

Kitakado's Lecture Series

Clustering methods: Part (1) Lecture

Toshihide Kitakado (TUMSAT)

Released on 2021/01/17



このスライドは部分的に次の授業素材として共通利用しています。一部、日本語と英語が混ざっていますが御容赦下さい。

- 「生物資源モデリング」 (学部)
- 「生物資源解析学演習」 (学部)
- 「資源動態・管理学(E)」 (大学院)
- 「データサイエンス概論(E)」 (卓越大学院)



Machine Learning 機械学習

機械学習とは？ データから関係性や特徴を学ぶこと.

- 教師あり学習 (Supervised learning):
 - 入力 x と出力 y のようにデータがペアになっている
 - この入出力データから関係性を見出す, あるいは入力 x から y を予測する為に学習することが目的
 - y が離散値のとき, 判別または分類とよぶ
 - y が連続値のとき, 回帰分析とよぶ
- 教師なし学習 (Unsupervised learning):
 - 入力 x のみがデータとして観測される
 - データの次元縮小やクラスタリングなどを通して, データの特徴を取り出すことが目的
- 強化学習 (Reinforcement learning):
 - 利得が最大になるように行動を逐次的に選択するルールを学習することが目的
 - 動的最適法, マルコフ決定過程, 深層強化学習など



For better prediction of future observations.

- 回帰分析 (Regression)
 - Linear regression
 - Non-linear regression
 - Kernel smoothing
 - Neural network
 - Regression tree
 - Lasso, ridge, elastic net
 - ..
- 分類 (Classification)
 - Linear discriminant analysis
 - Nonlinear discriminant analysis
 - Support vector classification
 - Random forest
 - ...



教師なし学習 (unsupervised learning)

教師なし学習のタイプ

Principal Component Analysis (PCA) 主成分分析:

- a method used for dimension reduction of data, visualization of data, and pre-data-processing
- データの次元縮小, 可視化, 事前処理などの目的で利用

Multi-dimension scaling (MDS) 多次元尺度法:

- a method used for 2 or 3-dim visualization of data by dimension reduction
- 次元縮小によりデータを2次元平面あるいは3次元空間で可視化する方法

Clustering クラスタリング:

- a broad class of methods for discovering “unknown” subgroups in data
- データの類似度からグルーピングを行い, データの特徴を取り出す方法)
- ここではクラスタリング法について紹介します.



Clustering and Classification (分類の違い)

Clustering:

- The aim of clustering is to look for homogeneous subgroups among the observations using the concept of “similarity” and “dissimilarity” (データの類似度から、似た者通しのグループに分割する方法)

The clustering is not same as the classification:

- the classification is one of “Supervised learning” and requires the paired information of predictors and label (分類は教師となるデータがある。例えば、サンプルはキハダかメバチのどちらかで、答えの分かっている形態データから分類方法を作成したい)
- the clustering is one of “Unsupervised learning” and does not assume the availability of information of labels (クラスタリングは教師となるデータがない。どの種のデータか分からない形態データから似た者通しのグループに分けたい)



Clustering クラスター法

Clustering methods

Several methods for the clustering:

- k-means method (k平均法)
- Hierarchical clustering method (階層的クラスタリング法)
- Model-based clustering method (モデルベースのクラスタリング法)
- ...



k-means method (k平均法)

Concept:

- To partition a data set into K distinct and non-overlapping clusters (K個のクラスターにデータを分割したい)
- The number of clusters is pre-specified (Kの値は事前にきめている)
- Each observation is assigned exactly to one of K-clusters (各データはK個のクラスターのどこか1つに含まれる)
- The idea behind the K-means method is “a good clustering is one for which the within cluster variation is as small as possible” (アイデアとしては、クラスター内の違いがなるべく小さくなるような均質クラスターを作りたい)



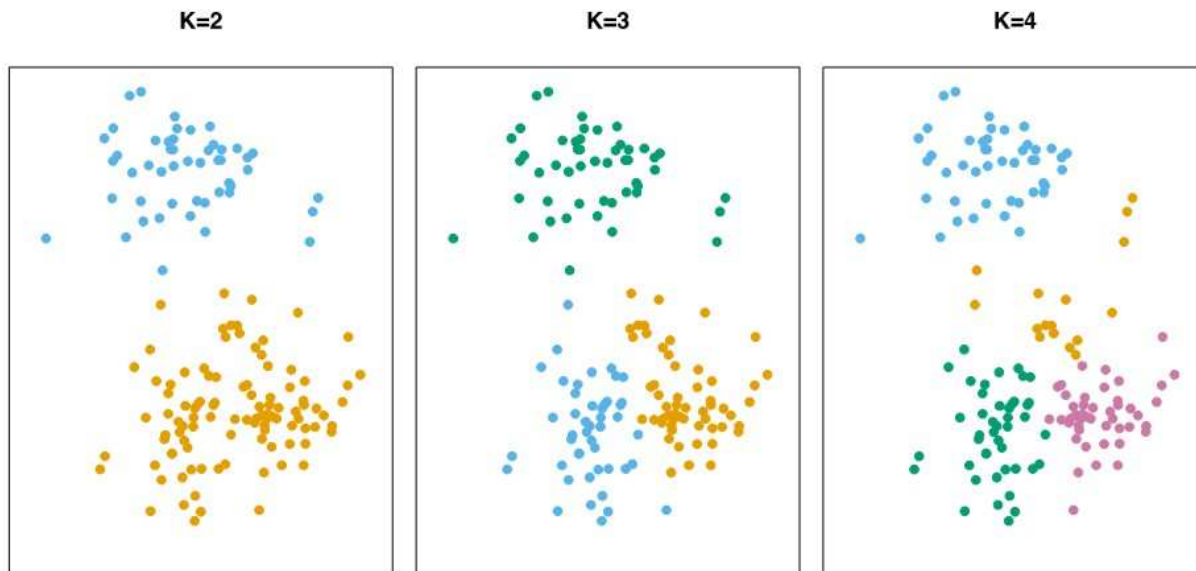


Image of clustering (from ISLR)



Mathematically speaking... (1)

- n : the number of observations (データの数)
- $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ (p is the dimension of variable, データの次元)
- C_k : a set of indices of observations $\{1, 2, \dots, n\}$
- $|C_k|$: C_k の大きさ

$$1) C_1 \cup C_2 \cup \dots \cup C_K = \{1, 2, \dots, n\}$$

$$2) C_k \cap C_{k'} = \emptyset \quad \text{for } k \neq k'$$

- 例えば $n = 8$ で, $C_1 = \{1, 2, 4, 6, 7\}$, $C_2 = \{3, 5, 8\}$ などのようにインデックスの部分集合
- $|C_1| = 5, |C_2| = 3$



Mathematically speaking... (2)

Then, to find a good partition, the total within-cluster variation $\sum_{k=1}^K W(C_k)$ is minimized, where $W(C_k)$ is the within-cluster variation for cluster C_k and normally defined as

$$W(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} d(x_i, x_{i'}),$$

where d is the distance between two observations and typically the following Euclidean distance is used.

$$d(x_i, x_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2.$$

- クラスタ C_k の変動を $W(C_k)$ で定義し、その合計が $\sum_{k=1}^K W(C_k)$ となるような C_1, C_2, \dots, C_K の組み合わせを探索する



For finding a local minimum:

- 1) Randomly assign a number from 1 to K to each of n observations
(n このデータをランダムに1から K のグループに割り当てる)
- 2) For each of the K clusters, compute the cluster centroid
(上記の仮のクラスター毎に, データの中心を計算する)
- 3) Assign each observation to the cluster whose centroid is closest
(ステップ2で K 個の中心ができ, n 個のデータを一番近い中心のクラスターに割り当てなおす)
- 4) Repeat until the centroids become stable
(中心が変動が小さくなり安定するまで, この2と3を繰り返す)



Graphically speaking...

