

# ETC3250/5250: k-means clustering

Semester 1, 2020


Professor Di Cook


Econometrics and Business Statistics

Monash University

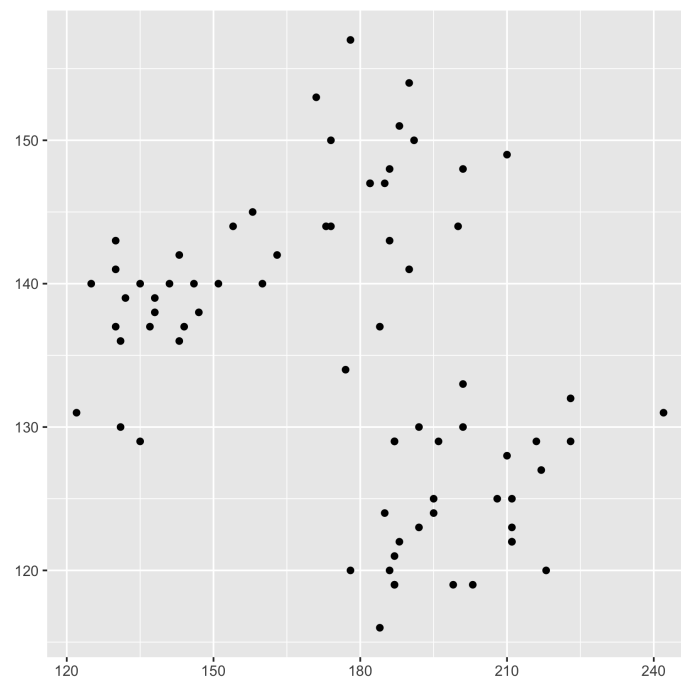
Week 10 (b)

## Cluster analysis

 The aim of cluster analysis is to group cases (objects) according to their similarity on the variables. It is also often called unsupervised classification, meaning that classification is the ultimate goal, but the classes (groups) are not known ahead of time.

 Hence the first task in cluster analysis is to construct the class information. To determine closeness we start with measuring the interpoint distances.

Cluster this!



## k-means clustering - algorithm

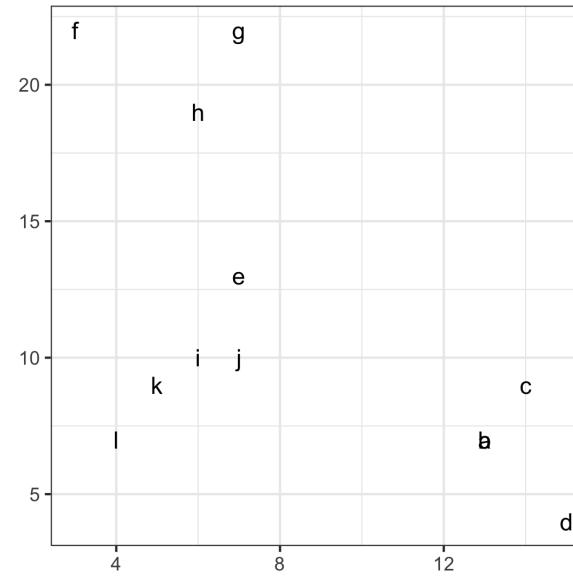
This is an iterative procedure. To use it the number of clusters,  $k$ , must be decided first. The stages of the iteration are:

- Initialize by either (a) partitioning the data into  $k$  groups, and compute the  $k$  group means or (b) an initial set of  $k$  points as the first estimate of the cluster means (seed points).
- Loop over all observations reassigning them to the group with the closest mean.
- Recompute group means.
- Iterate steps 2 and 3 until convergence.

Thean C. Lim's blog post

Some data 🧑

lbl	x1	x2
a	13	7
b	13	7
c	14	9
d	15	4
e	7	13
f	3	22
g	7	22
h	6	19
i	6	10
j	7	10
k	5	9
l	4	7



Select  $k = 2$ , and set initial seed means  
 $\bar{x}_1 = (8, 13)$ ,  $\bar{x}_2 = (1, 4)$   
Compute distances  $(d_1, d_2)$  between each  
observation and each mean.

lbl	x1	x2	d1	d2
a	13	7	7.8	12.4
b	13	7	7.8	12.4
c	14	9	7.2	13.9
d	15	4	11.4	14.0
e	7	13	1.0	10.8
f	3	22	10.3	18.1
g	7	22	9.1	19.0
h	6	19	6.3	15.8
i	6	10	3.6	7.8
j	7	10	3.2	8.5
k	5	9	5.0	6.4
l	4	7	7.2	4.2

Select  $k = 2$ , and set initial seed means  
 $\bar{x}_1 = (8, 13)$ ,  $\bar{x}_2 = (1, 4)$   
 Compute distances  $(d_1, d_2)$  between each  
 observation and each mean.

