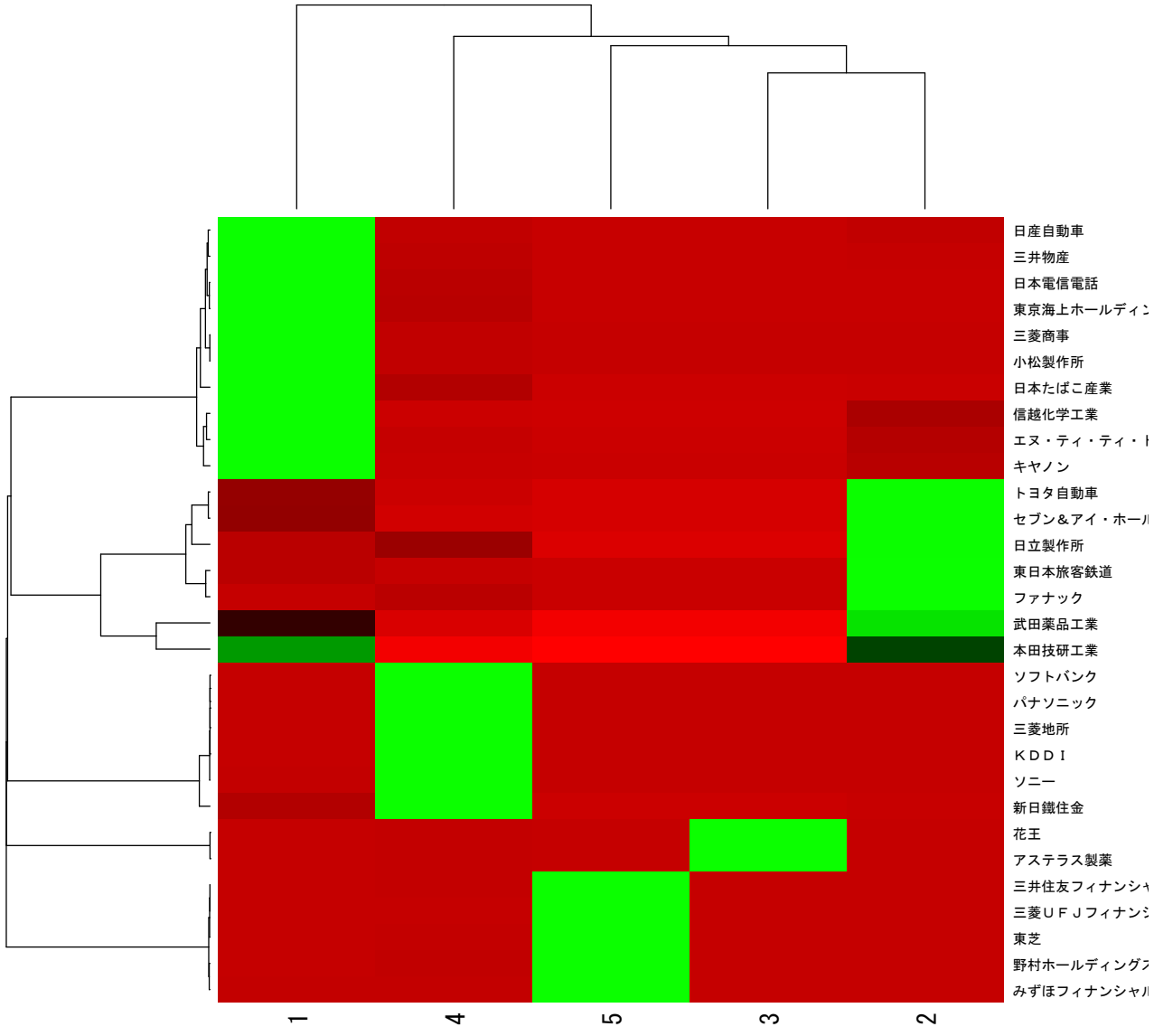
TOPIX CORE 30: 主成分分析

混合ガウスを当てはめるのに 250 次元は高すぎるので，主成分分析で低次元に落としてからクラスタリング.



TOPIX CORE 30: 混合ガウス分布

第三主成分スコアまでを用いてクラスタリング.



5 つのクラスタへの寄与率を計算. そのヒートマップをプロット.

講義情報ページ

http://www.is.titech.ac.jp/~s-taiji/lecture/2015/dataanalysis/dataanalysis.html