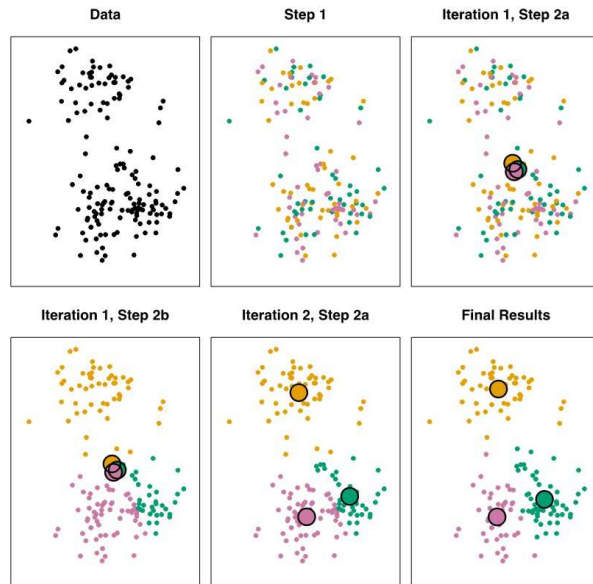


Image of clustering (from ISLR)



Algorithm (Continued)

For finding the global minimum:

- Use multiple initial assignments
- Select the best for which the target function $\sum_{k=1}^K W(C_k)$ is smallest
- 最初の割り当てに結果が依存するため、よりよい分割を求めて、何通りか初期割り当てを変えてみて、目的とする $\sum_{k=1}^K W(C_k)$ の値が最小となる分割を探索する。



Graphically speaking...



Image of clustering (from ISLR)



Example data (Iris data アヤメ データ)

The following 3 difference species 次の3種のアヤメ データ (50本ずつ)

- Setosa
- Versicolor
- Virginica

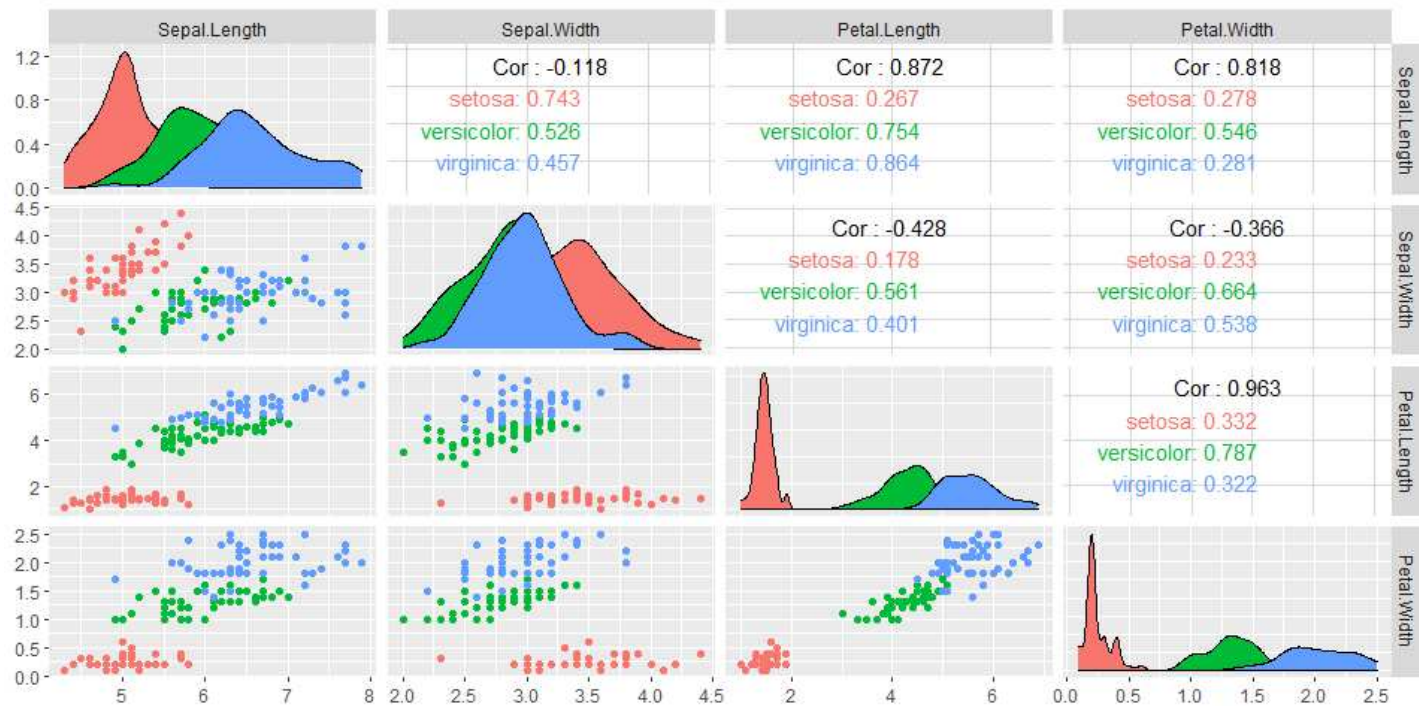
Data (4 measurements 4次元の測定値, $3 \times 50 = 150$ individuals 150本の花)

- Length and width of "Sepal" (がく 片の長さ と 幅)
- Length and width of "Petal" (花弁の長さ と 幅)