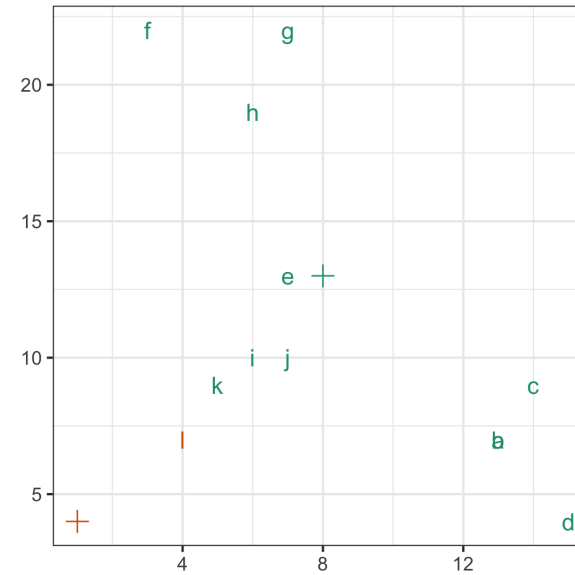
lbl	x1	x2	d1	d2
a	13	7	7.8	12.4
b	13	7	7.8	12.4
c	14	9	7.2	13.9
d	15	4	11.4	14.0
e	7	13	1.0	10.8
f	3	22	10.3	18.1
g	7	22	9.1	19.0
h	6	19	6.3	15.8
i	6	10	3.6	7.8
j	7	10	3.2	8.5
k	5	9	5.0	6.4
l	4	7	7.2	4.2

Assign the cluster membership

lbl	x1	x2	d1	d2	cl
a	13	7	7.8	12.4	1
b	13	7	7.8	12.4	1
c	14	9	7.2	13.9	1
d	15	4	11.4	14.0	1
e	7	13	1.0	10.8	1
f	3	22	10.3	18.1	1
g	7	22	9.1	19.0	1
h	6	19	6.3	15.8	1
i	6	10	3.6	7.8	1
j	7	10	3.2	8.5	1
k	5	9	5.0	6.4	1
l	4	7	7.2	4.2	2

Assign the cluster membership

lbl	x1	x2	d1	d2	cl
a	13	7	7.8	12.4	1
b	13	7	7.8	12.4	1
c	14	9	7.2	13.9	1
d	15	4	11.4	14.0	1
e	7	13	1.0	10.8	1
f	3	22	10.3	18.1	1
g	7	22	9.1	19.0	1
h	6	19	6.3	15.8	1
i	6	10	3.6	7.8	1
j	7	10	3.2	8.5	1
k	5	9	5.0	6.4	1
l	4	7	7.2	4.2	2