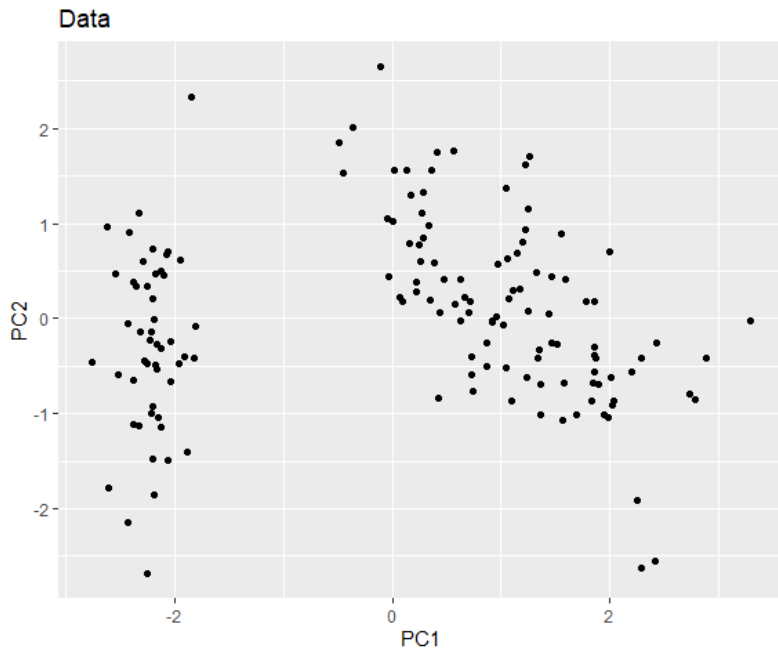
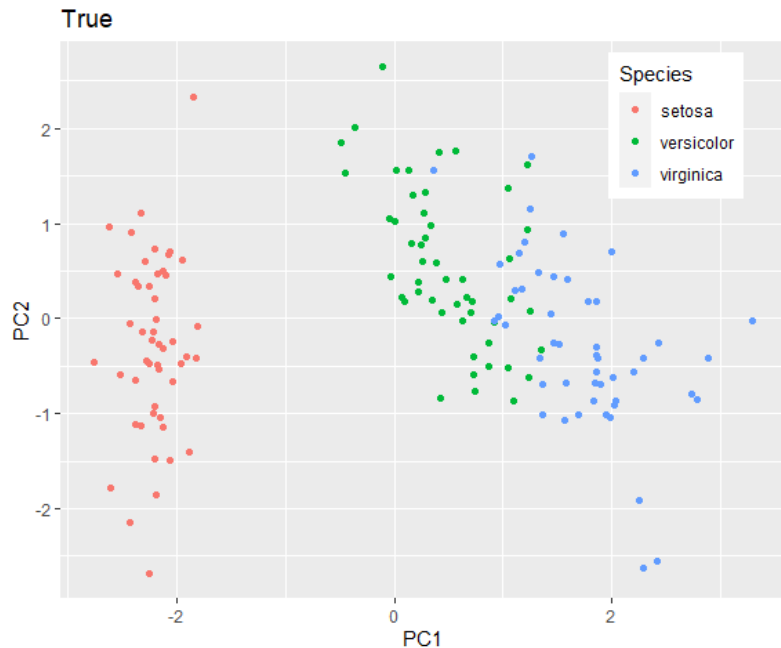
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica



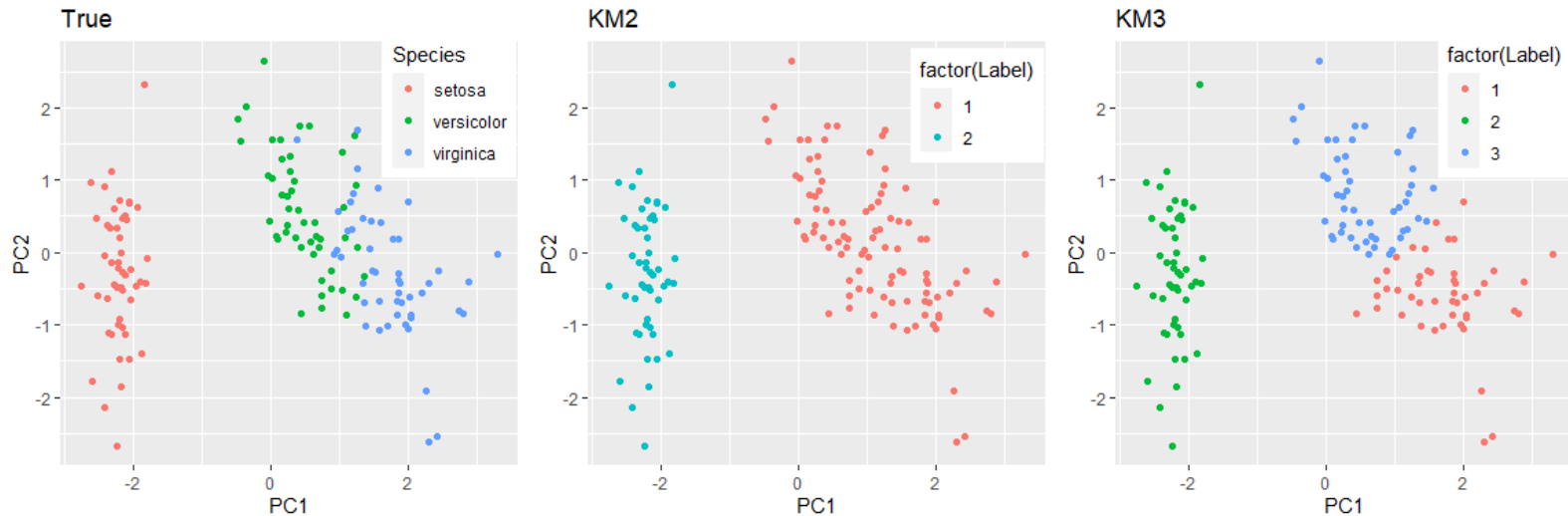
Data visualization データ可視化



Data visualization with dimension reduction 次元縮小で可視化



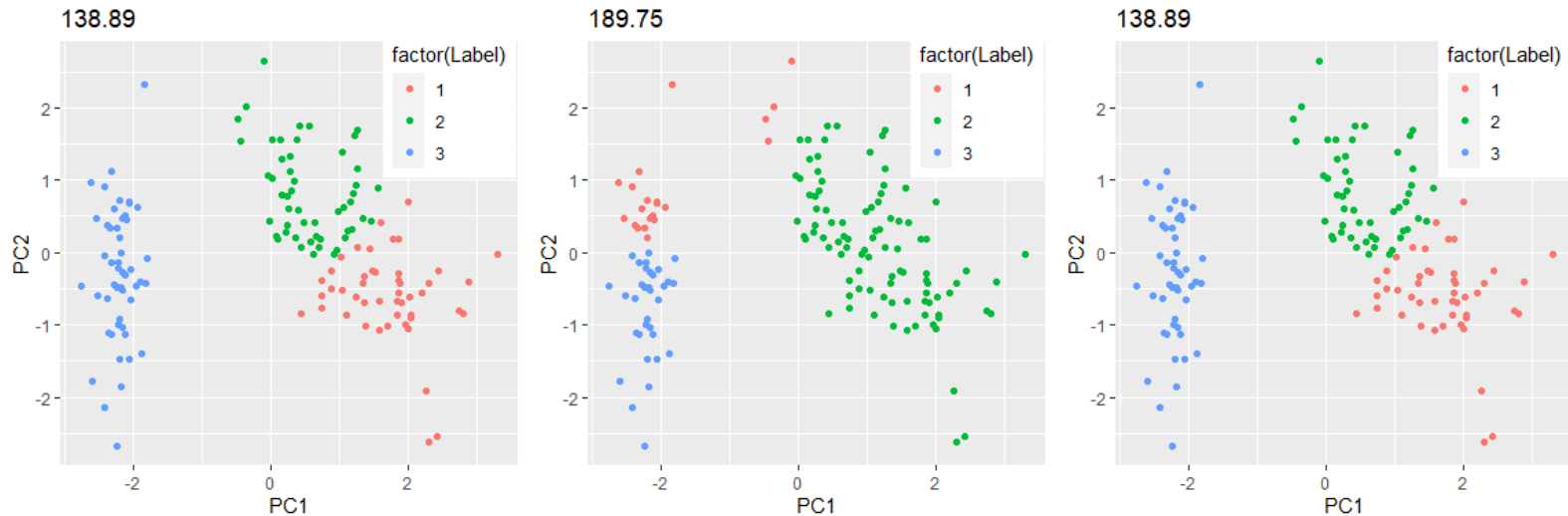
Clustering: K-means method with K=2 & 3 (10 repeats)



A total of 10 initial assignments were used. (10通りの初期割り当てで10回計算, 最良の結果を選択)



Clustering: K-means method with K= 3 (only one initial assignment)



もし初期割り当てを1度しか行わないと、アルゴリズムが変な分割を提示してしまうこともあるので、前シートのように何度か繰り返すことが重要

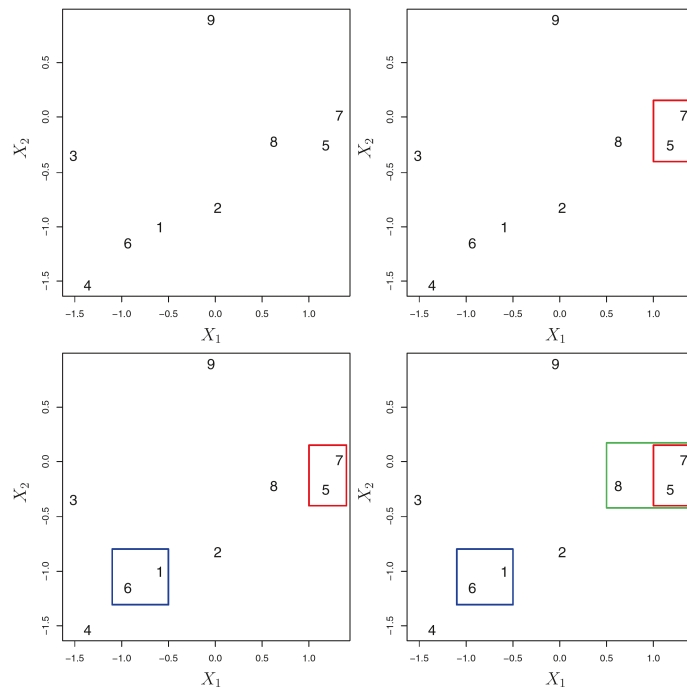


Hierarchical clustering method (階層的 クラスタリング法)

Hierarchical clustering method

Concept:

- Bottom-up clustering
- 近いものを順番にグループ化していく



Sequence of grouping (from ISLR)



Hierarchical clustering

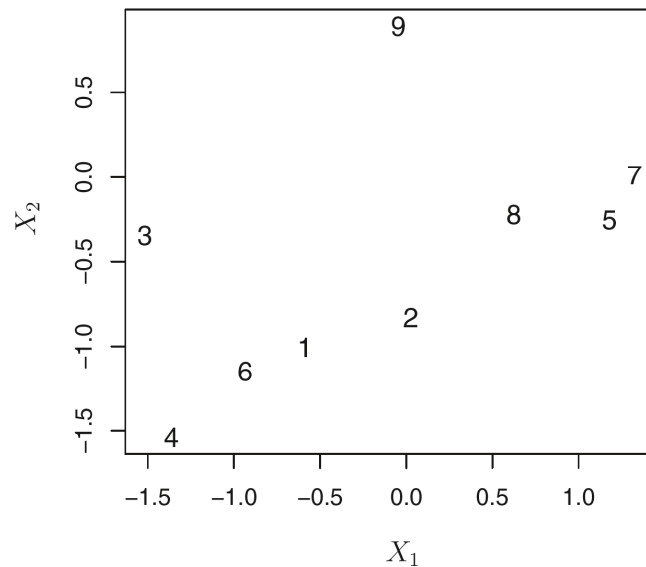
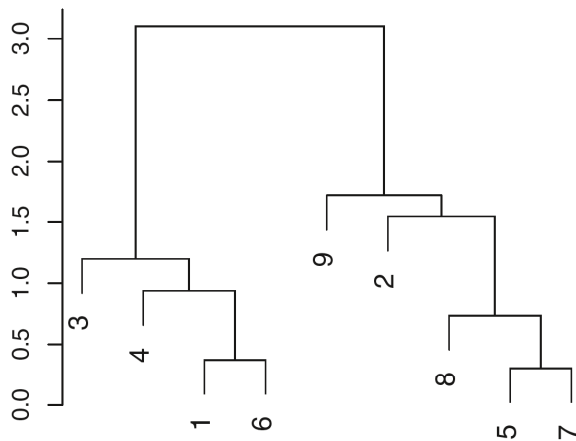


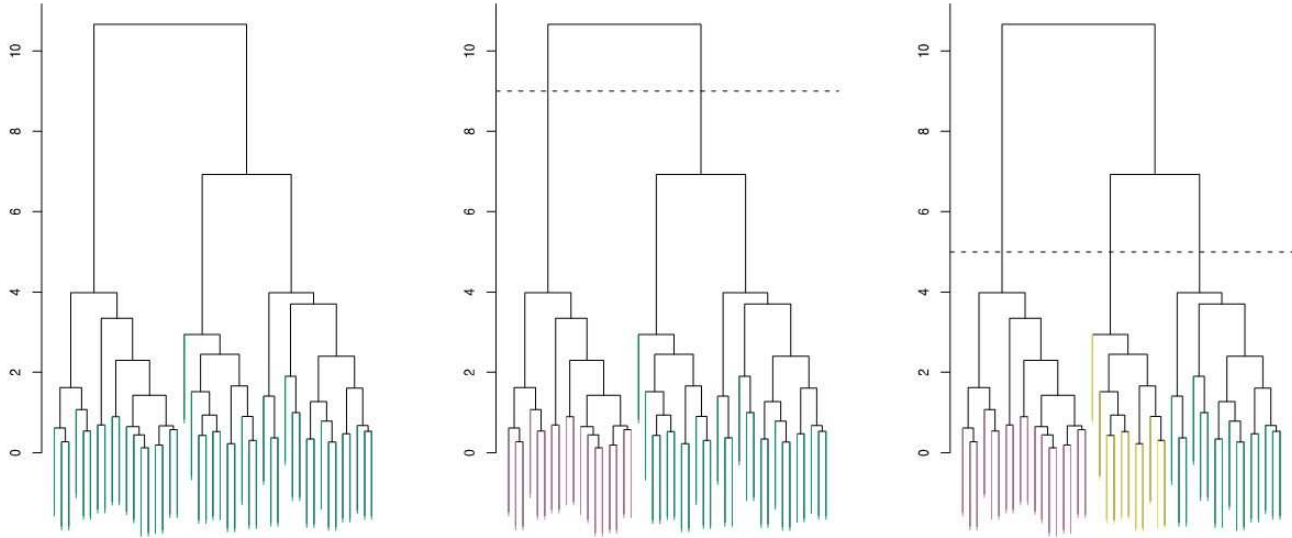
Image of dendrogram after hierarchical clustering (from ISLR)