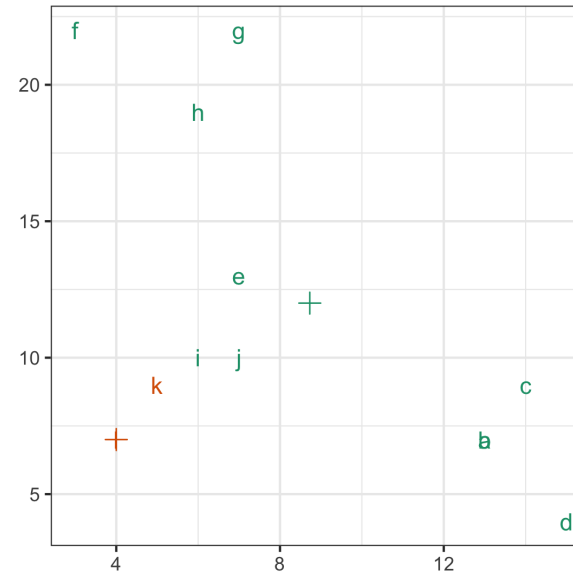




Recompute means, and re-assign the cluster membership

$$\bar{x}_1 = (9, 12), \bar{x}_2 = (4, 7)$$

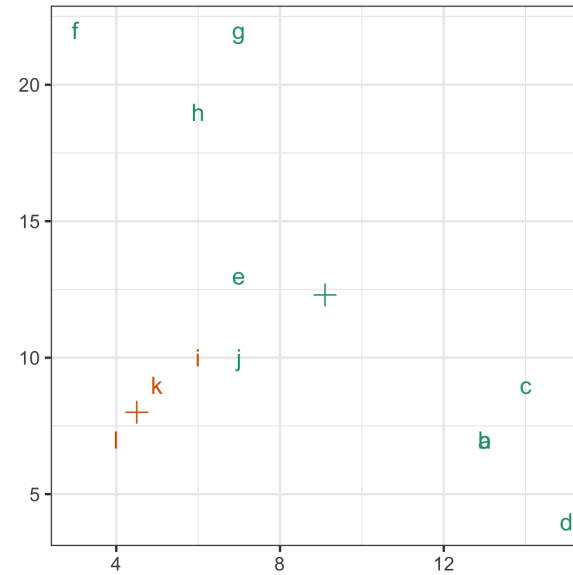
lbl	x1	x2	d1	d2	cl
a	13	7	6.6	9.0	1
b	13	7	6.6	9.0	1
c	14	9	6.1	10.2	1
d	15	4	10.2	11.4	1
e	7	13	2.0	6.7	1
f	3	22	11.5	15.0	1
g	7	22	10.1	15.3	1
h	6	19	7.5	12.2	1
i	6	10	3.4	3.6	1
j	7	10	2.6	4.2	1
k	5	9	4.8	2.2	2
l	4	7	6.9	0.0	2



Recompute means, and re-assign the cluster membership

$$\bar{x}_1 = (9, 12), \bar{x}_2 = (4, 8)$$

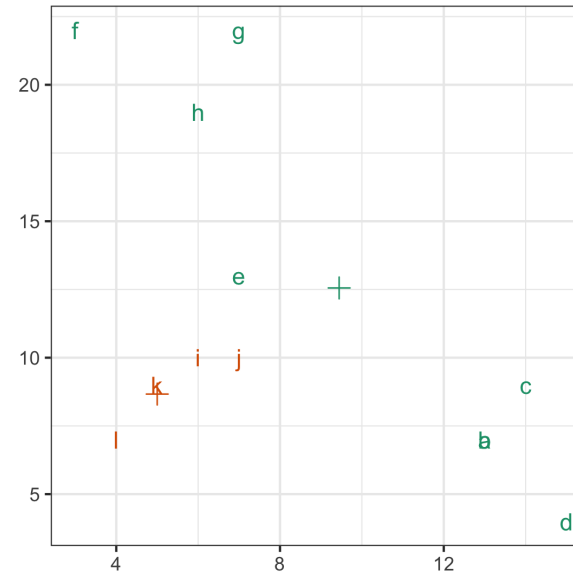
lbl	x1	x2	d1	d2	cl
a	13	7	6.6	8.6	1
b	13	7	6.6	8.6	1
c	14	9	5.9	9.6	1
d	15	4	10.2	11.2	1
e	7	13	2.2	5.6	1
f	3	22	11.5	14.1	1
g	7	22	9.9	14.2	1
h	6	19	7.4	11.1	1
i	6	10	3.9	2.5	2
j	7	10	3.1	3.2	1
k	5	9	5.3	1.1	2
l	4	7	7.4	1.1	2



Recompute means, and re-assign the cluster membership

$$\bar{x}_1 = (9, 13), \bar{x}_2 = (5, 9)$$

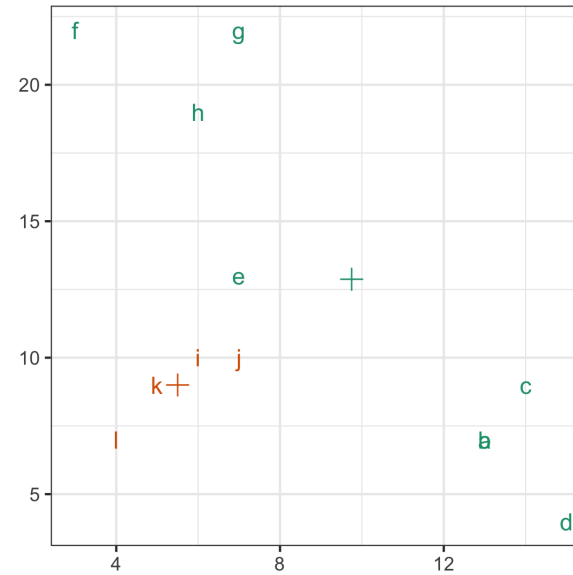
lbl	x1	x2	d1	d2	cl
a	13	7	6.6	8.2	1
b	13	7	6.6	8.2	1
c	14	9	5.8	9.0	1
d	15	4	10.2	11.0	1
e	7	13	2.5	4.8	1
f	3	22	11.4	13.5	1
g	7	22	9.8	13.5	1
h	6	19	7.3	10.4	1
i	6	10	4.3	1.7	2
j	7	10	3.5	2.4	2
k	5	9	5.7	0.3	2
l	4	7	7.8	1.9	2