- Choice of similarity metric (距離の定義の仕方)
 - Euclidean distance (ユークリッド距離)
 - Correlation (相関)
- Choice of linkage (複数の点からなるクラスター間の距離の定義)
 - single (最短)
 - complete (最長)
 - average (平均)
 - centroid (中心)
 - ...
- Choice of K (いくつのクラスターに分けるか?)



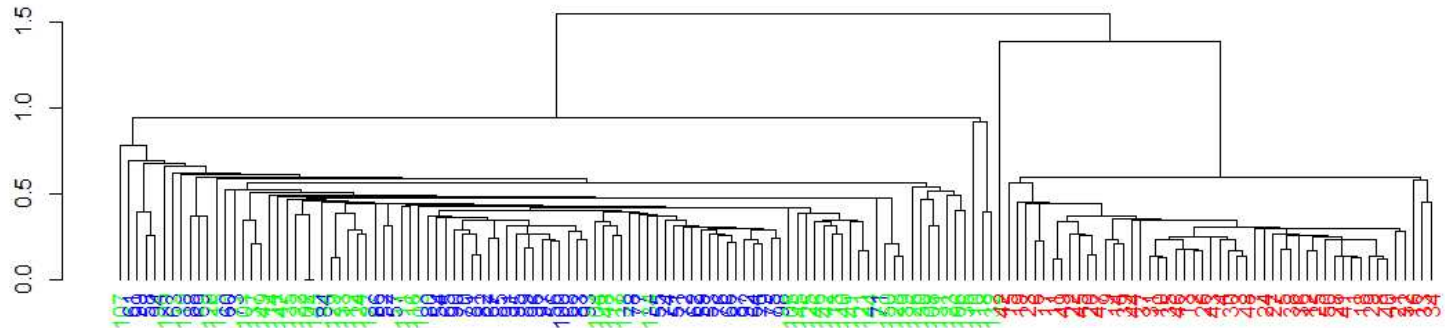
Outcomes as dendrogram (系統樹)



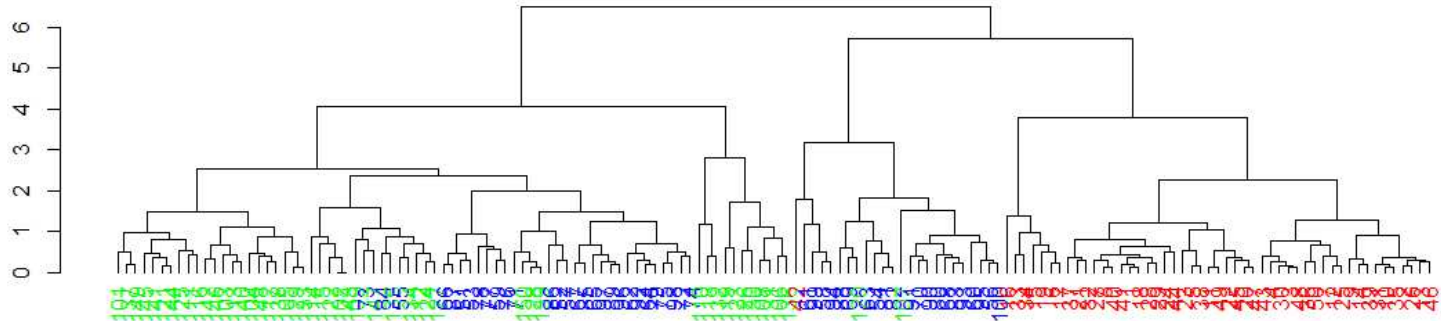
Grouping after hierarchical clustering (from ISLR)



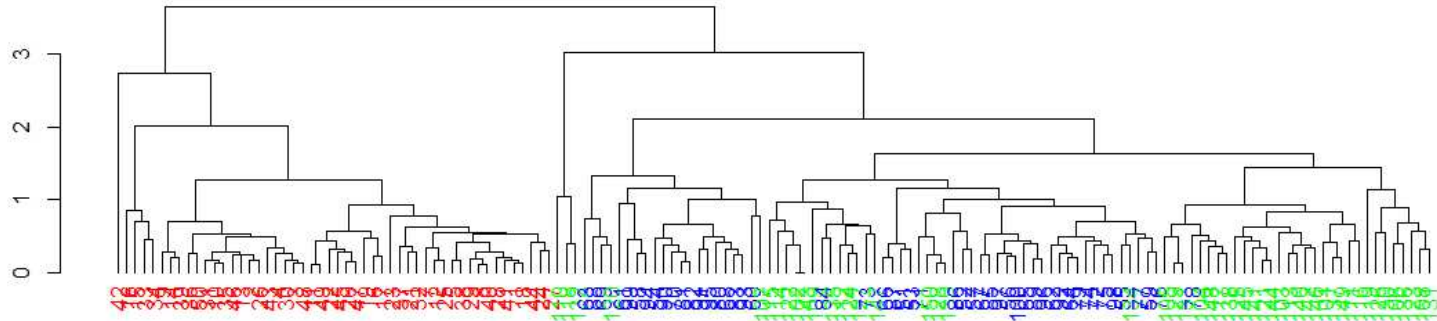
Iris again: Hierarchical clustering with “single” linkage



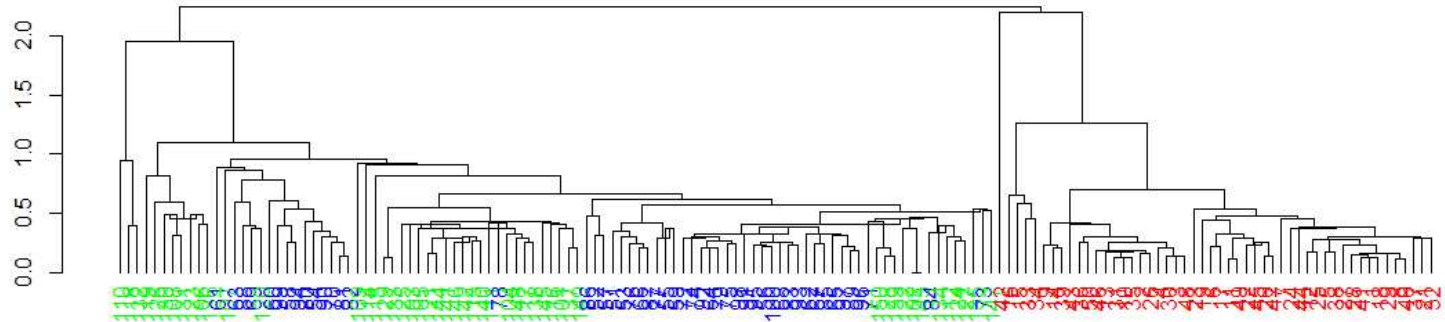
Iris again: Hierarchical clustering with “complete” linkage



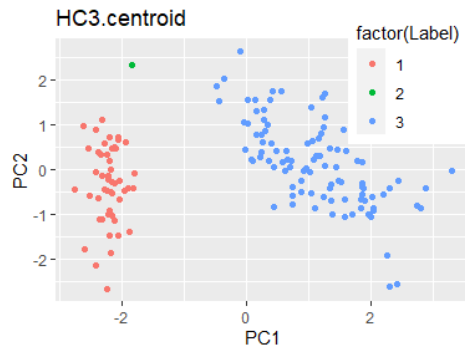
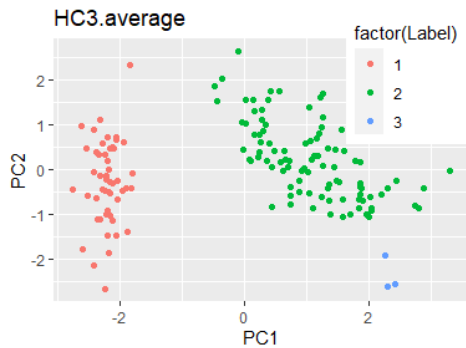
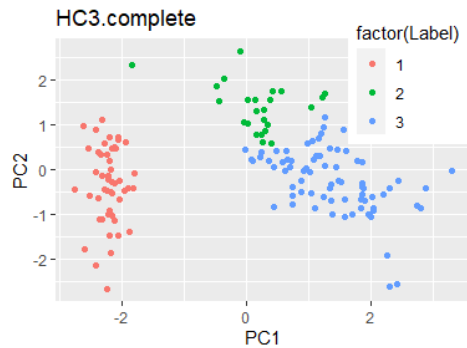
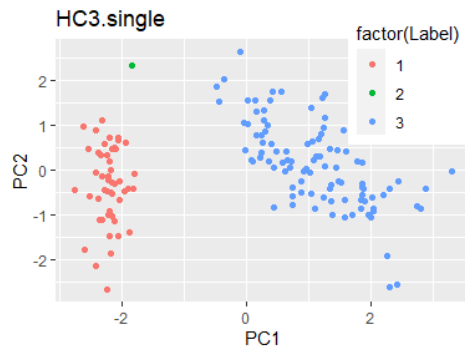
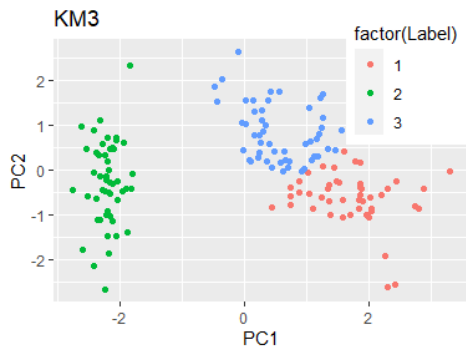
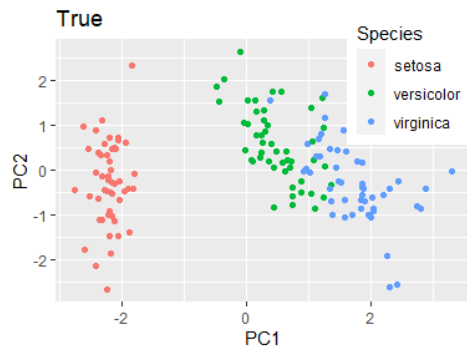
Iris again: Hierarchical clustering with “average” linkage



Iris again: Hierarchical clustering with “centroid” linkage



Iris again: Comparison of results



Model-based clustering method (モデルベースのクラスタリング法)

“Model-based clustering” is another approach using model-based framework.

- More probability and statistical descriptions with probability distributions and estimation procedure
- The method can deal with
 - determination of the number of clusters as a model selection
 - characterization of clusters through definition of probability distributions



(1-dim) Finite mixture model

Let us assume that

- you have data consisting of n observations of one dimensional measurement, as y_1, y_2, \dots, y_n from a population
- there are multiple potential groups (say, K) in the population
- each observation belongs to one of K groups
- z_i is an indicator variable as a label which the i -th individual belongs to
- $z_i = k$ means th i -th individual belongs to the k -th group
- we do don't know the cluster patterns, and therefore Z_i is a latent (unobserved) variable

Let's further assume that

- each group has its own probability distribution for the measurement as $f_k(y|\theta_k)$



(1-dim) Finite mixture model (continued)

We suppose that for $i = 1, 2, \dots, n; k = 1, 2, \dots, K$,

- $y_i | z_i = k \sim f_k(y | \theta_k)$
- $P(Z_i = k) = \pi_k \ (\geq 0)$

Then the marginal distribution of y_i is a mixture distribution expressed as follows:

$$f(y_i) = \sum_{k=1}^K \pi_k f_k(y | \theta_k) \quad (i = 1, 2, \dots, n)$$



Example: mixture of length of fish

Example

- You observe a length of fish for n samples
- There are multiple ages of fish in the population
- Therefore the observed length composition consists of multiple ages of fish
- We shall consider the following distribution for each age group:

$$f_k(y|\theta_k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_k)^2}{\sigma_k^2}\right\}$$

- then the marginal distribution is a normal mixture model as

$$f(y_i) = \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu_k)^2}{\sigma_k^2}\right\}$$