Recompute means, and re-assign the cluster membership

$$\bar{x}_1 = (10, 13), \bar{x}_2 = (6, 9)$$

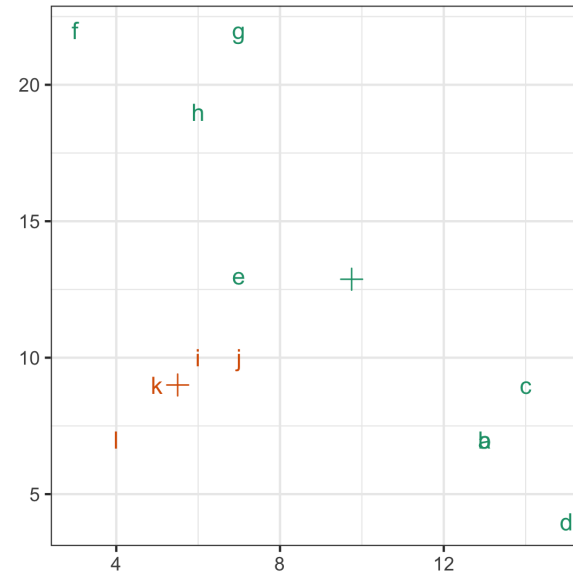
lbl	x1	x2	d1	d2	cl
a	13	7	6.7	7.8	1
b	13	7	6.7	7.8	1
c	14	9	5.8	8.5	1
d	15	4	10.3	10.7	1
e	7	13	2.8	4.3	1
f	3	22	11.4	13.2	1
g	7	22	9.5	13.1	1
h	6	19	7.2	10.0	1
i	6	10	4.7	1.1	2
j	7	10	4.0	1.8	2
k	5	9	6.1	0.5	2
l	4	7	8.2	2.5	2



Recompute means, and re-assign the cluster membership

$$\bar{x}_1 = (10, 13), \bar{x}_2 = (6, 9)$$

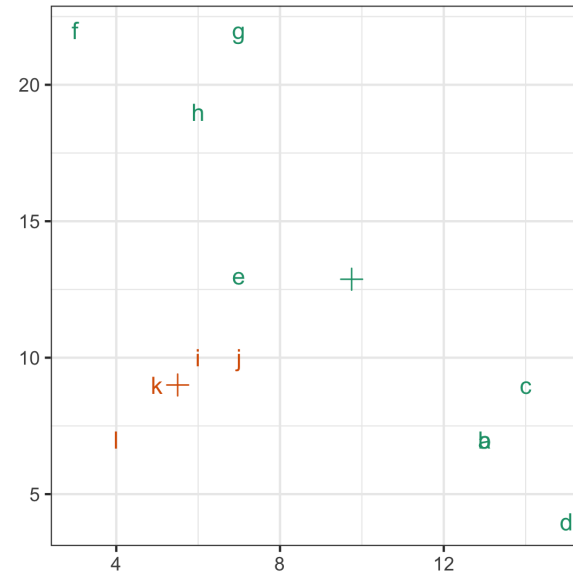
lbl	x1	x2	d1	d2	cl
a	13	7	6.7	7.8	1
b	13	7	6.7	7.8	1
c	14	9	5.8	8.5	1
d	15	4	10.3	10.7	1
e	7	13	2.8	4.3	1
f	3	22	11.4	13.2	1
g	7	22	9.5	13.1	1
h	6	19	7.2	10.0	1
i	6	10	4.7	1.1	2
j	7	10	4.0	1.8	2
k	5	9	6.1	0.5	2
l	4	7	8.2	2.5	2