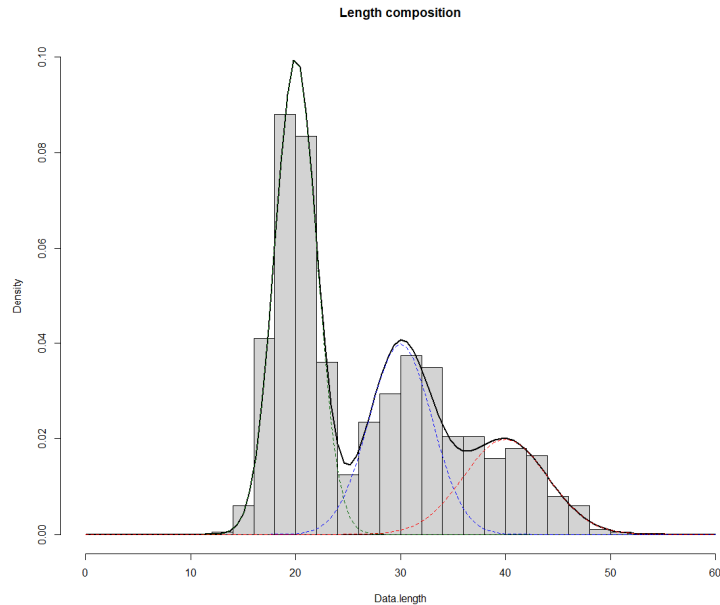


- Estimation by the maximum likelihood method via EM-algorithm
 - E-step: Calculate conditional probability of assignment via conditional expectation
 - M-step: Maximize the likelihood determined with expected mean of latent variables
- Model selection for choosing the number of groups
 - Use information criteria
 - In a famous R package “mclust”, BIC is used (but for maximization since the sign is different from the normal BIC definition)
 - The determination of the number of groups is not easy though



Mixture of 3 different age groups

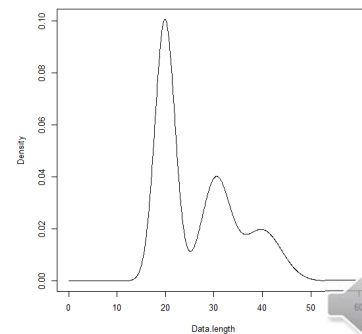
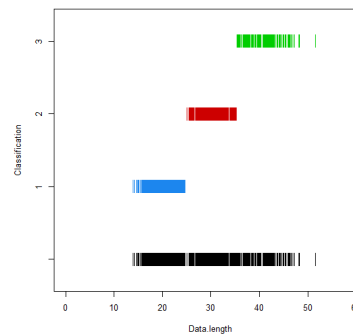
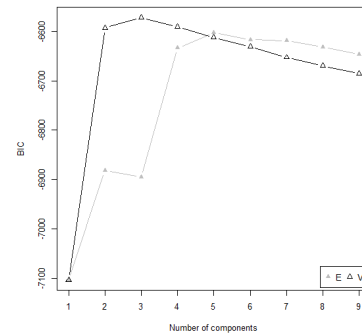
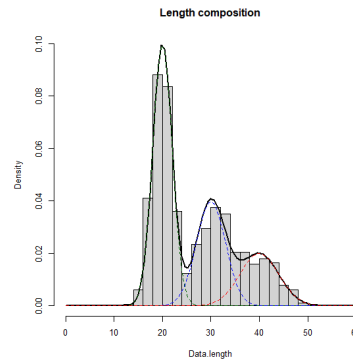
```
mu <- c(20, 30, 40)
sigma <- c(2, 3, 4)
pi <- c(0.5, 0.3,
0.2)
Nsample <- 1000
```



Estimation and model selection

“Model” for univariate distribution

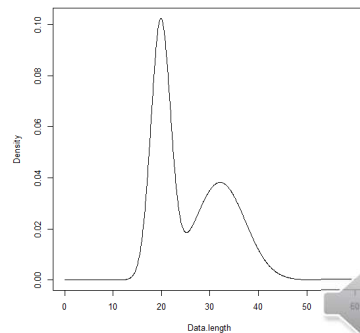
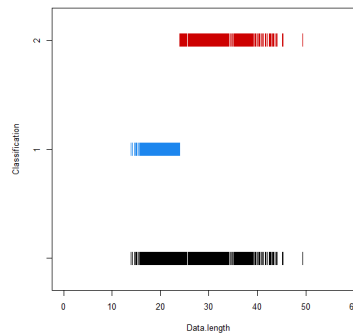
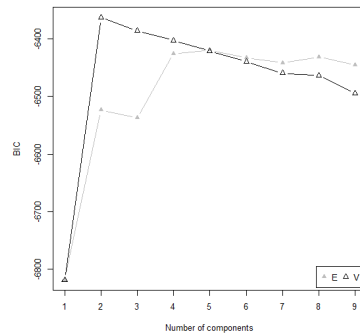
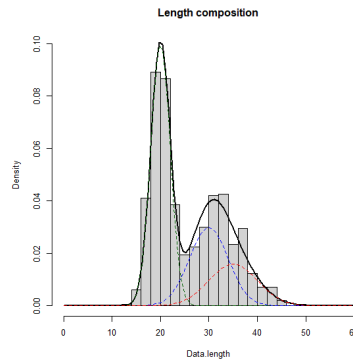
- “E”: equal volume (common variance)
- “V”: equal volume (different variances)



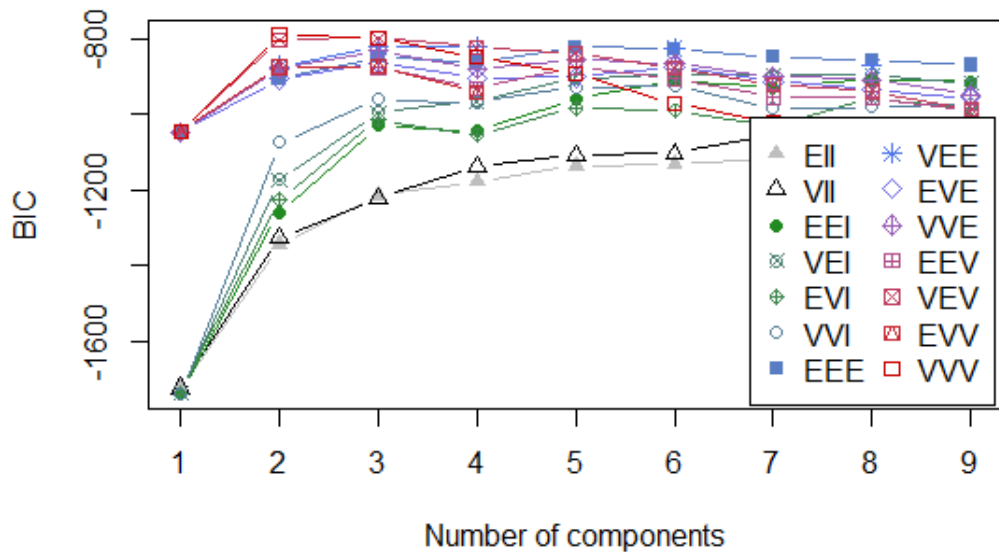
BUT...

In real examples, the determination of the number of groups is not so easy.
See example below:

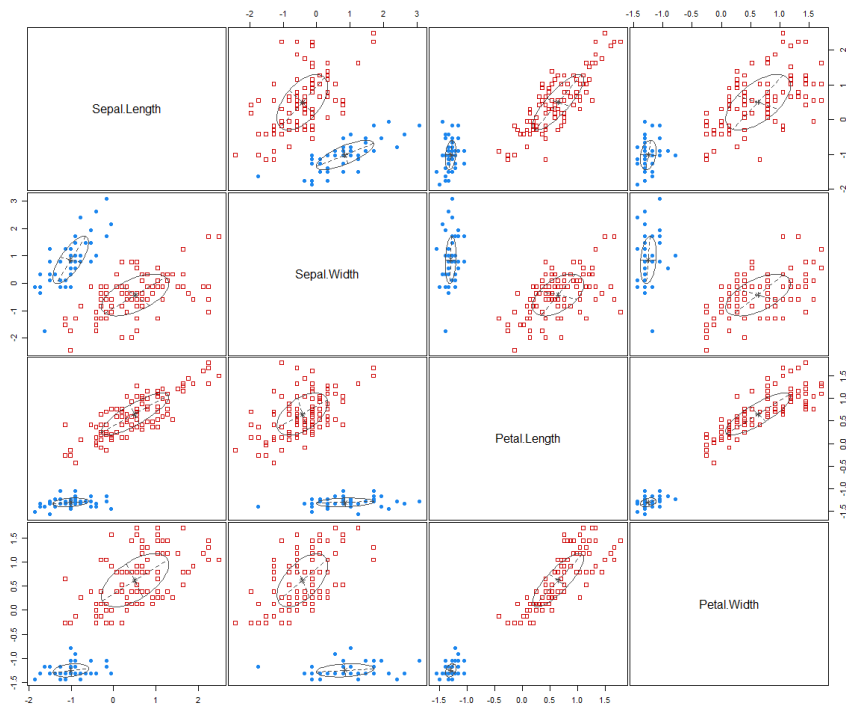
```
mu <- c(20, 30, 35)  
sigma <- c(2, 4, 5)  
pi <- c(0.5, 0.3,  
0.2)
```



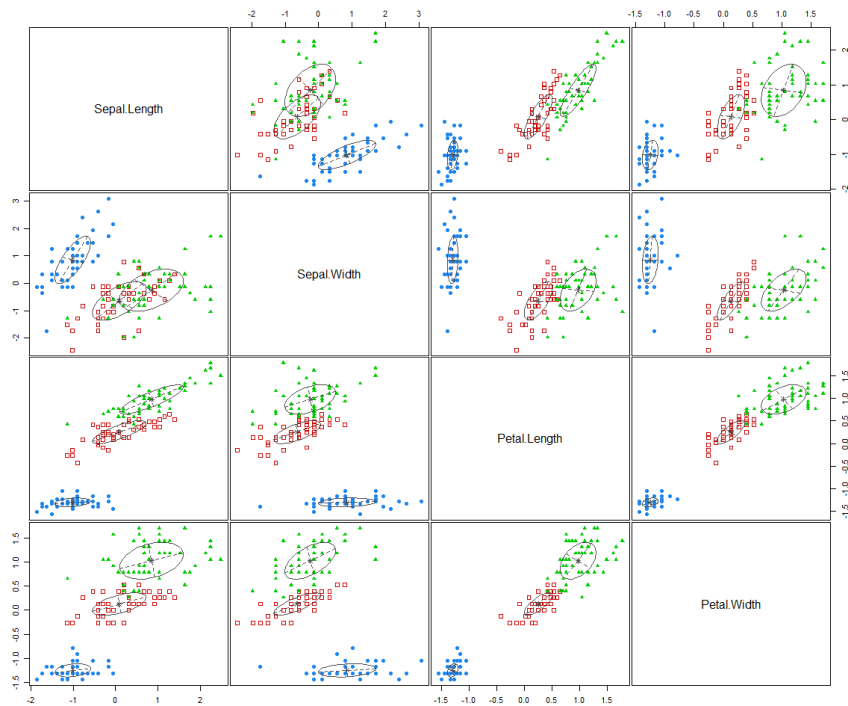
Application of mixture models to iris data (1) Best model (G=2)



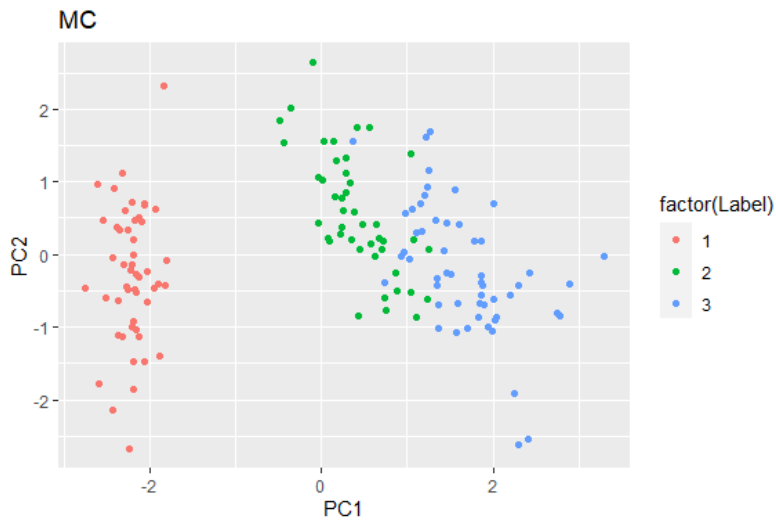
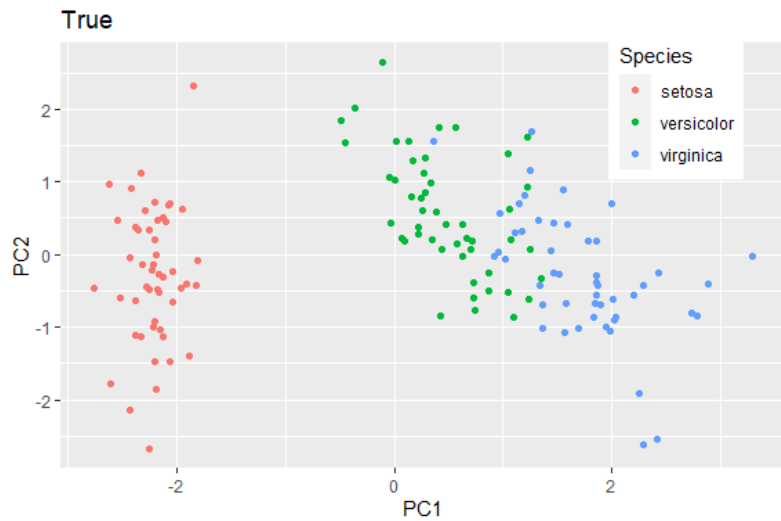
Mixture models (1 Continued) Best model (G=2)



Mixture models (2) Model with $G=3$



Mixture models (3) plots



Summary

What have we learned?

- Ideas of clustering were introduced
- The two methods shown here are just basic methods, and these have been extended to several directions
- Here is a set of useful references:
 - Hastie et al. (2016) "The Elements of Statistical Learning"
 - James et al. (2013) "An Introduction to Statistical Learning with applications in R (2nd ed). [Freely downloadable from <https://statlearning.com/>]
 - 金森敬文 (2017) "Rによる機械学習入門"