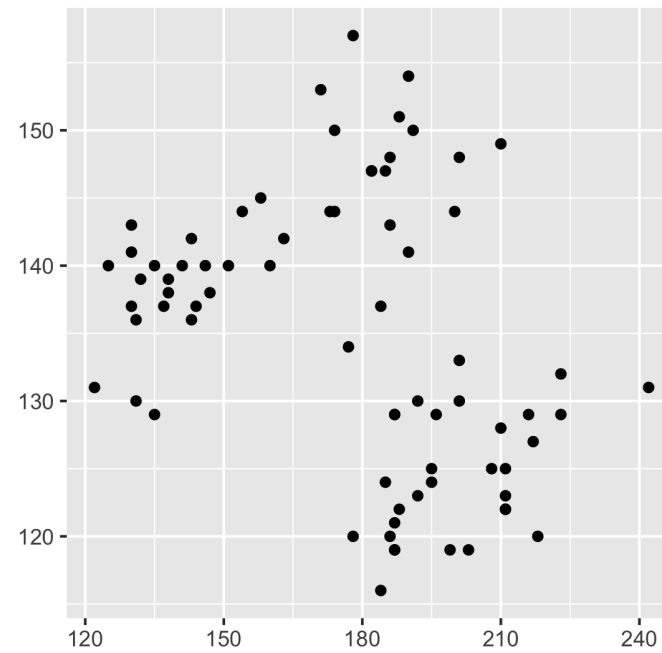
Recompute means, and re-assign the cluster membership

$$\bar{x}_1 = (10, 13), \bar{x}_2 = (6, 9)$$

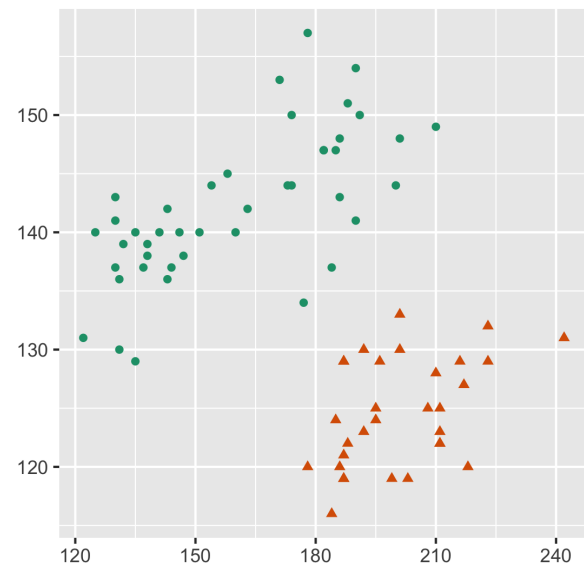
lbl	x1	x2	d1	d2	cl
a	13	7	6.7	7.8	1
b	13	7	6.7	7.8	1
c	14	9	5.8	8.5	1
d	15	4	10.3	10.7	1
e	7	13	2.8	4.3	1
f	3	22	11.4	13.2	1
g	7	22	9.5	13.1	1
h	6	19	7.2	10.0	1
i	6	10	4.7	1.1	2
j	7	10	4.0	1.8	2
k	5	9	6.1	0.5	2
l	4	7	8.2	2.5	2



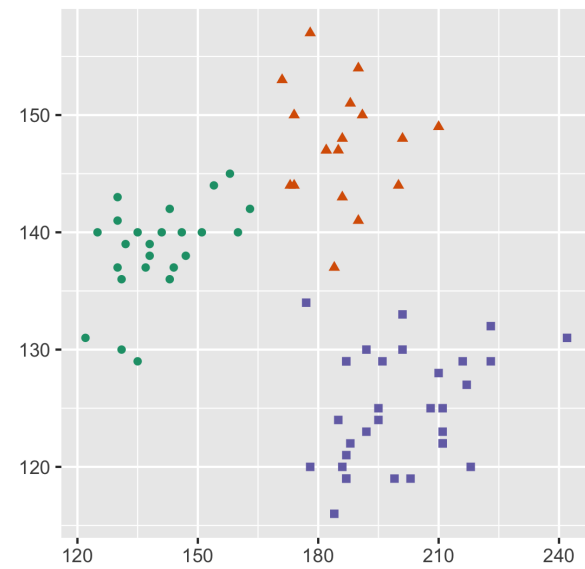
## Example



$k = 2$

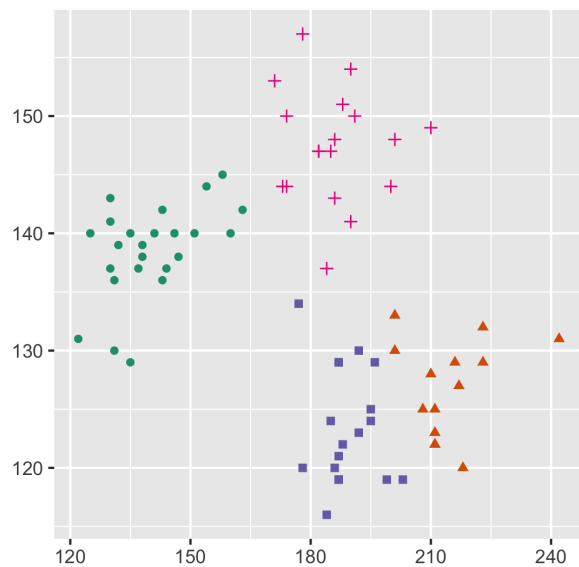


$k = 3$

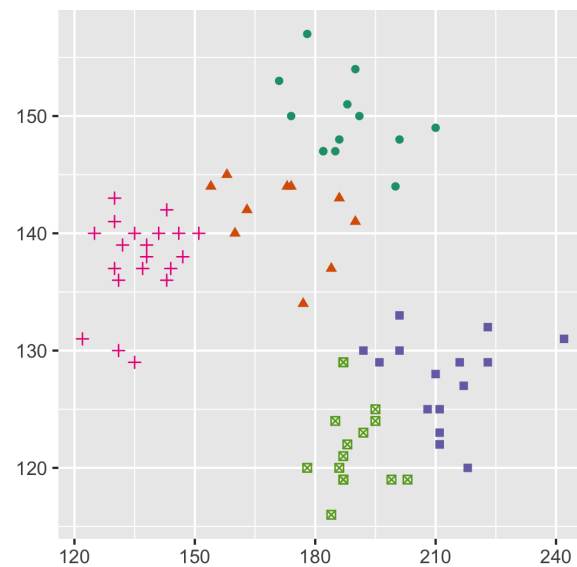




$k = 4$





$k = 5$





# Choosing k

## Cluster statistics

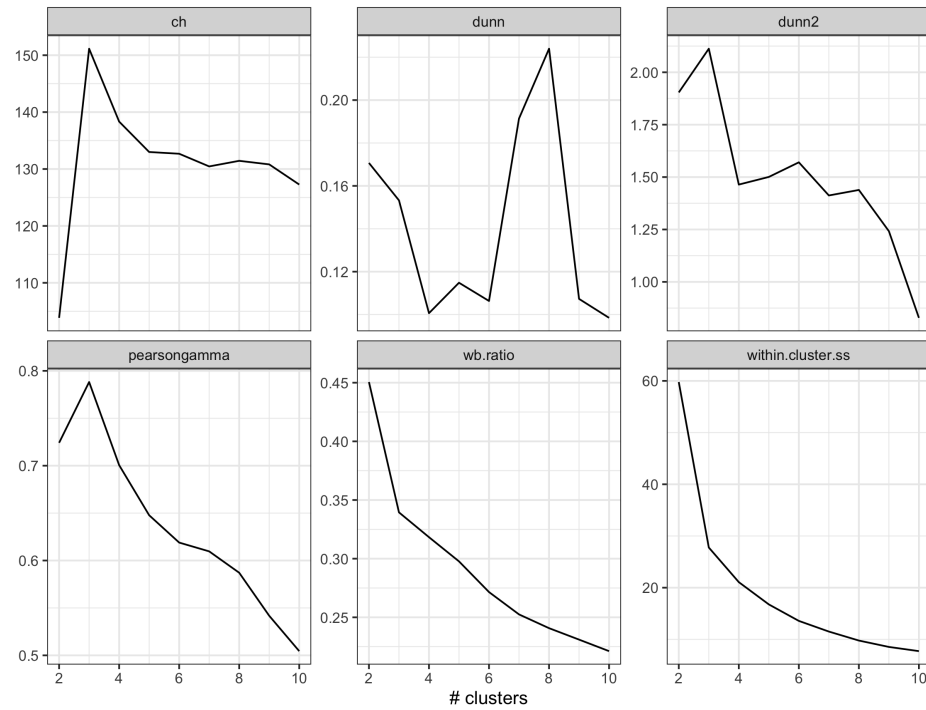
 **WBRatio**: average within/average between want it to be low, but always drops for each additional cluster so look for large drops

 **Hubert Gamma**:  $(s_+ - s_-)/(s_+ + s_-)$  where  $s_+$  = sum of number of within < between,  $s_-$  = sum of number within > between, want this to be high

 **Dunn**: smallest distance between points from different clusters/maximum distance of points within any cluster, want this to be high

 **Calinski-Harabasz Index**:  $\frac{\sum_{i=1}^p B_{ii}/(k-1)}{\sum_{i=1}^p W_{ii}/(n-k)}$  want this to be high

## Choosing k

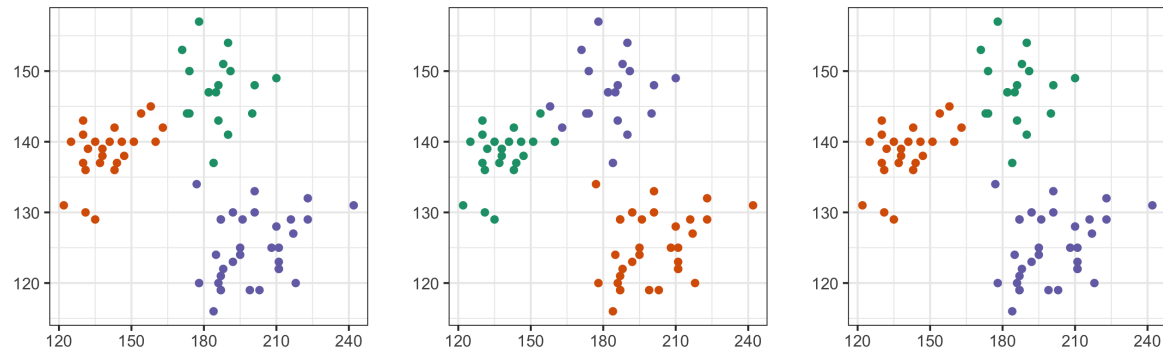


## k-means caveats

### Effect of seed

||| The k-means algorithm can yield quite different results depending on the initial seed.

||| Example runs used 5 random starts, and used the `within.cluster.ss` metric to decide on the best solution.



## Interpoint distance measures

### Euclidean

Cluster analysis depends on the interpoint distances, points close together should be grouped together

Euclidean distance was used for the example. Let  $A = (x_{a1}, x_{a2}, \dots, x_{ap})$ ,  $B = (x_{b1}, x_{b2}, \dots, x_{bp})$

$$\begin{aligned} d_{EUC}(A, B) &= \sqrt{\sum_{j=1}^p (x_{aj} - x_{bj})^2} \\ &= ((X_A - X_B)^T (X_A - X_B)) \end{aligned}$$

## Other distance metrics

▮▮▮ Mahalanobis (or statistical) distance

$$\sqrt{((X_A - X_B)^T S^{-1} (X_A - X_B))}$$

▮▮▮ Manhattan:

$$\sum_{j=1}^p |(X_{aj} - X_{bj})|$$

▮▮▮ Minkowski:

$$\left( \sum_{j=1}^p |(X_{aj} - X_{bj})|^m \right)^{1/m}$$

## Distances for count data

▮▮▮ Canberra:

$$\frac{1}{n_z} \sum_{j=1}^p \frac{X_{aj} - X_{bj}}{X_{aj} + X_{bj}}$$

▮▮▮ Bray-Curtis:

$$\frac{\sum_{j=1}^p |X_{aj} - X_{bj}|}{\sum_{j=1}^p (X_{aj} + X_{bj})}$$

## Interpoint distance measures - Euclidean

Rules for any metric to be a distance

1.  $d(A, B) \geq 0$
2.  $d(A, A) = 0$
3.  $d(A, B) = d(B, A)$
4. Metric dissimilarity satisfies  $d(A, B) \leq d(A, C) + d(C, B)$ , and an ultrametric dissimilarity satisfies  $d(A, B) \leq \max\{d(A, C), d(C, B)\}$

# Made by a human with a computer

Slides at <https://iml.numbat.space>.

Code and data at <https://github.com/numbats/iml>.

Created using R Markdown with flair by [xaringan](#), and [kunoichi](#) (female ninja) style.



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).



